Physics Department
Faculty of Pure and Applied Sciences
University of Cyprus

Application of an Efficient and Accurate Generalized Born Model to High Throughput Computational Protein Design

Phd thesis by
Savvas Polydorides

Under the supervision of Georgios Archontis

May 2011

Abstract

Computational Protein Design (CPD) is a very promising *state of the art* methodology for high-throughput protein and ligand mutagenesis studies. It has been continuously developed in recent years by numerous studies, which address its essential ingredients: the algorithms employed to search the vast sequence / conformational space and the scoring functions, used to assess different sequences and conformations.

In the present thesis we introduce and test methodological advances in both of the above ingredients. With respect to the scoring function, we implement an accurate and computationally efficient model of solvent effects, based on the generalized Born approximation. With respect to the exploration of the sequence / conformation space, we implement and test two criteria: (i) an absolute affinity criterion, which identifies protein sequences that minimize the association free energy of a protein and a specific molecule (ligand); (ii) a relative affinity criterion, which identifies protein sequences that minimize the association free energy difference betwen two complexes.

The accuracy of our solvent model is first tested against a benchmark Poisson model of continuum electrostatics, by calculations that introduce conformational changes and a broad set of mutations in a series of proteins. We then combine our solvent treatment with an atomic-detail representation of solvent interactions and test it further on binding-affinity calculations for several point mutants of the Aspartyl-tRNA synthetase and Tyrosyl-tRNA synthetase. We finally apply it on a specific design problem, the change in amino acid specificity of the protein Asparaginyl-tRNA synthetase (AsnRS). The engineered amino acid binding site of AsnRS contains physically reasonable mutations and is structurally robust, as verified by the conformational stability of the designed sequences during multi-ns molecular dynamics simulations. Furthermore, several of the sequences demonstrate a reverse specificity, favoring the target amino acid aspartic acid over the native amino acid asparagine. These results suggest that our model and the combined stability / affinity criteria employed here constitute improvements to earlier CPD studies of modified AsnRS specificity. The design is not absolutely successful, as experimental activity measurements with some of the proposed sequences fail to show activity for aspartic acid. Nevertheless, these experimental results are consistent with our design.

A second, application of our solvent treatment studies a problem that is related to CPD, protein acid / base equilibria. For a test set of six proteins and 78 titratable groups, the model performs well, with a reasonable rms error.

Acknowledgements

First and foremost I want to thank my advisor Georgios Archontis. It has been an honor to work with him all these years.

I am especially grateful to Tom Simonson and all his group members at the Ecole Polytechnique: Alexey Aleksandrov, Anne Lopes, Najette Amara and Marcel Schmidt am Busch for good advice and collaboration. I would also like to acknowledge Pierre Plateau and Caroline Aubard for their experimental contribution.

I would like to thank my committee members: Ilias Iliopoulos, Fotios Ptohos, Nicolaos Toumbas and Vasilis Promponas for their time, interest, and helpful comments and lastly, I would like to thank my parents for all their support and encouragement.

This thesis is dedicated to
Maria

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Computational Protein Design (CPD) is a set of *in silico* methods for the systematic search in the conformational and sequence space of proteins, and the identification of sequences and/or folds with desirable properties. CPD methods compare different states (e.g. a series of protein sequences, a set of protein conformations, a series of biomolecular complexes involving a protein with different ligands). CPD methods have been developed intensively in recent years [1–13]. Several applications of CPD methods have already succeeded in stabilizing specific protein folds [14–18], designing new proteins [19–25] and enzyme active sites [26–30] with altered activity [31; 32], improving the binding affinity of proteins for their native ligands [33–38], introducing novel specificities [39–43], optimizing ligand entrance and escape pathways [44–46], creating water-soluble variants of membrane proteins [47–49], redesigning protein-protein interfaces [50–55], assembling all-in-one proteins with multiple properties [56; 57], creating optimized protein libraries [58–60].

The success of CPD methods depends on two key factors: (i) the ability to search efficiently states from a large, representative portion of the available structure / conformational space of the molecule targeted; indeed, a major obstacle faced by CPD methods results from the extremely large number of degrees of freedom. In a typical CPD calculation, selected subsets of protein residues are allowed to change "chemical state" (chemical type and/or conformation). Chemical types are usually selected from the list of twenty natural amino acids, and conformations from a "rotamer library", which includes a set of distinct combinations of side chain torsional angles for each chemical type. Consider a moderate protein of 100 residues; assuming that 20 positions change both chemical type (from the 20 natural amino acids) and side chain conformation (with an average of 10 average conformations per amino acid chemical type), and the remaining 80 amino acids change only conformation, the total number of resulting sequences and conformations is $200^{20} \times 10^{80}$. Although computational resources available for research grow rapidly, such astronomical numbers preclude an exhaustive search of the sequence/conformational space, even for state-of-the-art su-

percomputers. (ii) A second key factor for the success of CPD calculations is the ability to weigh accurately the relative probabilities of the considered states, by a well-chosen "scoring" function. Usually these two factors are oppositely related (speed is increased at the expense of accuracy) and recent methodological improvements aim for an optimum compensation.

In the present thesis we introduce novel modifications, which target both methodological aspects of CPD. With respect to the "scoring" function, we develope an efficient free-energy model, that incorporates solvent effects into protein design and is based on the generalized Born approximation [61]. With respect to the exploration of the sequence/ conformational space, we implement and test two criteria. The first criterion searches for sequences that lower the free energy of association between a protein and a specific ligand (the protein affinity); the second searches for sequences that lower the difference in association free energy difference between two complexes (the affinity of a protein for a molecule, *relative* to a second molecule). The modifications are applied to an important biological problem, the change in amino acid affinity of the protein Asparaginyl-tRNA synthetase (AsnRS). The scoring function is also applied to the related problem of computing protein acid/base constants, or pKas. These novelties are further detailed below.

Physical scoring functions usually describe protein interactions via energy terms from molecular mechanics force fields [62; 63]. The surrounding solvent also affects the intra- and intermolecular interactions, and plays a critical role on the structural stabilization of proteins [64–67] and their functionality. An explicit representation of water molecules in atomistic models that describe a protein or a biomolecular complex is not tractable in CPD, as it leads to an enormous increase of the system degrees of freedom. For this reason, computational studies often use an implicit representation of the solvent, via the incorporation in the energy function of free-energy terms that describe the influence on intramolecular interactions due to the surrounding solvent [68; 69].

An essential requirement for physical scoring functions is their residue-pairwise decomposability, i.e. the ability to express the total energy as a sum of terms, which depend on the coordinates of individual residue pairs. This property allows the calculation of interactions between pairs of residues (in all possible chemical types) *prior* to the actual design, and their storage in an "interaction-energy" matrix [14; 70–73]. During the design, the energy of a particular sequence / conformation (i.e. a specific selection of chemical types) can be reconstructed and updated efficiently, by choosing appropriate elements from this matrix.

Continuum models approximate the highly inhomogeneous dielectric medium of a protein and its surrounding environment, as a low-dielectric cavity (protein) embedded in a high dielectric medium (solvent). The major representatives of continuum models are the Poisson-Boltzmann (PB) [74] approximation, and the generalized Born (GB)

[61] approximation, which compute solvation energies by numerical methods (PB) or analytical expressions (GB). The PB model is considered as the benchmanrk of accuracy, but is computationally more expensive than the GB approximation. Simpler models which express the solvation energy as a combination of a Coulomb term and a term proportional to the solvent accessible surface area (CASA model) [75] are also employed in CPD [76–80], preferably for applications of low energy resolution.

Many successful applications of the GB model in calculations of protein solvation [81], protein dynamics [82; 83], ligand binding [84; 85] and protein folding [86–89] encourage its employment in protein design. However, the generalized Born model (and the Poisson-Boltzmann approximation) are many-body quantities, that depend on the shape of the entire molecule and cannot be written in pairwise-decomposable form. One of the novelties of this thesis is to apply a residue-pairwise approximation to the Hawkins - Crammer - Truhlar generalized Born model [90].

In 2005, Archontis and Simonson derived a residue-pairwise generalized Born approximation [91] applicable in CPD. Here, we combine this model with the AMBER all-atom force field of protein interactions [92] and apply it for the first time in CPD. We conduct a series of tests, in which we introduce extensive conformational (rotamer) changes and chemical mutations into a diverse protein set. The tests show that the function is able to reproduce free-energies from a more accurate (but computationally inefficient) solution of the Poisson equation (Chapter 5). The combined function is tested further on binding-affinity calculations for several point mutants of the Aspartyl-tRNA synthetase and Tyrosyl-tRNA synthetase (Appendix B). We then apply the model to a specific, important biological problem: the modification of amino acid specificity in the protein Asparaginyl-tRNA synthetase (AsnRS). This protein recognizes a specific amino acid (asparagine) and the corresponding transfer RNA molecule ($tRNA^{Asn}$) and catalyzes the creation of a complex, which contributes to protein synthesis (Chapter 8). Change in the amino acid specificity of AsnRS and the synthetases in general helps to understand better the biological role and mechanisms of synthetases and contributes to the modification of the genetic code [93].

The quality of our design is compared with a previous effort, that utilized a simpler scoring function, combining a Coulomb/Accessible Surface Area (CASA) model of solvent effects [94] with a polar-hydrogen energy function [76] (Chapter 6). Our design produces more physically reasonable sequences, with inverted affinity (bind more strongly aspartic acid, with respect to the native AsnRS ligand, asparagine), and good structural stability, as verified by multi-ns molecular dynamics simulations in explicit solvent (Chapter 9). In contrast, the sequences resulting from the CASA model contained negatively charged residues in the vicinity of the (also negatively charged) target ligand (aspartic acid). Molecular dynamics simulations with several of the designed AsnRS:AsnAMP complexes showed that these negative-charge insertions destabilized the complexes, causing significant structural distortions in the binding site and the loss

of important protein-protein and protein-ligand interactions (Chapter 6, [77]).

An essential ingredient of our approach, which produced sequences of modified affinity and high stability (as manifested by the MD simulations), was the implementation and use of combined criteria by the sequence/conformational search algorithm, which took into account both structural stability and affinity. The theoretical foundation of these criteria and their implementation method is described in Chapter 7. The criteria are combined with a post-design filtering of the resulting sequences, consisting of a multi-step computational protocol of conformational (rotamer) exploration and energy minimization calculations. The application of the criteria to the AsnRS design problem is detailed in Chapter 7.

The methodological question of an improved treatment of aqueous solvent through a generalized Born model has broad relevance for protein modelling in general. In particular, the CPD problem has a close relationship to the problem of computing protein proton binding through acid/base constants, or pKas. Indeed, whereas CPD explores different amino acid side chain types and their preferred conformations, pKa calculations explore different side chain titration states and their preferred conformations. The metholodogy and software described in Chapters 5-7 was also applied to the pKa problem. For a test set of six proteins and 78 titratable groups, the model performs well, with a reasonable rms error.

The outline of the thesis is as follows. In Chapter 2, we outline the basic theoretical framework of biomolecular free-energy calculations, with emphasis on high-accuracy, low-throughput CPD methods. In Chapter 3 we describe the elements of high-throughput CPD methods, including the types of employed scoring functions and the classes of searching algorithms. We also describe illustrative applications, which demonstrate the current sophistication of high-throughput CPD methods. In Chapter 4 we summarize the most important implicit-solvent models used in computational studies, focusing on continuum electrostatics approximations. In Chapter 5 we present the residue-GB approximation, which constitutes the implicit-solvent model applied in this thesis. In Chapter 6 we compare the residue GB model with CASA, in various applications of protein design, and describe the predictions of CASA in an earlier attempt of AsnRS design. Chapter 7 describes the Proteus sequence/conformation searching program, and our implementation of absolute and relative affinity criteria. Chapter 8 describes the structural properties and the action of aminoacyl-tRNA synthetases. The next Chapter 9 presents our design of amino acid specificity on Asparaginyl-tRNA synthetase. The Appendices contain computational information, parts of the XPLOR scripts and three publications, related to our work. The first (Appendix A) present examples of XPLOR files, used to compute interaction energy matrices used in the design. Appendix B includes our residue-GB application to the amino acid design of AsnRS. Appendix C presents the application of the residue-pairwise GB model in proton binding. Appendix D presents work from the early stages of my doctoral studies,

which focused on an application of the MM-PBSA end-point free energy approximation (Chapter 2) to complexes of Ribonuclease A with a series of dinucleotide inhibitors.

Chapter 2

Computational Protein Design - Low Throughput Methods

This chapter contains a brief presentation of more accurate, "low-throughput" methods, which compare a small number of systems (e.g. a set of protein conformations, or a series of protein-ligand complexes). In the next chapter we focus on high-throughput methods, which can be used to routinely evaluate large numbers of states.

2.1 Low Throughput Methods

Low-throughput methods evaluate the relative probabilities of two states A and B in terms of their corresponding free energies [$p(B)/p(A) = \exp[-\beta(G_B - G_A)]$; $\beta = 1/(k_B T)$]. These methods usually represent the energy of each particular state by an atomic-detail biomolecular effective energy function [e.g. Eq. (3.2) in Chapter 3], and evaluate the corresponding free energy difference by an appropriate statistical-mechanical expression. To achieve this, they perform an exhaustive exploration over the phase space of the states considered (e.g. they generate a representative ensemble of protein and/or solvent conformations by a Monte Carlo or Molecular Dynamics algorithm). This is a computationally demanding procedure, which restricts the practical use of these methods to a small number of states at a time.

2.1.1 Pathway Methods

Pathway methods compute the free energy difference $G_B - G_A$ between two distinct states A and B, by following a reversible pathway that converts one state to the other. Typical applications compute the solvation free-energy of a biomolecule, the association free energy of a biomolecular complex or the folding free energy of a protein. The computation is based on the fact that the free energy difference between the two states

A and B can be expressed as an expectation value:

$$\begin{aligned}\Delta G = G_B - G_A = -k_B T \ln[Q_B/Q_A] &= -k_B T \ln\langle \exp[-\beta(H_B - H_A)]\rangle_A \\ &= +k_B T \ln\langle \exp[+\beta(H_A - H_B)]\rangle_B \quad (2.1)\end{aligned}$$

In the above Free-Energy Perturbation (FEP) expression, H_A and H_B are the Hamiltonians of states A and B; Q_A and Q_B are the corresponding partition functions, which in the isothermal-isobaric ensemble are given by the expression [95]

$$Q_{A/B}(T, p, N) = \int \mathrm{d}V \, \exp[\beta p V] \int \mathrm{d}r^N \, \exp[-\beta H_{A/B}] \quad (2.2)$$

If the two states A and B have an identical number and type of atoms, the difference $H_B - H_A$ reduces to the difference in the corresponding potential energies $U_B - U_A$.

Eq. (2.1) involves ensemble averages $\langle \ldots \rangle_A$ or $\langle \ldots \rangle_B$, respectively over state A or B, of the quantity $\exp[\pm\beta(H_B - H_A)]$. In principle, these averages can be computed by a Monte Carlo algorithm, which generates a representative number of equilibrium microstates of the initial (A) or final (B) state, in the appropriate ensemble. Under the ergodic hypothesis, the same averages can also be computed by Molecular Dynamics (MD) simulations in a suitably defined ensemble (e.g. under the assumptions of constant energy and volume, or temperature and pressure). The MD simulation will solve the equations of motion for system A and will generate a time series of K equilibrium conformations ($\{r_i(t_j)\}$, $j = 1, \ldots, K$). These conformations can be used to compute a time series of the quantity $\exp[-\beta(H_B - H_A)]$. Then, the expectation value entering in the free energy difference can be computed by the following time average:

$$\begin{aligned}\langle \exp[-\beta(H_B - H_A)]\rangle_A &= \frac{1}{T}\int_{t_0}^{t_0+T} dt' \, \exp[-\beta(H_B(t') - H_A(t'))]\rangle \\ &\approx \frac{1}{K}\sum_{j=1}^{K} \exp[-\beta(H_B((\{r_i(t_j)\})) - H_A(\{r_i(t_j)\}))] \quad (2.3)\end{aligned}$$

In the same way, a simulation with the final Hamiltonian (H_B) can be used to compute the expectation value $\langle \exp[+\beta(H_B - H_A)]\rangle_B$.

The expectation values obtained by Eq. (2.1) are accurate only if the two end-states A and B are very similar. To deal with the more general case of sufficiently different end-states, the transformation is accomplished through a series of N successive intermediate steps, in which the Hamiltonian is progressively changed from the functional form of state A (H_A) to that of state B (H_B). At each step, a separate simulation is conducted with a corresponding intermediate (hybrid) Hamiltonian H_i. The total free energy change can then be evaluated from the multi-step FEP expression

$$\Delta G = G_B - G_A = -k_B T \sum_{i=0}^{N} \ln\langle \exp[-\beta(H_{i+1} - H_i)]\rangle_i, \tag{2.4}$$

where $H_0 \equiv H_A$ and $H_{N+1} \equiv H_B$.

A similar approach, termed "thermodynamic integration" (TI), introduces a series of non-physical intermediate systems, described by a hybrid Hamiltonian H(λ) [95]. The parameter λ connects the initial and final states [i.e. H($\lambda_{\rm in}$) = H_A and H($\lambda_{\rm fi}$) = H_B]; varying λ from $\lambda_{\rm in}$ to $\lambda_{\rm fi}$, the system changes from state A to state B. The free energy change can be computed by the following TI expression:

$$G_B - G_A = \int_{\lambda_{\rm in}}^{\lambda_{\rm fi}} d\lambda \, \frac{\partial F}{\partial \lambda} = -k_B T \int_{\lambda_{\rm in}}^{\lambda_{\rm fi}} d\lambda \, \frac{d}{d\lambda} \ln Z(\lambda) = \int_{\lambda_{\rm in}}^{\lambda_{\rm fi}} d\lambda \, \langle \frac{\partial H}{\partial \lambda} \rangle_\lambda \tag{2.5}$$

In the simplest formulation, $\lambda_{\rm in} = 0$, $\lambda_{\rm fi} = 1$, and the hybrid energy function depends linearly on λ: H(λ) = $(1-\lambda)H_A + \lambda H_B$. In this case, Eq. (2.5) becomes:

$$G_B - G_A = \int_0^1 d\lambda \, \langle H_B - H_A \rangle_\lambda \tag{2.6}$$

In practice, separate runs are performed with Hamiltonians defined at a finite, discrete set of values $\{\lambda_i\}$. Using the trajectories from these runs, the corresponding time series $\langle H_B - H_A \rangle_{\lambda_i}$ are computed and the results are summed with Eq. (2.7):

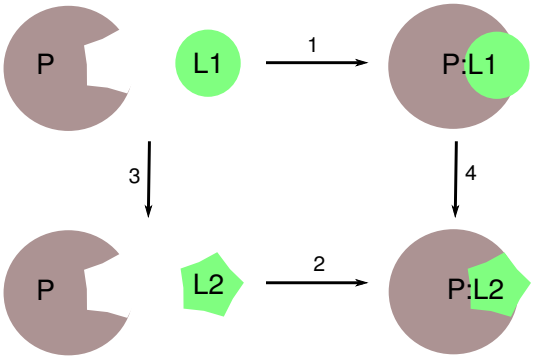
$$G_B - G_A \approx \sum_i (\lambda_{i+1} - \lambda_i) \langle H_B - H_A \rangle_{\lambda_i} \tag{2.7}$$

The FEP [Eq. (2.4)] and TI [Eq. (2.6)] expressions are equivalent in the limit of infinitely small changes. Indeed, rewriting Eq. (2.4) for the hybrid potential and assuming small changes in λ between adjacent steps, the logarithmic term can be simplified by keeping only first-order terms in a Taylor expansion:

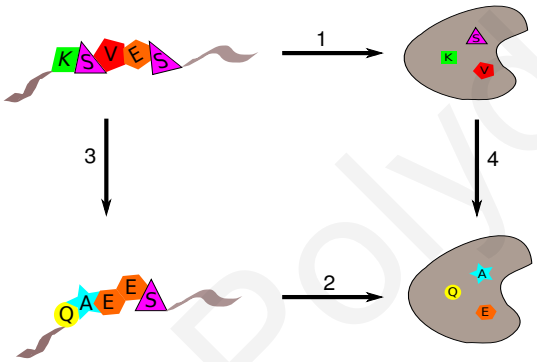
$$\begin{aligned} \Delta G &= -k_B T \sum_i \ln\langle \exp[-\beta \frac{\partial \mathrm{H}(\lambda)}{\partial \lambda} (\lambda_{i+1} - \lambda_i)]\rangle_{\lambda_i} && (2.8) \\ &\approx -k_B T \sum_i \ln\langle (1 - \beta(H_B - H_A))(\lambda_{i+1} - \lambda_i)\rangle_{\lambda_i} \\ &\approx \sum_i (\lambda_{i+1} - \lambda_i) \langle H_B - H_A \rangle_{\lambda_i} && (2.9) \end{aligned}$$

Computational Alchemy

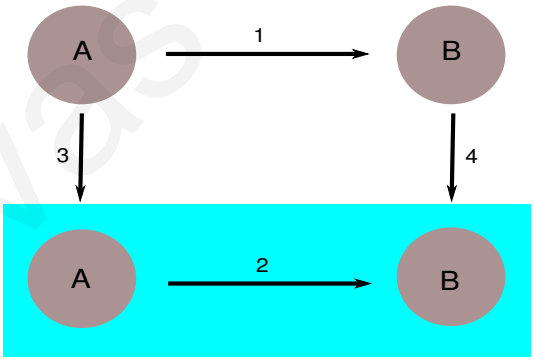
Using the above FEP or TI formulation, the formation of solvated complexes, folding transitions or solvation processes can be compared with the aid of suitably chosen thermodynamic cycles, such as the ones shown in Fig. 2.1.



(a) Complex formation



(b) Folding transition



(c) Solvation

Figure 2.1: Thermodynamic cycles describing (a) the formation of two complexes (P:L1 and P:L2), of a protein (P) and two ligands L1 and L2. (b) the folding of two proteins and (c) the solvation of two solutes A and B.

The above thermodynamic cycles can be used to choose suitable pathways, along which the computation of a free-energy change can be accomplished with higher accuracy. For example, consider the thermodynamic cycle 2.1a, which connects the complexes between a protein (P) and two different small molecules (ligands L1 and L2). The experimentally measurable binding processes are described by the horizontal arrows 1 and 2. The vertical arrows 3 and 4 describe the "alchemical" transformation of one solute (L1) to the other (L2), or one complex (P:L1) to the other (P:L2). Since the free-energy is a state function, the net free-energy change in the full cycle is zero; this implies that the free energy differences between the vertical arrows (step 2 - step 1) or horizontal arrows (step 4 - step 3) are equal:

$$\Delta\Delta G = \Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3 \tag{2.10}$$

Thus, to compute the relative association free energy $\Delta\Delta G$ of the two complexes, it is sufficient to follow the horizontal paths. Even though the vertical transformations do not have a physical meaning and are experimentally inaccessible, they can be easily accomplished in a computer, by varying the Hamiltonian from the functional forms corresponding to the complex P:L1 (or ligand L1) to the one of the second complex P:L2 (or ligand L2). This was described above and is often termed as "computational alchemy" [91; 96–99].

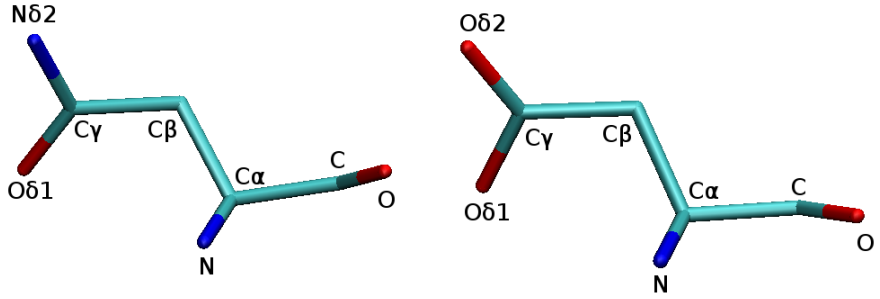
In a typical implementation of this approach, the biomolecular system (e.g. the protein-ligand complex) can be partitioned into three groups of atoms: Group "1" contains the part that remains invariant in the transformation (e.g. the protein, solvent and part of the ligand that is common in the initial and final state); Group "2" contains the set of atoms that constitute part of the initial state (in ligand L1) but are not part of the final state (missing from ligand L2); Finally, group "3" contains atoms that belong to the final state (in ligand L2) but are absent from the initial state (ligand L1). The total system is described by the hybrid Hamiltonian

$$\mathrm{H}(\lambda) = H_{11} + (1-\lambda)(H_{12} + H_{22}) + \lambda(H_{13} + H_{33}) \tag{2.11}$$

where "1" describes the invariant group of atoms, and "2", "3" describe the groups of atoms that constitute part of the initial (P:L1) and final state (P:L2), respectively; then, H_{XY} describes the interaction terms between groups X and Y.

An example is shown in Fig. 2.2, which describes the transformation from the aminoacid asparagine to aspartic acid. Atoms C_γ, $O_{\delta 1}$,$N_{\delta 2}$ correspond to group 2. Atoms C_γ, $O_{\delta 1}$,$O_{\delta 2}$ correspond to group 3. Atoms C_γ and $O_{\delta 1}$ belong to both groups 1 and 2, because they have different charges in Asp and Asn. The remaining atoms correspond to group 1.

With the notation of Eq. (2.11)



(a) Asparagine (b) Aspartic acid

Figure 2.2: Molecular representation of two similar sidechains, asparagine and aspartic acid. Atoms C_α, C_β, C, O and N are common for both sidechains. Atoms C_γ, $O_{\delta 1}$,$N_{\delta 2}$ are part of asparagine and C_γ, $O_{\delta 1}$,$O_{\delta 2}$ part of aspartic acid.

$$H_A \equiv H_{11} + H_{12} + H_{22}, \qquad H_B \equiv H_{11} + H_{13} + H_{33},$$
$$H_B - H_A = H_{13} + H_{33} - H_{12} - H_{22} \tag{2.12}$$

In the "dual topology" scheme, all three groups of atoms are simultaneously present, but their interactions are weighed by a suitable parameter λ, as shown in Eq. (2.11). The hybrid function $H(\lambda)$ describes the alchemical transformation $P : L1 \rightarrow P : L2$ as a function of the slowly varying parameter λ. Starting from $\lambda = 0$, L2 is "growing in" while L1 is "growing out"; for intermediate values of λ, a mixture of the two ligands is simultaneously present. The transformation can also be performed in the opposite direction $P : L2 \rightarrow P : L1$. The relative free-energy difference is computed by Eq. (2.8), using the Hamiltonian difference of Eq. (2.12).

The horizontal paths of Fig. (2.1) are more challenging computationally; nevertheless, we describe below methods by which they can also be studied.

Annihilation and Separation

Such pathway methods can be employed to compute the absolute binding affinity of a ligand to a protein. The "separation pathway" follows the horizontal paths of Fig. 2.1a, [100; 101], in which the two members of a biomolecular complex are gradually distanced. This can be accomplished by a series of simulations, in which the ligand is restrained (e.g. with the aid of a harmonic potential) at different positions along a linear path, defined by the translation vector $\vec{r}_{lig}(\lambda) = \vec{r}_{lig}(A) + \lambda\, r\, \hat{u}$, where r is the distance between the end positions ($r \equiv |\vec{r}_{lig}(B) - \vec{r}_{lig}(A)|$) and $\hat{u}$ is a unit vector, parallel to the difference vector $\vec{r}_{lig}(B) - \vec{r}_{lig}(A)$. The parameter λ takes N distinct values between 0 and 1. For each value λ_i a separate simulation is performed, with the ligand restrained around the reference position $\vec{r}_{lig}(\lambda)$. This restraint ensures that each simulation will sample adequately the space around the reference position, in the

philosophy of the "umbrella sampling" method [102]. Using the N biased simulations, a set of N position probability distributions $\rho_i^b(\lambda)$ is obtained. The bias is removed and a single unbiased probability can be obtained by a suitable linear combination of the biased probabilities, using the Weighted Histogram Analysis Method (WHAM) [103]. The unbiased distribution can be used to compute a potential of mean force (PMF) [95; 100], i.e. a free-energy profile as a function of the parameter λ (equivalently, the separation distance $\lambda\, r$) between the ligand and the protein:

$$W(\lambda) = -k_B T \ln[\rho(\lambda)] \tag{2.13}$$

The absolute protein-ligand binding free energy is then given by the expression

$$\Delta G_{bind} = -k_B T \ln \left[C^0 \int \mathrm{d}\lambda \, \exp[-\beta W(\lambda)] \right] \tag{2.14}$$

The standard-state concentration C^0= 1 M = $1/V^0$ renders the argument of the logarithm dimensionless (the volume integral is over the 3-dimensional space). The separation method is applicable to solvent-exposed binding sites, where the two binding partners can be moved away from each other in a simple (e.g. linear) path, without inducing large structural deformations [104; 105].

A different pathway which does not require a solvent-exposed binding site is the "double decoupling pathway" [100; 101; 106], shown in the following diagram:

$$L_{sol} \xrightarrow{\Delta G_2} L_{gas}$$

$$P : L_{sol} \xrightarrow{\Delta G_1} P_{sol} + L_{gas}$$

$$P_{sol} + L_{sol} \xrightarrow{\Delta\Delta G = \Delta G_2 - \Delta G_1} P : L_{sol}$$

The pathway requires the simulation of two systems (the solvated complex and the solvated, dissociated ligand) and consists of two parts [105; 107–111]: in the annihilation part, described by the top arrow in the above diagram, the ligand is transferred from solution to the gas phase (annihilation); in the decoupling part, described by the middle arrow, the ligand is transferred from the binding site to the gas phase, by gradually eliminating its interactions with the surrounding system (protein and solvent). During this transformation, the ligand is constrained (e.g. via a harmonic potential) to remain inside the binding cavity. The free energy changes are computed by FEP/TI methods, including correction terms to account for the constrained ligand [100]. The absolute binding affinity (lowermost arrow) is determined by the difference $\Delta G_{bind} = \Delta G_2 - \Delta G_1$.

A double decoupling pathway can calculate the absolute binding free energy of a

series of ligands to a given protein [106; 107; 112]. It can also be applied to determine the binding free energy of specific water molecules, strongly interacting with the protein [107; 108].

2.1.2 End-Point Methods

End-point methods compute directly the free energies of the two end-states (e.g. a solvated complex and the dissociated, solvated molecules). They are based on certain assumptions, which reduce their accuracy compared to pathway or "alchemical" methods. At the same time, they are more efficient.

MM-PB(GB)SA Model

The Molecular Mechanics - Poisson Boltzmann Surface Approximation (MM-PBSA) method [109; 113; 114] uses a continuum electrostatics (Poisson-Boltzmann) approximation to evaluate the solvation free energy of a biomolecular system. A detailed example of an application of the MM-PBSA method to determine the relative affinities of a series of dinucleotide inhibitors for Ribonuclease A is presented in Appendix D. In its most general formulation, it requires three independent MD simulations, of the solvated free protein, the solvated free ligand and the solvated protein-ligand complex. These simulations are usually done with an atomic-level representation of the protein, ligand and water environment in a suitable equilibrium ensemble (e.g. constant temperature and/or constant pressure). In this way, a representative ensemble of biomolecular conformations at equilibrium is generated. The explicit solvent is then replaced by a continuum approximation, which enables the fast evaluation of the electrostatic solvation free energies of the complex and isolated protein / ligand, for the various conformations obtained from the MD.

The binding energy calculation is described by the thermodynamic cycle shown in Fig. 2.3. Step 1 corresponds to the association of protein and ligand in vacuum. The accompanying energy change has two contributions. The first originates from the formation of intermolecular non-bonded (electrostatic and vdW) interactions in the complex. The second reflects the possible structural reorganization of the protein and ligand in the complex, and affects both inter- and intramolecular non-bonded and bonded (bond, angle, dihedral, improper dihedral) energy terms. The various energy contributions are computed directly from the simulation, using the molecular mechanics energy function and the atomic coordinates.

Steps 3 and 4 correspond to the solvation of the complex and dissociated protein and ligand. The free-energy changes in these two steps, $G^{solv}_{P:L}$, G^{solv}_{P} and G^{solv}_{L}, are computed by a continuum-electrostatics (Poisson-Boltzmann, PB, or generalized Born, GB) approximation with an extra accessible surface-area (ASA) term, that accounts for non-polar solvation effects. The protein and ligand are treated as dielectric cavities

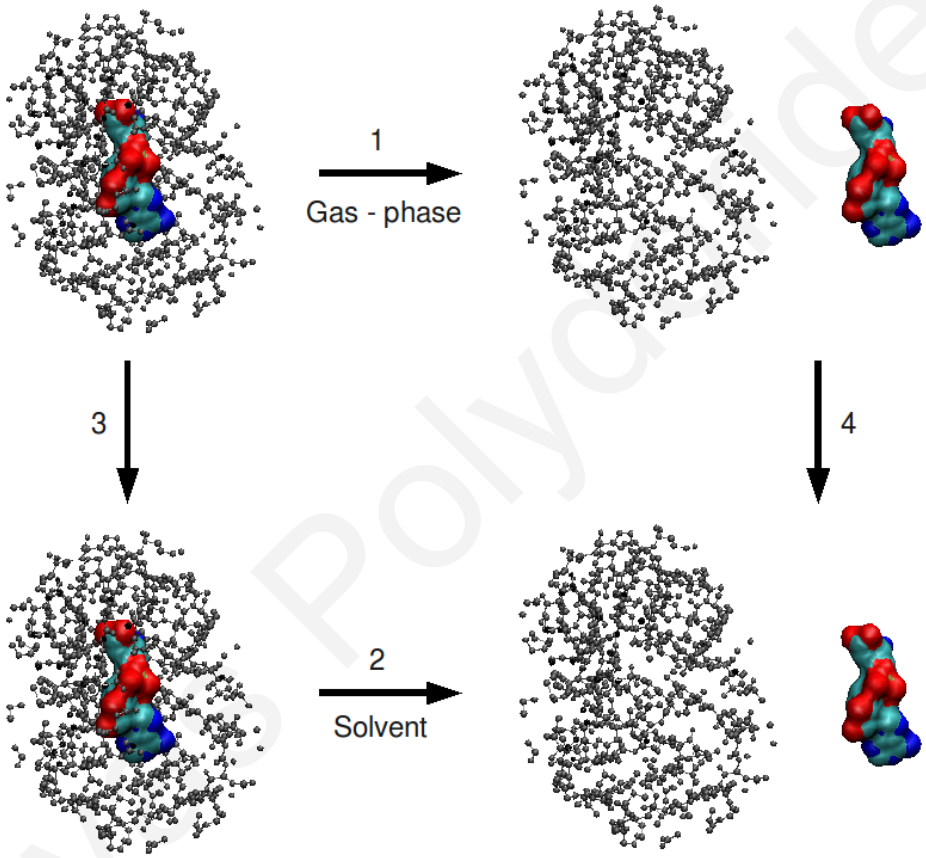


Figure 2.3: Thermodynamic cycle used to calculate the association free energy of a protein-ligand complex. In steps 1 and 2, the association takes place, respectively, in the gas phase and in solution. In steps 3 and 4, the complex and dissociated molecules, respectively, are transferred from the gas-phase into solution.

with embedded charges, immersed in a high dielectric medium (the solvent). In the case of the PB approximation, the electrostatic potential at each atom i of biomolecular state X (X=PL, P, or L) is determined in vacuum and solution by an arithmetic solution of the Poisson-Boltzmann equation (e.g. via a finite-difference or finite-element method [66]); subsequently, the change in solvation free energy of state X due to transfer from vacuum to solution is computed by the equation

$$G_X^{\text{solv}} = \frac{1}{2}\sum_{i\in X} q_i(V_i^{\text{sol}} - V_i^{\text{vac}}) + \sum_{i\in X}\sigma_i\, A_i \tag{2.15}$$

The first and second terms on the right hand side of the above equation correspond to the electrostatic and non-polar part of the solvation free energy change; q_i are the atomic charges, $V_i^{\text{sol/vac}}$ is the electrostatic potential at position i, respectively, in vacuum and in solution. A_i is the solvent accesible surface area of atom i and σ_i is a suitably parameterized surface-tension coefficient, which reflects the change in free-energy per unit surface area of solvent-exposure for the particular atom i.

In the case of the MM-GBSA approximation, the solvation free energy is expressed as an analytical (Generalized Born) expression of the biomolecular atomic coordinates [61; 115; 116]. Although the PB approximation is in general more accurate, GB models are significantly faster; recent implementations are of comparable accuracy with the PB approximation [67; 117; 118].

The MM-PBSA and MM-GBSA estimates often suffer from uncertainty introduced by the structural fluctuations in the complex and dissociated states. In the "single-trajectory" approximation, the assumption is made that the protein and ligand have identical structures in the complex and separated states (i.e. that there is no structural relaxation accompanying the formation of the complex). In this case, the only contributions to the association free energy are due to the formation of intermolecular non-polar and polar interactions. Despite the fact that structural relaxation is neglected, often this approximation yields more accurate results.

Even though the results are often dependent on the force field and the protein / ligand dielectric constant, several applications of the MM-PB/SA or MM-GB/SA method are able to reproduce experimental binding affinities [100; 119; 120]. Other applications of the method are the calculation of relative affinity ($\Delta\Delta G_{bind}$) for low molecular weight inhibitors [121] (see Appendix D) or relative stability (ΔG) in structural transitions [122; 123], the determination of key role residues in protein - ligand binding using computational alanine scanning (CAS) method [113; 124] and binding free energy decomposition (BFED) analysis [115].

Linear Interaction Energy (LIE) Models

The linear interaction energy (LIE) model [125; 126] approximates the protein-ligand binding energy as a linear combination of the electrostatic and van der Waals inter-

actions of the ligand with its surrounding environment in solution and in the protein-ligand complex.

$$G_{bind} = \alpha[\langle U_{vdw}\rangle_{\text{bound}} - \langle U_{vdW}\rangle_{\text{free}}] + \beta[\langle U_{elec}\rangle_{\text{bound}} - \langle U_{elec}\rangle_{\text{free}}] + \gamma \quad (2.16)$$

Each energy term in Eq. (2.16) is an expectation value of the corresponding interaction energy, that is extracted from two independent MD simulations: the free ligand in solution and the solvated protein-ligand complex. Water molecules are explicitly represented during MD, though recent studies[127] employ implicit models to speed up the calculation. The empirical parameters α, β, γ are fitted to experimental values to compensate for the absence of any (protein,ligand,water) internal energy contributions.

Despite its simplicity, this approximation works well [125; 126; 128], if the parameters are carefully calibrated. The nonpolar parameter α is usually set to 0.18; parameter β is usually adjusted between the linear-response value 0.5 [125] and a slightly smaller value (0.33). The parameter γ is a system-dependent constant, used to correct estimates of absolute binding energies.

Chapter 3

Computational Protein Design - High Throughput Methods

High-throughput computational protein design (CPD) or drug design methods evaluate routinely an extremely large number [O(10^6-10^9)] of possible sequences and/or structures of the targeted protein or complex. Usual objectives are to identify sequences and conformations with low folding free energies, and/or high affinity for a specific biomolecular partner. To accomplish this, CPD methods rank sequences and conformations by approximate, very efficient scoring functions; furthermore, they reduce the conformational space into a discrete set of states, and guide the search through the structure/sequence space with the aid of efficient searching algorithms.

CPD calculations usually fix the chemical identity and conformation for part of the protein (e.g. the entire backbone and selected side chains), and modify the orientation of all (non-fixed) side chains and the chemical identity of selected side chains (Fig. 3.1(a)). [16; 73; 129–132]. Chemical types of mutatable side chains are usually chosen from the set of twenty natural aminoacids, even though non-natural aminoacids have also been employed in protein design calculations [93; 133; 134]. Side chain conformations (rotamers) are taken from a discrete set (a finite rotamer library [135]); the side chain dihedral angles in employed rotamers are usually set to values of a favorable set of dihedral angles, frequently observed in experimental protein structures, or can be extracted from MD simulations. A number of discrete rotamers for each amino acid range from a single conformation for glycine or alanine, to $\backsim$50 for bigger side chains like arginine or lysine, compose the rotameric library employed in protein - ligand docking.

The sequence / structure space increases exponentially with the number of rotamers; hence, the method is feasible only for a reasonable number of discrete states. On the other hand, a high-resolution design requires a large rotameric library, which includes not only side chain rotamers but also a number of main chain scaffolds, which account for the flexibility of the protein main chain. This improves the accuracy of the design,

but also increases the computational demand. Although computational methods are continuously developed to reduce the time limit of such problems, it is obvious that a compensation between speed and accuracy must be made.

The considered sequences and/or conformations are assessed with the aid of a target function; this function could be the folding free-energy of a protein or a biomolecular complex (i.e. the difference $G_N - G_D$ between the free-energies of the folded (N) and denatured (D) states), or the association free energy of a biomolecular complex (the difference between the free energy of the complex and the dissociated molecules). The employed target functions can be classified in three broad categories; physical, empirical and statistical. These categories are further discussed in Section 3.1.

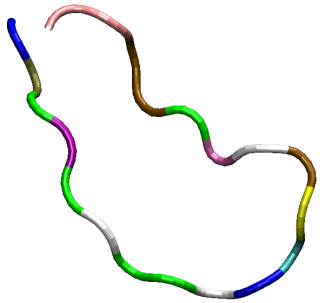
Even with the discretization of conformations, the remaining sequence/conformation space is enormous: assuming an average of 10 possible rotamer conformations per chemical type and 20 chemical types, each mutatable position has 200 distinct sequence/conformation states. For 10 mutatable positions, the total number of possible sequences and conformations is $200^{10} \approx 10^{23}$. Since the exhaustive consideration of all possible states is impossible, optimized sequences and conformations are typically selected by efficient deterministic or heuristic searching algorithms. Examples of such algorithms are further discussed in Section 3.2.

Fixing part of the system into a definite structure is an approximation that reduces the quality of the design. For this reason, efforts have been made toward introducing flexibility. Pecore et al. designed sequences targeting the WW domain [136], which were compatible with an ensemble of backbone folds. Another way is to allow small changes of the backbone, by rotating tripeptides through an axis joining the C_α atoms at the ends [8; 137].

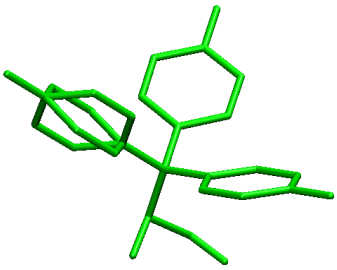
3.1 Scoring Functions

Scoring functions are used in CPD to compare different states (e.g. sequences and/or conformations compatible with a certain protein fold). Accuracy and speed are of crucial importance in CPD. A difficulty arises from the fact that these factors are inversely related; for this reason, the efficiency of the design method depends on the judicious combination of scoring function and searching algorithm. Current progress in this area, focuses on developing more efficient searching methods, with faster algorithms and more accurate potential energy functions.

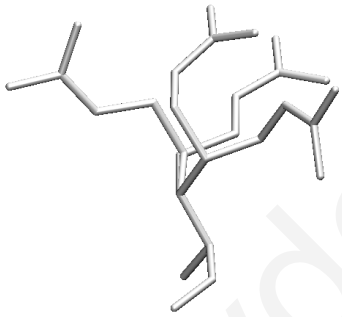
Scoring functions are divided into three categories: (i) physical (ii) empirical and (iii) knowledge-based.



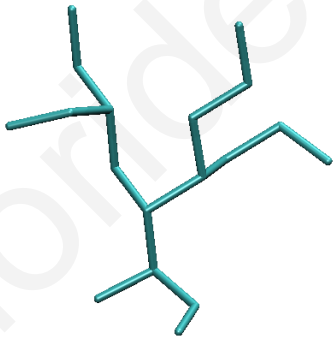
(a) Design scaffold



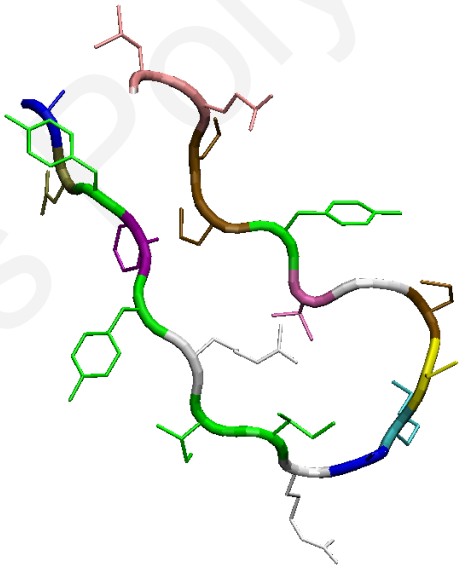
(b) Tyrosine rotamers



(c) Arginine rotamers



(d) Lysine rotamers



(e) Optimized structure

Figure 3.1: (a) The main chain of a protein is assumed to be fixed in a typical CPD calculation. It represents a "scaffold" on which a suitable, optimized set of side chain chemical types and/or rotamer conformations has to be fitted. Different colors show main chain segments of individual residues. (b)-(d) Each amino acid side chain is associated with a suitable number of high-probability conformations (rotamers). Examples of rotamers are shown for tyrosine, arginine and lysine. (e) Example of an optimized conformation, with a set of side chain chemical types/rotamers chosen in conjunction with the protein main chain scaffold.

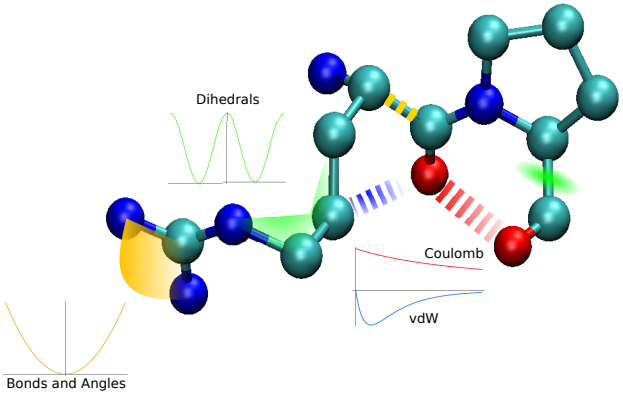


Figure 3.2: A schematic representation of a typical biomolecular potential energy function.

3.1.1 Physical Scoring Functions

Physical effective energy functions are based on molecular mechanics force field models, such as the ones used in biomolecular simulations [62; 63]. They are all-atom or coarse-grained additive functions of bonded contributions (bond streches, angle bends, dihedral torsions, improper angles) and non-bonded contributions (van der Waals and electrostatic interactions).

$$E = E_{bonds} + E_{angles} + E_{dihe} + E_{impr} + E_{vdW} + E_{elec} + E_{solv} \tag{3.1}$$

The last term models solvation contributions to the energy. Eq. (3.2) presents a typical all-atom physical energy function, in which the solvent contribution is absorbed in the dielectric constant ϵ_p, that is used to scale the Coulombic interactions.

$$\begin{aligned} E \;=\; & \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1+cos(n\phi-\delta)] \\ & + \sum_{\text{impropers}} k_\omega(\omega-\omega_0)^2 + \sum_{\text{non-bonded}} \left[\epsilon[(\frac{R_{min}}{r_{ij}})^{12} - (\frac{R_{min}}{r_{ij}})^6] + \frac{q_i\, q_j}{\epsilon_p\, r_{ij}}\right] \end{aligned} \tag{3.2}$$

The first four terms of Eq. (3.2) reflect the dependence of potential energy on (i) the lengths of covalent bonds, (ii) the angles between adjacent covalent bonds, (iii) the rotation around specific covalent bonds; (iv) An additional energy term (improper dihedral) maintains the chirality and planarity of specific groups (Fig. 3.2). Bond, angle and improper terms are approximated by harmonic oscillator potentials.

The functional form of Eq. (3.2) is adopted by several widely used biomolecular force-fields, such as CHARMM [62] and AMBER [138]. It contains a number of pa-

rameters [force constants $\{k_b, k_\theta, k_\phi, k_\omega\}$, equilibrium geometries $\{b_0, \theta_0, \phi_0, \omega_0\}$, atomic charges $\{q_i\}$, van der Waals energy (ϵ) and length scales ($R_{\min}$). These parameters are determined with the aid of structural, spectroscopic and thermodynamical data, as well as high-level quantum mechanical calculations of suitable model compounds.

The last two terms of Eq. (3.2) describe non-bonded *van der Waals* (vw) interactions (typically modelled by a Lennard-Jones (LJ) 6-12 potential), and *Coulombic* interactions. In typical force-fields, these interactions apply between atoms at least three covalent bonds apart in the same molecule, or in different molecules; hence they are referred to as "non-bonded" interactions. vw and Coulombic interactions between atoms separated by less than three covalent bonds in the same molecule are absorbed into the bonded energy terms.

Solvent effects play a key role in biological processes, such as protein stability and function [64–67], ligand binding [96; 139–143], protein-protein association [50]. For example, such terms favor the placement of hydrophobic side chains in the interior, and polar or charged side chain at the surface of a protein; they also account for the modulation in the interaction of charged atom pairs, due to the polarization of the surrounding solvent. For this reason, they must be taken into account by the scoring energy function. Explicit representation of water molecules is rarely attempted in high-throughput CPD studies [144; 145], due to combinatorial explosion, except for special cases where a small number of water molecules play important structural or functional roles (e.g. in an active site, a binding pocket, or at the protein - protein interface). The last term of Eq. (3.1) represents the contribution of solvent effects to the total (free) energy. Its inclusion in the energy function implies that the solvent is not represented explicitly but *implicitly*.

The simplest implicit-solvent (Coulomb/Accessible Surface Area or CASA) model contains screened Coulombic energy, combined with a term proportional to the solvent-exposed surface-area [94; 146; 147]:

$$E_{\text{solv}}^{\text{CASA}} = (\frac{1}{\epsilon_p} - 1) \sum_{i<j} \frac{q_i\, q_j}{r_{ij}} + \sum_i \sigma_i\, A_i \tag{3.3}$$

where ϵ_p is the common dielectric constant employed to scale electrostatic interactions.

With this solvation free-energy, the total electrostatic free energy of the system becomes

$$E_{\text{elec}}^{\text{tot}} = E_{\text{Coul}} + E_{\text{solv}}^{\text{CASA}} = \sum_{i<j} \frac{q_i\, q_j}{\epsilon_p\, r_{ij}} + \sum_i \sigma_i\, A_i \tag{3.4}$$

This model [76; 77] is routinely used in CPD calculations, due to its simplicity and computational efficiency [76–80]. Its success in a particular design problem, the modification of the amino acid specificity of the protein Asparaginyl-tRNA synthetase, is discussed in detail in Chapter 6.

More accurate approaches approximate the solvation energy by continuous electrostatic models such as the Poisson - Boltzmann (PB) approximation [74; 148] and the Generalized Born (GB) model [61]. Both methods treat the protein as a low dielectric cavity with embedded charges at the atomic positions, surrounded by a high dielectric medium, the solvent. In the PB approximation, the PB equation is solved arithmetically and provides the electrostatic potential V as a function of position. The electrostatic free energy of the system can then be computed by the equation:

$$\Delta G_{\text{solv}}^{\text{PB}} = \frac{1}{2}\sum_i q_i\, V_i \tag{3.5}$$

where q_i are the atomic charges and V_i is the PB-derived electrostatic potential at the location of the charges.

Alternatively, the solvation free-energy can be approximated by the "generalized-Born" (GB) approximation, which is based on an analytical function of the biomolecular coordinates first proposed by Still [149]:

$$\Delta G_{\text{solv}}^{\text{GB}} = \frac{1}{2}(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p})\sum_{i,j}\frac{q_i\, q_j}{(r_{ij}^2 + b_i b_j exp[-r_{ij}^2/(4b_i b_j)])^{1/2}} \tag{3.6}$$

In the above equation, $\{q_i\}$ are the atomic charges and ϵ_p, ϵ_w are the protein and water dielectric constants; the parameters $\{b_i\}$ are referred to as "atomic Born radii" and are related to the distance between the charges and the protein-solvent interface.

The diagonal terms ($i = j$) in the GB solvation free energy are " self energies", which express the interaction of each charge i with the solvent polarization potential induced by the same charge:

$$\Delta G_{\text{self,i}}^{\text{GB}} = (\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p})\frac{q_i^2}{2\, b_i} \tag{3.7}$$

For a spherical solute with a charge q_i at the center (or a spherically-symmetric charge density), the GB self energy coincides with the Born solvation energy and the GB radius b_i coincides with the radius of the solute [150].

The non-diagonal ($i \neq j$) terms in Eq. (3.6) are referred to as "GB interaction energies", and express the interaction between a particular charge i and the solvent polarization potential due to a different charge j.

To apply Eq. (3.6), the solvation Born radii of all atoms must be known. In the general case of a single charge inside a solute without spherical symmetry, each radius can be computed from the equation

$$\Delta G_i = \Delta G_{\text{self,i}}^{\text{GB}} \equiv (\frac{1}{\epsilon_w} - 1)\frac{q_i^2}{2\, b_i} \Longrightarrow b_i = \frac{1}{2}(\frac{1}{\epsilon_w} - 1)\frac{q_i^2}{\Delta G_i} \tag{3.8}$$

The quantity ΔG_i on the right-hand side of the above equation is the solvation free energy of a biomolecular system with a single charge q_i at position i and can be com-

puted in the PB approximation. To avoid this impractical approach, the Born radii are computed with two approximations:

i) in the first approximation, the electrostatic field due to the embedded charge in the dielectric cavity of the solute is assumed to be radial ($\propto 1/r^2$), even in the absence of spherical symmetry; this is the "Coulomb-field approximation" [151–153]. In this case, the electrostatic-field density $\propto 1/r^4$ and the Born radius can be expressed as the following integral over the volume of the solution [61]:

$$b_i^{-1} = \alpha_i^{-1} - \frac{1}{4\pi} \int_{in, r>\alpha_i} \frac{1}{r^4} \mathrm{dV} \tag{3.9}$$

The integral over the interior space [(in) inside the solute] contains a singularity at $r \to 0$ due to the use of atomic point charges. GB formulations handle this singularity by spreading atomic charges over small spheres of radius a (proportional to the atomic radius) and excluding the integration over the space of these atomic volumes.

ii) The integration over the solute volume is computationally expensive, and needs to be carried out at every energy evaluation step. A second approximation replaces the above integral by a sum of terms, which depend (usually) on the atomic coordinates of pairs of atoms. The functional form of these terms can differ, resulting in a variety of GB variants (e.g. the "Analytical Continuum Electrostatics" or ACE model [154], the "Hawkins-Crammer-Truhlar" or HCT model [90], etc [61]).

An essential property of the employed functional forms (g) is that they are pairwise-decomposable, i.e. they depend on the atomic coordinates of pairs of atoms. In this case:

$$b_i^{-1} = \sum_j g(\vec{r}_i - \vec{r}_j) \equiv \sum_j g_{ij} \tag{3.10}$$

When this is true, the total self-energy of a biomolecule is also pairwise-decomposable:

$$\Delta G_{\mathrm{self}} = (\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p}) \sum_i \frac{q_i^2}{2\, b_i} = \frac{1}{2}(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p}) \sum_i q_i^2 \sum_j g_{ij} \tag{3.11}$$

Eq. (3.10) shows that the radii b_i depend on the coordinates of all atoms. Thus, the GB expression in Eq. (3.6) is not pairwise-decomposable, despite its appearance. This point, along with the electrostatic and solvation terms employed in physical energy functions is discussed in detail in Chapter 4

Since the expression of Eq. (3.2) (combined with a suitable solvation free-energy function such as CASA or GB) is an analytical function of the atomic coordinates, it can provide the energy of a biomolecule, provided the three dimensional structure is known.

3.1.2 Statistical Scoring Functions

Statistical effective energy functions [155–157] use large protein databases of high-resolution crystal structures to extract information such as the relative probabilities of specific side-chain conformations, or the probability density of the distance between specific residues. The probabilities are extracted as observation frequencies and can be converted to free energies via the Boltzmann relation $-k_\beta T \ln[f_{\text{obs}}/f_{\text{obs}}^{\max}]$.

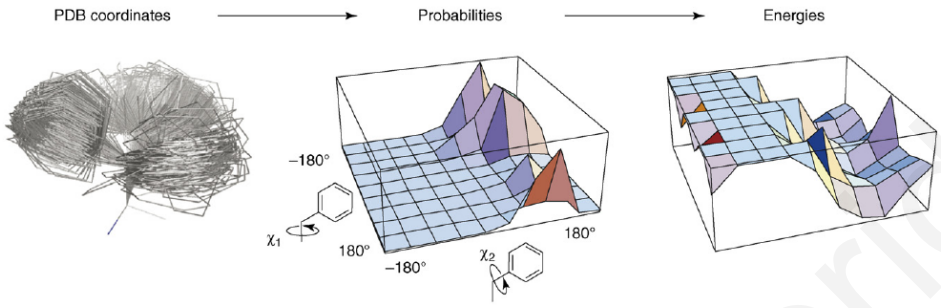


Figure 3.3: Observation frequencies of side chain conformations are derived from the coordinates of the rotamers. The extracted probabilities are converted to energies according to the Boltzmann equation. The figure is taken from Ref. [1].

The computation avoids the complexity of the atomic-detail (or coarse-grained) representation employed in physical energy functions. Each residue is usually defined by two points, mapped onto a three dimensional grid: (1) the position of the C_α atom and (2) the side chain center of mass. All relevant residue - residue distances are computed based on these points.

The characterization of the energy function as "effective" implies that all ignored degrees of freedom (representing the protein and solvent) are integrated out. A typical statistical effective energy function consists of various knowledge-based potential terms, describing short-range interactions of neighbouring residues and long-range interactions of residues more than 3 positions apart. Short-range interactions include energy terms describing hydrogen-bonds, dihedral angles, side chain rotamers and side chain packing terms. Long-range interactions include energy terms of buried residues and residue-residue interactions. Vieth et al. [158] joined the above contributions to a statistical energy function of the form:

$$E_{\text{tot}} = s_1 E_{\text{hb}} + s_2 E_{\text{R14}} + s_3 E_{\text{rot}} + s_4 E_\beta + s_5 E_{\text{one}} + s_6 E_{\text{pair}} \tag{3.12}$$

The model predicts a hydrogen bond between residues **i** and **j** solely on geometric criteria, such as the $C_\alpha - C_\alpha$ distance $|r_{ij}|$ of the two residues and the orientation of the h-bond (E_{hb}). To account for the correct distribution of dihedral angles, the energy function includes a sequence-dependent Ramanchandran potential-energy term (E_{R14}), which favors distributions observed in realistic protein conformations. Each side chain conformation (rotamer) is assocciated to a rotameric energy (E_{rot}). Another

term (E_β) accounts for orientational coupling between neighbouring side chains. A residue is "unburied" if its total number of side chain contacts with other residues is smaller than a specified threshold. An unburied residue is rewarded with a residue-specific energy contribution (E_{one}). Residue-residue interaction energies are expressed by the pair potential (E_{pair}), which contains a repulsive part (to penalize contacts) and a pair-specific potential energy. The individual terms are weighted by scaling factors ($s_1, s_2, s_3, s_4, s_5, s_6$), determined by parameterization of the energy function on a training set of proteins.

By construction, statistical energy functions have built-in the experimental information from known protein structures; on the other hand, these energy functions are unable to predict new effects not present in the training set. Statistical energy functions are residue-based and can be directly employed in CPD applications, but the limited representation of the protein structure makes them less sensitive to conformational/sequence changes, compared to physical energy functions.

3.1.3 Empirical Scoring Functions

Empirical energy functions use simplified approximations of the potential terms in molecular mechanics energy functions, to detect stable sequences during CPD.

Eq. (3.13) shows a typical empirical scoring function developed by Bohm et al. in Ref. [159].

$$\Delta G_{\text{tot}} = \Delta G_{\text{hb}} \sum_{\text{hb}} f(\Delta R, \Delta\alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta\alpha) + \Delta G_{\text{lipo}} |A_{\text{lipo}}| + \Delta G_{\text{solv}} \quad (3.13)$$

Each hydrogen bond is identified by a distance/orientation criterion and contributes to the total energy by an empirically determined parameter. Ionic interaction energies are calculated in a similar way; non-polar contributions are proportional to the contact surface area of the lipophilic parts of the protein.

The free-energy terms contain adjustable parameters, which are calibrated by fitting to experimental results on a selected "training set" of protein structures. The parameter values vary with the training set and are therefore somewhat system-dependent. On the other hand, the use of these simpler expressions accelerates the computation of the scoring function.

In Eq. (3.13) ΔG_{solv} is the contribution of the solvation energy, that is described by a solvent-accessible surface area (SASA) model; that is, a linear function of the solvent-exposed surface, scaled by an empirically determined factor. The surface area is calculated by the Lee and Richards method [75], by rolling a probe sphere along the van der Waals surface of the protein. The probe radius is approximating the radius of a water molecule close to 1.5Å. The surface area is a function of all the atomic

coordinates and must be modified to form a pairwise-decomposable approximation. Street et al. [160] propose a two-body approximation for the pair-wise calculation of the surface area.

$$A_{\text{exposed}}^{\text{pairwise}} = \sum_i A_{i_r t} - s \sum_{i<j} (A_{i_r t} + A_{j_s t} - A_{i_r j_s t}) \tag{3.14}$$

where $A_{i_r t}$ represents the one-body contribution to the total exposed surface area of rotamer r at position i in the presence of the whole template (t). Inside the parenthesis of the second sum, the exposed surface area for a pair of residue positions i, j and rotamers r, s, respectively, in the presence of the template, is subtracted from the exposed surface areas of each side chain intependently. The parenthesis represents the buried surface area between the two side chains. The second sum of residue-residue buried area contributions is subtracted from the total exposed surface are of each individual side chain. The scaling factor s is employed to compensate for overcounting of the buried area, and is adjusted to provide agreement with the total surface area when this is calculated as a function of the entire structure.

3.1.4 The Importance of Residue-Pairwise Additivity of the Scoring Function

For CPD calculations to be efficient, it is essential that the employed scoring function can be expressed in a residue-pairwise form. For example, in the case of a physical energy function, the total potential energy (or free energy) needs to be written in the form

$$G = \sum_R \sum_{i \in R} G(\vec{r}_i) + \sum_{R \neq R'} \sum_{i \in R, j \in R'} G(\vec{r}_i, \vec{r}_j) = \sum_R G_R + \sum_{R \neq R'} G_{RR'} \tag{3.15}$$

The first sum on the right-hand side of Eq. (3.15) contains all terms which depend on individual residues; the second sum contains interactions between residue pairs.

The residue-pairwise form of Eq. (3.15) ensures that the interaction energy between two residues can be calculated without any knowledge of the surrounding environment. This is essential, since during the design of a new sequence/conformation, the environment around a specific residue pair may change. If the energy function is residue-pairwise, then the individual interaction energies between all residues (in all possible chemical types and conformations) can be pre-calculated and tabulated. For example, consider a protein with N residues. Out of these residues, K are considered to be "fixed" in a design calculation (they retain the chemical type and conformation in the original, native protein). This could be due to the fact that the native residues have specific chemical types (e.g. cysteins, prolines or glycines), or due to the fact that a specific chemical identity and conformation are required for a specific function.

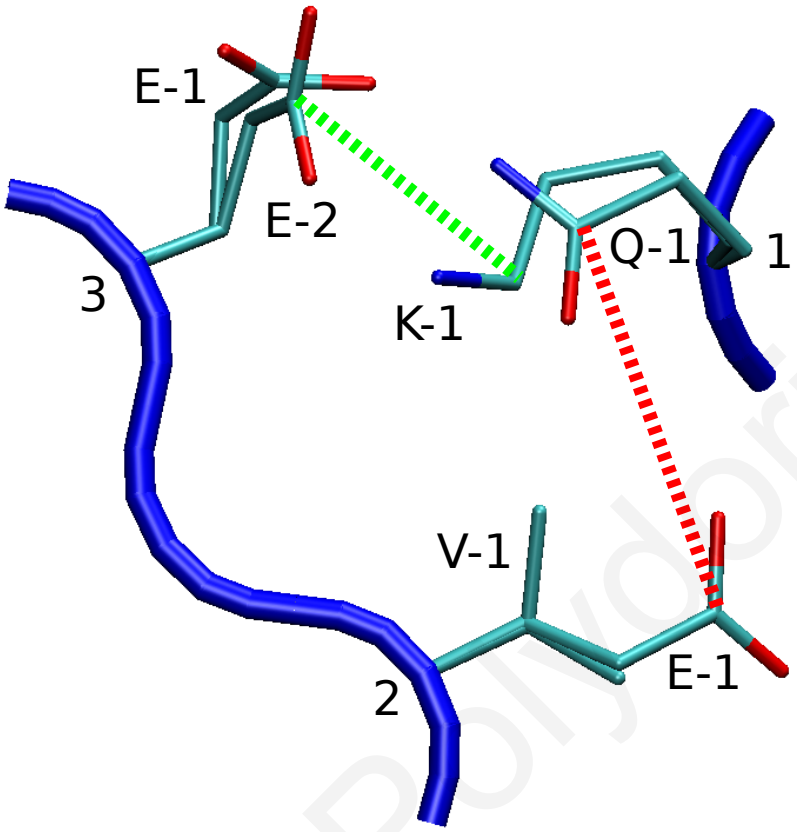
Another M residues are "inactive", i.e. retain their chemical type but are allowed to change side chain orientation; finally, $N - M - K$ residues are "active", i.e. change both chemical type and orientation. If an "inactive" residue i with chemical type $C(i)$ has $P(C(i))$ possible rotamer conformations, and each "active" residue selects L possible chemical types/rotamers, the total number of possible residue states will be $D \equiv K + \sum_{i \in M} P(C(i) + (N - M - k) \times L$. This yields $D \times (D-1)/2$ possible interaction energy terms between different residues, plus D self-terms, which contain the interactions between atoms within each residue (these self-terms can also absorb the interactions between the residue and the fixed part of the system). The $D + D \times (D-1)/2$ possible interaction energy terms can be stored in a triangular matrix (Fig. 3.4).

The computation of such interaction energy matrices is very demanding computationally, but is performed only once prior to the actual design; furthermore, it is trivially parallelizable (i.e. it can be split into a large number of processors). The advantage of this precalculation is that the total energy of a particular sequence/conformation can be quickly assembled from this matrix, by adding together the appropriate elements, in the spirit of Eq. (3.15). During the design, specific parts of the sequence/conformation (e.g. individual residues) are modified; the total energy after the modification can be updated rapidly, by correcting only the matrix entries that correspond to interactions of the altered residue.

The question is then, in which case an energy function such as the one described by Eq. (3.2) can be expressed in the form of Eq. (3.15). Bonded energy terms of Eq. (3.2) depend on covalently-bonded groups of atoms and can be assigned to individual residues, with a suitable partitioning scheme; e.g. bond energies are assigned to a particular residue R if both atoms of the covalent bond belong to R, otherwise they are partitioned equally to adjacent residues. Coulombic and vw non-bonded energies depend on the coordinates of pairs of atoms; thus, these terms can always be assigned to specific residues or pairs of residues. Surface-area contributions to non-polar free energies (accessible surface area terms) can also be assigned to individual atoms or pairs of atoms, under a pairwise approximation [e.g. see Eq. (4.3)].

On the other hand, the GB energy is not pairwise-decomposable, because the interaction between two atoms i and j cannot be expressed solely as a function of the corresponding coordinates $\vec{r}_i$ and $\vec{r}_j$. In Eq. (3.6), the atomic Born radii depend on all atomic coordinates, as discussed in Section 3.1.1.

The energies of all possible side chain - side chain and side chain - backbone interactions are stored (Fig. 3.4(b)) and fed to the algorithm, which can rapidly gather a set of low energy amino acid / rotamer combinations (local minima).



(a) Sidechain - sidechain interactions

Position		1		2		3	
AA-Rotamer		Q-1	K-1	V-1	E-1	E-1	E-2
1	Q-1						
	K-1						
2	V-1						
	E-1						
3	E-1						
	E-2						

(b) Energy matrix

Figure 3.4: (a) All possible sidechain - sidechain and sidechain - backbone interactions are computed. (b) The computed interaction energies are stored in a matrix form and used by the searching algorithm.

3.2 Searching Protocols

The main obstacle faced by CPD calculations is the enormous sequence/conformation space, that prohibits an exhaustive search even for small proteins. In order to overcome this difficulty, CPD protocols usually assume that the protein backbone is fixed into a certain fold, and that side chains can take conformations from a small number of discrete, energetically preferred rotamers. Even with these approximations, the remaining sequence/conformation space is vast. An exhaustive conformational search of the entire space is not possible; to deal with this problem, CPD calculations use smart searching procedures, which allow the consideration of a small subset of all possible conformations.

Searching algorithms are classified into two main categories (deterministic and heuristic), although a combination of approaches can also be successfully applied. These categories are further analyzed in the next sections.

3.2.1 Deterministic Algorithms

Deterministic protocols perform semi-exhaustive explorations of the space and identify the optimum solution (e.g. the global minimum of the free-energy). The Dead-End Elimination (DEE) algorithm [161] performs a systematic deletion of high-energy regions in the sequence/conformation space, including unfavorable rotamers, until a single solution is left. The Self-Consistent Mean Field (SCMF) algorithm [162] is also deterministic, in that it converges to the same solution for a given set of running parameters (without a guarantee that the obtained solution is the global minimum). In what follows, we present in more detail these algorithms.

Self-Consistent Mean Field Optimization Methods

SCMF protocols optimize the side chain conformations for a given main chain fold; The total energy E_{i_r} of a side chain at position i and rotamer r is given by the following expression:

$$E(i_r) = E(i_r, i_r) + E(i_r, bb) + \sum_{j \neq i} E(i_r, j_l) \tag{3.16}$$

The first term on the right-hand side is the self-energy of the side chain; the next term is the interaction with the fixed backbone; the final term is the total interaction with all other side chains.

The possible side chain conformations are described by a global conformation matrix C. The elements $C(i,\ r)$ of this matrix correspond to the Boltzmann probability of residue i to adopt rotamer r:

$$p(i_r) = \frac{e^{-\beta E_{MF}(i_r)}}{\sum_{r=1}^{K_i} e^{-\beta E_{MF}(i_r)}} \tag{3.17}$$

The total energy of rotamer r at position i in the mean field approximation is a sum of the side chain self interaction, the interaction with the backbone, and an average (mean-field) interaction, exerted by all other side chains. The mean-field approximation [162] weighs all pairwise side chain interactions $E(i_r, j_l)$ of Eq. (3.16) by the Boltzmann probability of the rotamer l at residue j:

$$E_{MF}(i_r) = E(i_r, i_r) + E(i_r, bb) + \sum_{j \neq i} \sum_{l=1}^{K_j} E(i_r, j_l) p(j_l) \tag{3.18}$$

The resulting energy landscape is smoothed-out, assisting the algorithm to avoid getting trapped into local minima and facilitating the fast convergence to the minimum of the mean field energy. The optimum solution is determined iteratively: Initially, all rotamers of residue i are assigned the same probability $C(i, r) = 1/K_i$ (with K_i the total number of rotamers consistent with the chemical type of the side chain at position i). For each side chain and each rotamer, the mean-field energy is calculated by Eq. (3.18) and used to compute the probability with Eq. (3.17). In the next iteration, the algorithm calculates the mean field energies using the probabilities from the previous cycle. The procedure is repeated until the change in the obtained probabilities/energies between subsequent cycles is smaller than a specificed threshold. The optimum solution predicts for each residue the rotamer of maximum probability (equivalently, minimum energy).

The efficient implementation of the SCMF algorithm requires the use of a threshold energy, applied to cut off unfavorable rotamer interactions when calculating energies from Eq. (3.18), and an updating scheme for the probabilities. The updated probabilities are usually computed by the formula

$$p^{\text{update}}(i_r) = \lambda\, p^{\text{new}}(i_r) + (1 - \lambda)\, p^{\text{old}}(i_r) \tag{3.19}$$

which uses a linear combination of the probabilities from the two most recent cycles. The parameter λ is chosen so as to accelerate convergence. Although the SCMF protocol is not guaranteed to identify the global minimum, it will always end up to the same solution for the same running parameters. The main advantage of the algorithm is the linear dependence of the computational cost on the size of the system (number of side chains).

Dead-End Elimination

The Dead-End Elimination (DEE) algorithm [161] performs a systematic exploration of the sequence space, and identifies the global minimum (provided that it converges). The

algorithm functions only with a pairwise decomposable energy function. It identifies and excludes from further consideration high-energy ("dead-end") rotamers; in this way, the exchaustive exploration is progressively restricted to a smaller space, until the global minimum is eventually identified.

The DEE method uses a precalculated interaction energy matrix, whose elements contain the pairwise interaction between pairs of rotamers k, l at residue positions i, j. The energy of a protein with a given main-chain fold and N side chains in a conformation $\{1_{i_1},, 2_{i_2}, \ldots, N_{i_N}\}$ is given by the expression:

$$E = \sum_{i=1}^{N} E(i_r) + \sum_{i<j}^{N} E(i_r, j_l) \tag{3.20}$$

In the above expression, $E(i_r)$ is the interaction energy of rotamer r at position i with itself and the (fixed) backbone; $E(i_r, j_l)$ is the interaction energy between the two rotamers r and l at positions i and j, respectively.

The basic idea of the DEE algorithm is to identify those paths that lead to a "dead end" during exploration of the space, and exclude them from further consideration. To identify "dead-end" paths, the algorithm compares the energy of two rotamers r and l of a given residue i; if rotamer r has a higher energy, it is never going to be part of the global minimum. This is considered as a "dead-end", and all side chain combinations including r are rejected. This filter is implemented in a computation by the following condition:

$$E(i_r) + \sum_{i \neq j} \min_{l} E(i_r, j_l) > E(i_m) + \sum_{i \neq j} \max_{l} E(i_m, j_l) \tag{3.21}$$

When residue i is kept fixed at rotamer r, while every other side chain adopts the rotamer of minimum interaction energy with i_r, the energy of the protein is given by the left hand side of the inequality. On the right hand side, residue i is kept fixed at rotamer m, while every other side chain adopts the rotamer of maximum interaction energy with i_m. If the energy $E(i_r, \min(j))$ is higher than $E(i_m, \max(j))$, a protein state with rotamer r at residue position i will never be part of the global minimum solution, since it can always be replaced by the energetically favored rotamer m. The DEE procedure is performed on all residues in turn, and the cycle is repeated until no more dead-end rotamers can be identified. The algorithm converges to the global minimum of the protein conformational space, or to a limited space which can be explored exhaustively by other methods [163]. The method confirms that the converged solution is the optimal solution of the energy landscape.

The above procedure describes the simple DEE criterion of Fig. 3.5, which compares a pair of rotamers (g_α, h_α) of the same residue. An improved criterion minimizes the difference $[E(i_r, j_l) - E(i_m, j_l)]$, by employing a common side chain environment for both rotamers r, m. This criterion [164] completes the simple criterion by identifying

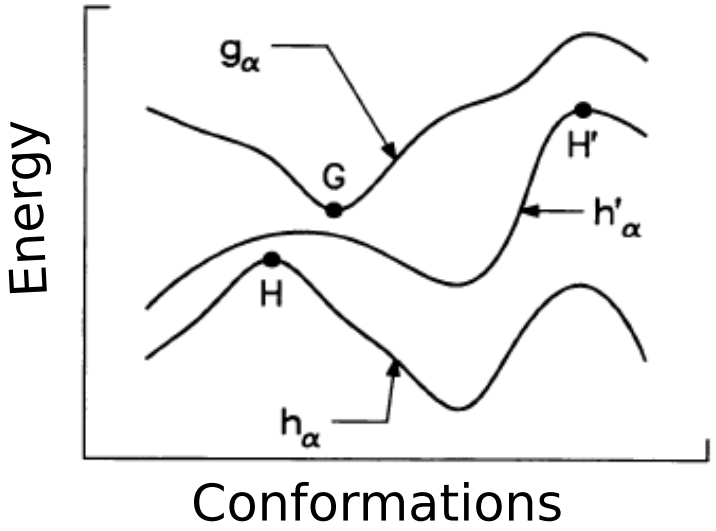


Figure 3.5: Energy profiles of protein conformations with three fixed rotamers of residue α, $g_\alpha, h'_\alpha, h_\alpha$. The lowest energy point of the conformation with fixed g_α is noted with G, while H and H' indicate the energy maxima of conformations h_α and h'_α respectively. Comparing G and H according to their difference G - H < 0 the algorithm identifies g_α as a dead end using Eq. (3.21). This simple criterion fails for cases like G and H' where G - H > 0. The solution is given by employing the limited criterion of Eq. (3.22). The figure is taken from Ref. [161].

dead-end paths with $\min[E(i_r, j_l)] - max[E(i_r, j_l)] < 0$ as shown in Fig. 3.5.

$$E(i_r) - E(i_l) + \sum_{i \neq j} \min_m [E(i_r, j_m) - E(i_l, j_m)] > 0 \tag{3.22}$$

The limited criterion can be expanded from a single residue of a pair of rotamers to comparisons of pairs or higher order clusters of residues.

3.2.2 Heuristic Algorithms

Heuristic methods rely on a random exploration of the sequence/conformation space, which reduces the computational requirements but does not guarantee that the optimum solution will be found (e.g. the coformation/sequence of globally minimum folding free energy); furthermore, the solutions identified by a heuristic algorithm depend on the initial conditions and do not always converge to the same solution. This is not a serious concern: repeating the search from a large number of different initial conditions, the identified solutions can be successively improved. Monte Carlo (MC), Genetic Algorithms (GA) and the Wernisch protocol belong to this category of searching algorithms.

Monte Carlo Methods

The main heuristic algorithm is the Monte Carlo (MC) method, which performs a random walk in the sequence/conformation space. During a MC exploration, a position is

chosen randomly and is filled with a randomly chosen chemical type (from a list of types, compatible with the position) and a randomly chosen side chain rotamer (compatible with the chemical type). The energy (or scoring function) of the resulting trial mutant is evaluated and is accepted or rejected, according to a Metropolis criterion. According to this criterion, the trial mutant is accepted if it is favored energetically compared to the sequence/conformation prior to the change ($\Delta E = E_{\text{trial}} - E_{\text{last-accepted}} < 0$); otherwise, the mutant is accepted only if the corresponding Boltzmann probability is greater than a randomly generated number **p** between 0 and 1 [$p < \exp(-\beta \Delta E)$]. The procedure is repeated until no new trial mutants with favorable energies are produced and the method converges. Although the method is fast enough, it is uncertain if the resulting mutant sequence after convergence corresponds to the global minimum, or to a local minimum. During the searching procedure the temperature T ($T = 1/k_B\beta$) is toggled between high and low values to overcome high energy barriers of the landscape.

Since the method is stochastic, the solution relies on the starting point (e.g. sequence/conformation, and seed of the employed random-number generator). Nevertheless, the computational speed of the algorithm allows running a large number of cycles starting from different initial points.

Genetic Algorithms

Genetic Algorithms (GA) use genetically constructed operators to refine a group of random sequences representing the protein under design. These methods follow an iterative sheme which ranks sequences and selects the best to continue onto the next stage. The procedure is repeated until the algorithm finds the optimum solution.

A GA [165] can be used to generate sequences that best fit a given protein fold. A protein sequence S of length m is defined by a series of m constituent residues, each one assigned with an amino acid chemical type chosen from a specific group of chemical types (e.g. the twenty natural amino acids $S : R_1, R_2, ..., R_m \, R_i \in 20$ aa). The method starts by generating a population of n random sequences, corresponding to many individual solutions S^n. The size of the population depends on the protein size (m), but is usually set to a very large number, to allow for adequate sampling of the space. Sometimes the generated population of sequences are not entirely randomized but follow predefined motifs, which constrain the exploration to neighborhoods of optimum solutions and speed up the procedure. The starting population is then subjected to three genetic algorithm operators: (a) selection, (b) crossover and (c) mutation. These operators undertake the generation of a new population from the previous one by an iterative scheme.

The crossover operator acts on the previously selected sequences, creating for each individual member a number of "parent" sequences (usually two), depending on the method. Objective is to produce "children" sequences, which share many of the characteristics of the parents. In other words, the crossover operator generates sequences

of the same amino acid composition by a recombination procedure. Such methods are the two-point crossover, cut and splice, etc.

A mutation event simply replaces the amino acid type of a given side chain with one of the remaining alternatives (unless constrained otherwise), to preserve diversity in the population and prevent the production of too similar sequences. At each generation step a number of mutations are performed at different positions k of the sequence, selected by the following formula:

$$k_{new} = 1 + k_{old}\, mod(mn) + \frac{ln(r)}{ln(1-p)} \tag{3.23}$$

where r is a random number (uniformly distributed) between 0 and 1 ($0 < r < 1$) and p the mutation probability, that is given as an input parameter at the beginning of the generation procedure. The modulo function in the equation returns the integer of the remainder of the divition $\frac{mn}{k}$. The generation process is repeated until it converges to an optimum solution, without neccessarilly being the global minimum.

At each successive generation, a group of "high scoring" sequences is selected for the production stage. The selection criterion is the propensity of the sequence towards an optimum solution and is measured with an objective function characteristic of the problem, similar to the scoring energy function discussed in earlier sections. The idea is to select more frequently the most favored sequences. This selection dependence is implemented by a variety of methods. Two of the most commonly used are the "rulette wheel" [166] and the "tournament" [167]. One advantage of GA is that any sequence modification performed by the genetic operators can overcome possible restrictions produced by high energy barriers of the energy landscape.

The Wernisch Protocol

Wernish and coworkers proposed a simplified heuristic method for searching the sequence / rotamer conformational space. The algorithm minimizes the folding free energy [Eq. (7.3)] of a protein sequence, by minimizing each residue position independently, given that the rest of the protein is kept fixed at a random rotamer combination. The Wernisch algorithm is described in detail in Section 7.1.3. Its advantage is that it can locate local minima very rapidly. For problems of moderate size, the Wernisch algorithm identifies the global minumum very fast, compared to deterministic methods (DEE, B&B [73]).

3.3 Examples of High-Throughput CPD Applications

In what follows, we present briefly some important applications of CPD calculations.

3.3.1 Stability Calculations

Stability calculations search for sequences that minimize the folding free energy of a particular known fold. The designed sequences can be highly homologous to the original sequence (since high-similarity sequences usually share the same fold), but may also be sufficiently different. For example, Mayo and coworkers redesigned the entire sequence of Drosophila melanogaster engrailed homeodomain (51 residues) and produced two high-stability sequences with midpoints of thermal denaturation at 99°C (50°C higher than the wild-type denaturation temperature). These sequences shared less than 25% identity with the native sequence [17]; Baker and coworkers redesigned the entire sequence of the activation domain of human pyrocarboxypeptidase A2, identifying a sequence that was more stable by 10 kcal/mol and had 22% similarity with the native protein [168].

CPD have also been used to identify sequences that stabilize novel protein folds. Calhoun and coworkers designed a monomeric four-helix bundle fold for due-ferro (DF) metalloproteins, that had increased stability with respect to the native tetrameric or dimeric four-helix bundle fold [20]. Kuhlman and Ambroggio optimized the aminoacid sequence of the peptide Sw2 (32 residues) to adopt a zinc finger fold in the presence of a zinc atom and a trimeric coiled-coil fold in its absence. Their algorithm produced compatible sequences to the backbone parts of a trimeric coiled coil from hemagglutin and the zing finger-DNA, identifying those able to switch between these two distinct protein folds in the presence or absence of a metal [16].

3.3.2 Affinity Calculations

Other CPD examples redesign a protein binding site, aiming to increase the affinity for a specific substrate. The scoring functions employed in such studies need to be able to reliably identify binding free energy differences on the order of a few kcal/mol. Handel and colleagues used affinity CPD methods to produce potent inhibitors of the enzyme SHV-1 $\beta-$lactamase. They obtained two sequences of the inhibitor protein BLIP with increased affinities, compared to the native; the predicted affinities were within 1 kcal/mol of the corresponding experimental values [36]. Sammond and coworkers used a combination of criteria to select optimum sequences of the ubiquitin conjucating enzyme (UbcH7) in complex with the associated protein (E6AP). They performed point mutations of residues not participating in the hydrogen-bond network of the protein - protein interface. Mutants with favorable binding free energy were also tested for distabilizing the protein structure [169].

Another work [52] increased by CPD methods the stability of the D44.1 antibody-lysozyme antigen complex. The design was initially guided by the optimization of electrostatic interactions. Low-energy sequences were then filtered by more precise energy calculations (in the Poisson - Boltzman approximation); the combined proto-

col identified optimum sequences/structures with 140-fold improved affinities. Looger and coworkers engineered the binding pocket of protein receptors which bind trinitrotoluene, L-lactate or serotonin [21] aiming better affinity and specificity. Their procedure replaced the ribose or amino acid ligands from proteins of the *Escherichia coli* periplasmic binding protein (PBP) superfamily with one of the three target ligands. The design method, mutated amino acid residues in direct contact with the native ligand, allowing the target ligand to adopt 10^8 different conformations. All designed protein receptors exhibited a detectable binding affinity to their target ligands without at the same time responding to the native ligand. Such examples of engineering the active site of the protein for binding a target ligand with little resemblance to the cognate ligand, result in modified binding pocket with switched specificity.

3.3.3 Specificity Calculations

Furthermore, CPD calculations can be used to enhance the relative binding free-energy of a specific ligand or protein. This can be achieved either by simply maximizing the stability of the target state (positive design) or by following a negative design approach aiming to destabilize alternative states. Positive and negative design were tested by Sauer and co-workers onto the SspB adaptor protein [40]. Positive design produced both heterodimer and homodimer mutant sequences with increased stability, compared to the native homodimer complex. Negative design employed by the free energy difference between the complex and the isolated states, predicted exclusively heterodimer mutants. The different mutations introduced in each monomer, favor the comlpex formation by destabilizing each isolated protein.

A negative design approach can also be employed to create sequences with altered specificity. Ashworth and coworkers used a negative design to engineer the specificity of the intron-encoded homing endonuclease (I-MsoI) recognizing DNA and cleaving selectively long target sites of base pairs[41]. To achieve an altered cleavage specificity of the protein, they replaced certain base pairs of the DNA chain and designed endonuclease sequences with increased relative affinity against the native specificity.

3.3.4 Design of Novel Functions

The design of novel enzymes with new desirable catalytic activities is of great interest. Apart from naturally occuring enzymes, biotechnology and biomedicine fields are in need of novel biocatalysts. CPD can be used to predict new enzymes to catalyze synthetic reactions. Rothlisberger et al. designed a new enzyme catalyst for the Kemp elimination; the transfer of a proton from a carbon [29]. They proposed two active sites to catalyze the reaction and saught backbone scaffolds to accommodate the specific arrangement of catalytic motif. Subsequently, they redesigned the residues surrounding

the catalytic site for increase stability and affinity. The novel designed enzymes exhibited increased catalytic activity.

Chapter 4

Treatment of Electrostatic Interactions in CPD

Electrostatic interactions play a central role in the structural stabilization of proteins [64–67] and the determination of protein function and protein specificity [96; 139–143; 170]. They can also play a key role in catalysis, by stabilizing a preorganized conformation of an enzyme active site that is optimized for the transition state of its substrate [171].

Electrostatic interactions between polar or charged residues may accelerate the formation of biomolecular complexes (through "electrostatic steering" [172]). Networks of electrostatic interactions may yield a net stabilization of protein-protein complexes, by compensating for the free-energy penalty due to the burial of charged residues upon the complex formation.

Electrostatic interactions may also be important for protein-ligand association, and even if their overall effect may not govern binding, mutants with favorable electrostatic interactions improve binding affinity [3]. An example is the protein aspartyl-tRNA synthetase (AspRS), which is highly specific for the negatively charged aminoacid aspartic acid (Asp). This protein is described in more detail in Section 8.3.2. The binding affinity of Asp to AspRS is dictated by a large network of conserved electrostatic interactions between the Asp side chain and proximal charged or polar residues. This is shown in Fig. 8.18 of Section 8.3.2. These networks stabilize the AspRS:Asp complex and favor Asp with respect to other ligands [142]. Overall, the AspRS specificity for Asp is achieved by a group of residue - residue and residue - ligand interactions, the protonation state of active side histidines, the presence of magnesium ions attached to the ligand and the structural shift of a flipping loop [140]. These multiple states of the system and the accurate treatment of electrostatic interactions are essential for the succesful application of CPD calculations, which aim to modify the stability or the function of a protein or a protein complex.

Even though the accurate treatment of electrostatic interactions is essential for ac-

curate computational studies of protein stability, function and design, it is far from trivial. Proteins are macromolecules that function usually inside aqueous solutions (or in membrane environments, which are in the vicinity of an aqueous medium). A protein and its surrounding environment constitutes a complex, highly inhomogeneous dielectric medium, which modifies in a non-trivial manner the intra- and intermolecular electrostatic interactions. Furthermore, during an association reaction (e.g. the formation of a protein-protein or protein-protein complex), the environment in the vicinity of the protein binding site changes significantly: a part of the highly polar aqueous solvent is replaced by chemical groups, which could be non-polar, polar or charged. This change in environment may introduce a structural reorganization in the components of the biomolecular complex, which contribute further to the changes in electrostatic interactions.

Ideally, the most accurate treatment of interactions is achieved by atomic-detail models, which represent explicitly all atoms of a biomolecule and its surrounding environment, including the aqueous solvent. In practice, the explicit modeling of water results in an enormous increase of the conformational space. To deal with this problem, computational studies often use an implicit representation of the solvent, via the incorporation in the energy function of free-energy terms that describe the influence on intramolecular interactions due to the surrounding solvent. In what follows, we outline the major implicit solvent models used in computational studies.

4.1 Implicit-Solvent Models

The effect of solvent on a system (e.g. a biomolecule or biomolecular complex) is expressed mathematically via the solvation free energy of the system, i.e. the change in free energy due to the transfer of the system from vacuum to solution. From the point of view of statistical mechanics, the solvation free energy of a biomolecule in a fixed conformation can be determined from the partition function of the total system, after integrating out the solvent degrees of freedom. Implicit-solvent models represent solvent effects by introducing terms in the Hamiltonian of the system, which describe the dependence on the solvation free energy in terms of the biomolecular coordinates.

Conceptually, most implicit solvent models decompose the solvation process into three sequential steps [95]: i) Creation of a cavity in solution to accommodate the biomolecule; ii) Switching-on dispersion interactions between the biomolecule and surrounding medium, while all atomic charges are set to zero; iii) Switching-on the biomolecular charges. The solvation free-energy is then given by the following expression:

$$\Delta G_{\text{solv}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdW}} + \Delta G_{\text{elec}} \tag{4.1}$$

The first two terms are usually assumed to depend linearly on the solvent-accessible surface area of the biomolecule, even though the validity of this approximation has been questioned for step ii) [173]. In Eq. (4.1) we assume that the solute's surface area is maintained during transfering from vacuum to solution. With a positive coefficient of proportionality, the increase in the solvent-accessible surface area is associated with an increase in solvation free-energy; thus, this term reflects the fact that the free energy increases with the exposure to solvent, and accounts partly for the tendency of non-polar residues to be solvent-excluded. Step iii) describes the change in free energy due to interactions of the molecular charges with the surrounding solvent.

Two major classes of implicit-solvent models are used in MD simulations; i) phenomenological models, which assume that the solvation free-energy can be expressed as a function of simple geometric properties of the solute, such as the accessible surface area in the ASA model [75]; ii) Continuum-electrostatic models such as the Poisson-Boltzmann (PB) [74] approximation and the generalized Born (GB) [61] approximation, which treat the solute as a low-dielectric cavity embedded in a high dielectric medium, and compute the solvation free energy either numerically or analytically. The PB model has been extensively used to calculate small-molecule and protein solvation energies, free-energies of biomolecular complex formation, as well as in ionization (pK) calculations. Its use in MD simulations is less frequent, due to the computational cost associated with calculating the corresponding forces. The GB model is more easily employed in MD simulations, because its energy is expressed as analytic and differentiable functions of the atomic coordinates.

High-throughput CPD calculations are based on efficient and accurate scoring functions, which allow the calculation and tabulation of residue-pair interaction energies prior to the design, as explained in Section 3.1. Thus, CPD calculations require the use of residue-pairwise energy functions. On the other hand, PB and GB solvation free-energies are many-body quantities, which depend on the solvent-solute dielectric boundary (i.e. on the atomic coordinates of all solute atoms). Thus, in order for PB or GB approximations to be applicable in CPD calculations, they have to be written in a pairwise-decomposable form, compatible with the optimization algorithms (DEE) employed to search the combinatorial problem. Pairwise-decomposable implicit solvent models are discussed in detail in the following paragraphs.

4.1.1 Empirical Models

The Coulomb Accessible Surface Area (CASA) Approximation

The simplest implicit-solvent (Coulomb Accessible Surface Area or CASA) models express the total solvation free energy as a sum of two terms: the first term is a screened Coulombic energy, which assumes that the protein/solvent medium can be represented by a common dielectric constant ϵ. The second term is a sum over atomic

contributions, which are proportional to the solvent-exposed surface-areas (SASA) A_i of individual atoms [94; 146; 147]:

$$E_{\text{solv}}^{\text{CASA}} = (\frac{1}{\epsilon} - 1) \sum_{i<j} \frac{q_i \, q_j}{r_{ij}} + \sum_i \sigma_i \, A_i \tag{4.2}$$

The proportionality coefficients σ_i (measured in kcal/mol/Å^2) depend on the nature of the individual atoms, and express the tendency of individual atoms to be buried or exposed to solvent. They have empirical values obtained from comparison with thermodynamical data and are optimized for a set of applications performed by MD simulations [174]. Simpler applications involve a common solvation parameter for all atoms [76]. The SASA surface areas A_i are usually calculated by the Lee and Richards method [75], using a probe sphere of radius 1.5 Å (approximately the water-molecule radius), that is rolled over the surface of the solute.

The arbitrary geometry of the solute makes the surface area a function of all the atomic coordinates. A residue-pairwise variant of the SASA model is based on a simple approximation by Mayo and coworkers [160]:

$$A_{\text{exposed}}^{\text{pairwise}} = \sum_i A_{i_r t} - s \sum_{i<j} (A_{i_r t} + A_{j_s t} - A_{i_r j_s t}) \tag{4.3}$$

where $A_{i_r t}$ represents the one-body contribution to the total exposed surface area of rotamer r at position i in the presence of the entire backbone (i.e. template (t)), as shown in Fig. 4.1. Similarly $A_{j_s t}$ is the corresponding surface for residue j in rotamer s. $A_{i_r j_s t}$ is the exposed area of the rotamer pair r,s at positions i,j in the presence of the whole backbone. The first term of Eq. (4.3) sums the exposed area of individual residues, buried by the backbone, whereas the second term subtructs the sum of buried area between pairs of residues. The factor s accounts for the overcounting of an overlapping area, when summing over multible residue pairs. This overestimation of the exact solvent exposed area originates from contributions of residues at the core of the protein which is more compact than the surface. A more refined expression of Eq. (4.3) uses s = 0.42 for core residues and s = 0.74 for residues at the surface.

The Distance-Dependent Dielectric (DDD) Approximation

The distance-dependence dielectric (DDD) model employs a distance-dependent dielectric constant in the Coulombic term. A simple functional form, linearly dependent on the interatomic distance ($\epsilon(r) = \epsilon r$) can be used; more advanced methods assign multiple dielectric constants for different groups of atoms in core, boundary and surface residues, based on their distance from the protein-solvent boundary [175]. The assignement takes into account the place and solvent-exposure of each atom, as shown in Fig. 4.2.

The total electrostatic interaction energy between all possible pairs of charges is

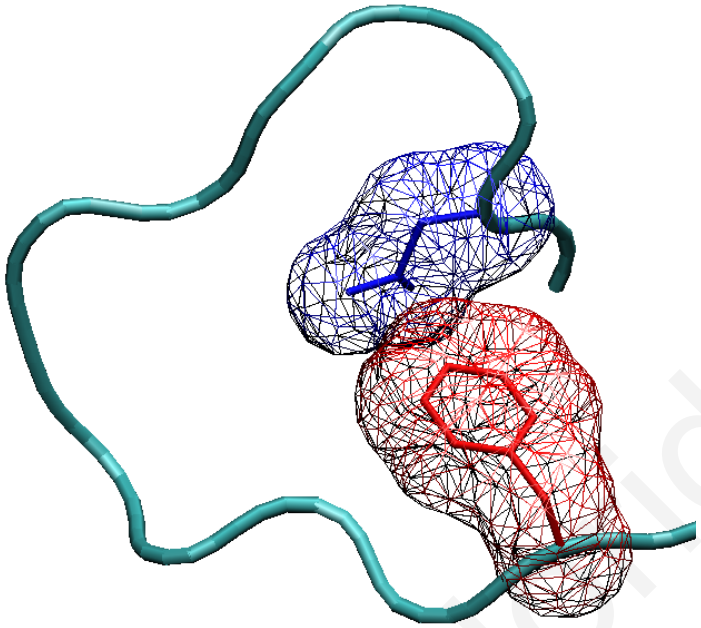


Figure 4.1: Solvent accessible surface area of residue positions i,j shown in blue and red colored wireframes, respectively. $A_{i_r t}$ and $A_{j_s t}$ are the backbone and side chain moieties of the residue pair i,j in rotamer r and s, in the presence of the entire protein backbone t. $A_{i_r j_s t}$ is the exposed area of the rotamer pair r,s at positions i,j in the presence of the whole backbone, shown by the overlaped surface.

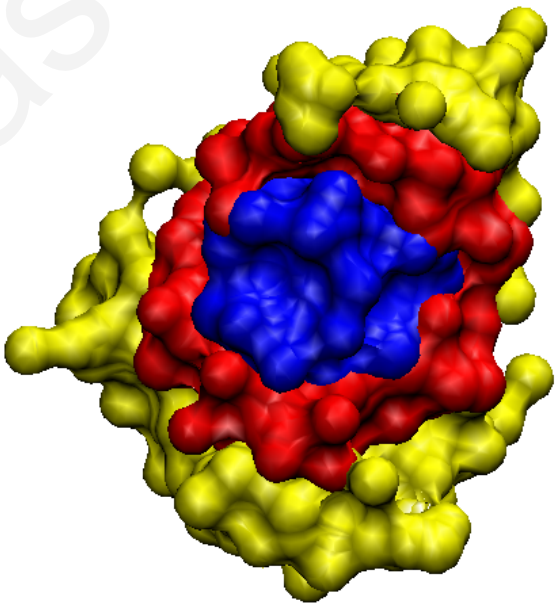


Figure 4.2: Residues are classified as core (blue) boundary (red) and surface (yellow) according to the distance between their C_β atom and the solvent accessible surface.

given by the sum of pairwise interactions:

$$\Delta G_{\text{solv}}^{\text{inte}} = \sum_{i<j} \frac{1}{\epsilon_p(r_{ij})} \frac{q_i q_j}{r_{ij}} \tag{4.4}$$

where the protein dielectric constant $\epsilon_p(r_{ij})$ associated with pair $(i,\ j)$ depends on the location of the two atoms i,j and their mutual distance. A set of different values are employed for interactions between atoms from different regions. An additional Born self-energy term (missing in the CASA approximation) takes into account the interaction between atomic charges with the solvent reaction field they induce:

$$\Delta G_{\text{solv}}^{\text{self}} = \frac{1}{2}\Big(\frac{1}{\epsilon_p(r_i)} - \frac{1}{\epsilon_w}\Big) \sum_i \frac{q_i^2}{r_i^{\text{Born}}} \tag{4.5}$$

where q_i is the charge of atom i, R_i^{Born} is its Born radius, and $\epsilon_p(r_i)$ is a region-dependent dielectric constant. Optimal values of the protein dielectric constants used for core and boundary atoms and their Born radii are empirically obtained, by fitting the model to a large data set of experimentally determined ionization constants. A low dielectric constant ϵ=19 is obtained for the core region and a higher value ϵ=32 for the intermediate boundary region [175]. Atoms in the surface region are considered to be fully exposed to solvent and assigned with a dielectric constant of ϵ_w.

The Lazaridis-Karplus (LK) Model

The Lazaridis and Karplus (LK) model [176] expresses the total solvation free energy of a particular protein conformation as a sum over contributions from individual groups of atoms, as shown in Eq. (4.6).

$$\begin{aligned} \Delta G^{\text{solv}} &= \sum_i \Delta G_i^{\text{solv}}, \\ \Delta G_i^{\text{solv}} &= \Delta G_i^{\text{ref}} - \sum_j \int_{V_j} d^3r\, f_i(r_{ij}) \end{aligned} \tag{4.6}$$

Each contribution reflects the change in the solvation free-energy due to the transfer of the corresponding group from the unfolded to the folded conformation. This transfer is accompanied by a partial or total replacement of the surrounding high-dielectric solvent by the less polar solute medium, a change in the solvent orientation around the solute and a modificiation in the solute-solvent interactions. The solvation energy of a fully solvent exposed group i is given by an empirically determined reference value ΔG_i^{ref}. A group i inside the solute is screened from solvent by the surrounding groups, each contributing to a reduction in the solvation energy of group i. This reduction is expressed by the integral over the volume of group i of a suitably defined energy-density function $f_i(r_{ij})$. In the LK model, the function f_i depends on the distance r_{ij} between

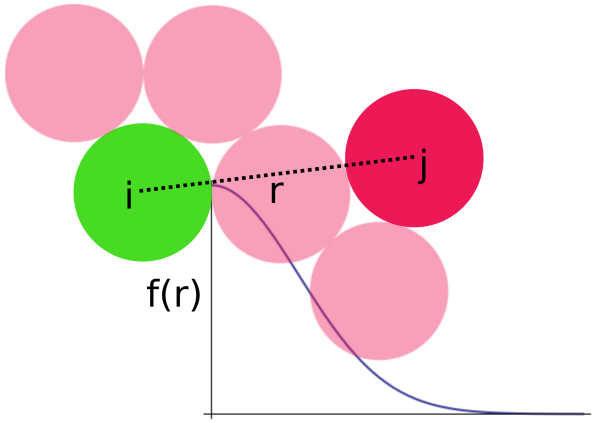


Figure 4.3: The gaussian free energy density of group i in the LK model [176]. The variable r_{ij} corresponds to the distance between i and any surrounding group j.

the group i and the surrounding solute groups j and is approximated by a Gaussian (Fig. 4.3).

The LK model approximates each integral of Eq. (4.6) as the product of the group volume V_j and the solvation energy density of group i, $f_i(r_{ij}) = \alpha \exp[-(\frac{r_{ij}-R_i}{\lambda_i})^2]$, where R_i is the vdW radius of group i, λ_i a correlation length and α_i a scaling coefficient. Eq. (4.6) can be written as:

$$\Delta G_i^{\text{solv}} = \Delta G_i^{\text{ref}} - \sum_{j \neq i} f_i(r_{ij}) V_j \tag{4.7}$$

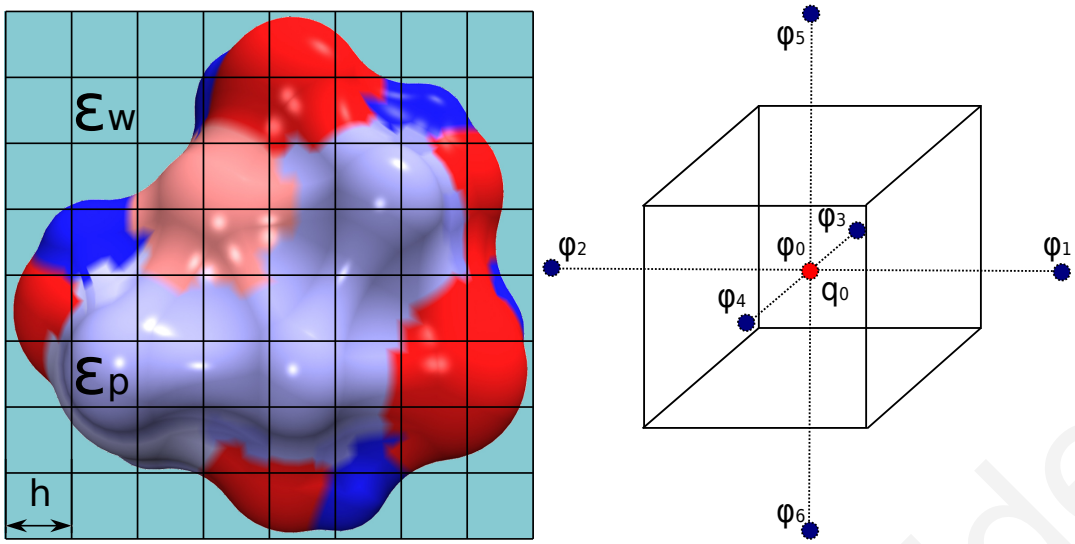
An advantage of the LK model is that the derivatives of the solvation energy can be obtained analytically, enabling its straightforward implementation in MD simulations. The LK solvation model is also pairwise-decomposable and can be included in scoring functions, employed in protein design calculations.

4.1.2 Models Based on the Continuum - Electrostatics Approximation

Continuum electrostatic models treat the solute as a low dielectric cavity embedded in a high dielectric medium. The solute charge distribution is described by a charge density function $\rho(r)$, which in the simplest and most common approximation is centered around the individual atoms: $\rho(r) = \sum_i q_i \delta(r_i - r)$. The electrostatic field obeys the Poisson equation (PE):

$$\nabla[\epsilon(r)E(r)] = 4\pi\rho(r) \tag{4.8}$$

where $\epsilon(r)$ is the position - dependent dielectric constant. Substituting $E = -\nabla\phi$ into Eq. (4.8), the scalar potential ϕ becomes the main variable of the Poisson equation:



(a) Continuum dielectric model (b) FDPB grid cell

Figure 4.4: (a) The continuum dielectric model uses a low dielectric value for the protein cavity $\epsilon_p \approx 2-20$, and a high dielectric value for the solvent $\epsilon_w = 80$. The protein surface is colored by atomic charge. (b) A schematic representation of a unit cell employed by the FDPB method. The nodes are shown with blue color.

$$\nabla[\epsilon(r)\nabla\phi(r)] = -4\pi\rho(r) \tag{4.9}$$

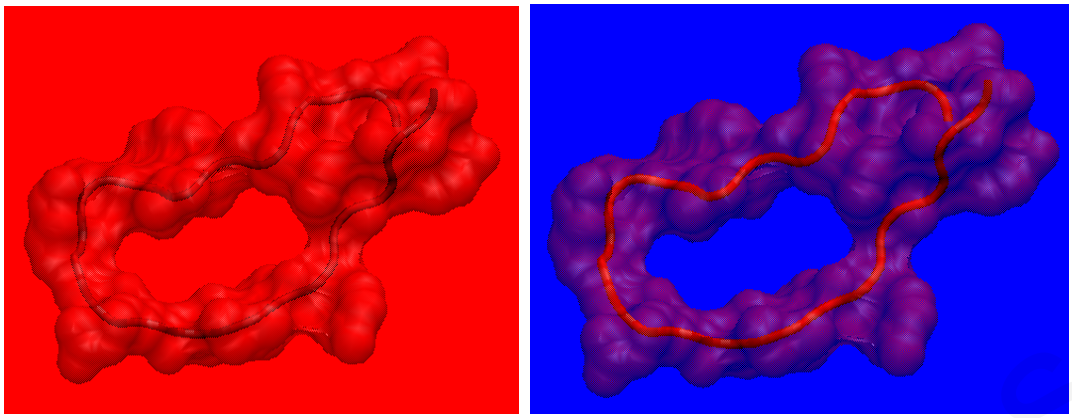
The solute charges polarize the dielectric medium, inducing a reaction field and a polarization charge at surfaces where the dielectric constant is discontinuous. The total electrostatic energy is given by the equation:

$$E^{\text{elec}}_{\text{total}} = \frac{1}{2}\sum_{i,j\in\text{protein}}\frac{q_i q_j}{r_{ij}} + \frac{1}{2}\sum_{i\in\text{protein}} q_i\phi_i^{\text{reaction}} \tag{4.10}$$

Continuum models describe the solute-solvent interactions including both non-polar contributions (creating the cavity) and electrostatics (screened field). The solute surface defines the boundary between the two continuum dielectric media. The Poisson or Poisson - Boltzmann (when solvent ions are included) equation can be solved analytically for simple solute geometries (e.g. a sphere) and numerically for arbitrary shapes like proteins.

In the finite-difference approach (FDPB), the space is discretized into a grid (Fig. 4.4(b)). The ∇ operator is written as the finite difference $\nabla\phi(r) = \frac{\phi_{i+1}-\phi_i}{h}$ and the atomic charges q_i are mapped onto the grid nodes. The electrostatic potential ϕ_i is expressed as a function of the potentials at the six neighbouring nodes ϕ_j with the appropriate dielectric constant ϵ_{ij}:

$$\phi_i = \frac{\sum_{j=1}^{6}\epsilon_{ij}\phi_j + 4\pi q_i/h}{\sum_{j=1}^{6}\epsilon_{ij}} \tag{4.11}$$



(a) High dielectric solvent (b) Low dielectric solvent

Figure 4.5: (a) The space surrounding the protein is filled with the same low dielectric solvent $\epsilon_s = \epsilon_p$. (b) The protein is surrounded by a high dielectric solvent $\epsilon_s >> \epsilon_p$.

The electrostatic free energy of the protein is given by the sum over all nodes of the grid of the products $q_i\phi_i$. The electrostatic potential at node i is the sum of the potentials produced by all the other atomic charges. $\phi_{j\to i}$ is the potential at node i due to the presence of the atomic charge q_j at node j.

$$\begin{aligned} G_{elec} &= \frac{1}{2}\sum_i q_i\phi_i, \\ \phi_i &= \sum_j q_j\phi_{j\to i} \end{aligned} \tag{4.12}$$

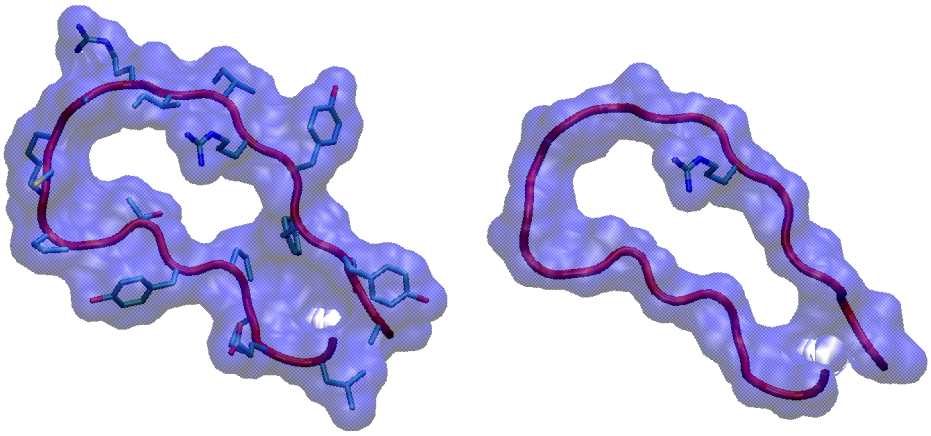
The solvation energy of a protein can by calculated by subtracting the electrostatic free energy of the protein inside the high dielectric solvent, from the electrostatic free energy of the protein when it is immersed in an infinite medium of dielectric constant ϵ_p (Fig. 4.5).

Continuum dielectric models are not pairwise-decomposable, because the solute-solvent boundary surface depends on the position of all atoms; thus, they cannot be directly used in CPD calculations. Efforts have been made to overcome this difficulty and derive a model that can by used in CPD. This approach is discussed in the next section.

The Pairwise Poisson - Boltzmann Approximation of Mayo

The Poisson (or Poisson-Boltzmann) approximation cannot be readily employed as a part of pairwise-factorable scoring functions in CPD calculations, since the boundary surface separating the solute and solvent dielectric regions depends on the atomic coordinates of the entire protein.

A pairwise decomposable FDPB method, developed by Mayo and coworkers [70; 177], expresses the total electrostatic energy of a protein as a sum over contributions



(a) Exact PB (b) One-body PB

Figure 4.6: (a) The exact calculation of the PB equation uses the entire protein to define the dielectric boundary surface (b) The "one-body" approximation employs a system consisting of the protein backbone and a single side chain at a time (in a specific rotamer), to define the low dielectric cavity.

from single side chains or side chain pairs, by assuming a simplified representation of the boundary dielectric surface.

The first order (or "one-body") approximation determines the solvation free energy of a side chain at position i and rotamer k, by assuming that the protein backbone is present, but all other side chains are absent (Fig. 4.6). The PB equation is solved for all side chain chemical types and rotamers compatible with each position i.

Turning off the side chain charges, the solution corresponds to the potential induced at the backbone by the backbone atomic charges $\phi_{bb}^{i,bb}$. Repeating the calculation in the absence of the side chain from the low dielectric cavity produces the backbone potential ϕ_{bb}^{bb} approximating the unfolded state. The solvation energy difference of the two states accounts for the contribution of the given side chain to the desolvation energy of the backbone, as shown by the thermodynamic cycle in Fig. 4.7

The total desolvation of the backbone is approximated by a sum of contributions over all side chains:

$$\Delta G_{\text{desolv}}^{bb} = \sum_i \frac{1}{2} \sum_t^{bb} q_t (\phi_{bb}^{i,bb} - \phi_{bb}^{bb}) \tag{4.13}$$

where t and q_t are, respectively, the backbone atoms and their atomic charges.

By turning off the backbone atomic charges, the method calculates the potential $\phi_i^{i,bb}$ due to the side chain atomic charges at the side chain atoms of residue i (in the presence of the total backbone - superscript [^{bb}]). Upon removing the backbone surface from the low dielectric medium and keeping just the atoms on the backbone fragment of the residue of interest, the calculation yields in the potential ϕ_i^{ib} at the unfolded state of the side chain, where it can interact only with itself and the local backbone. The

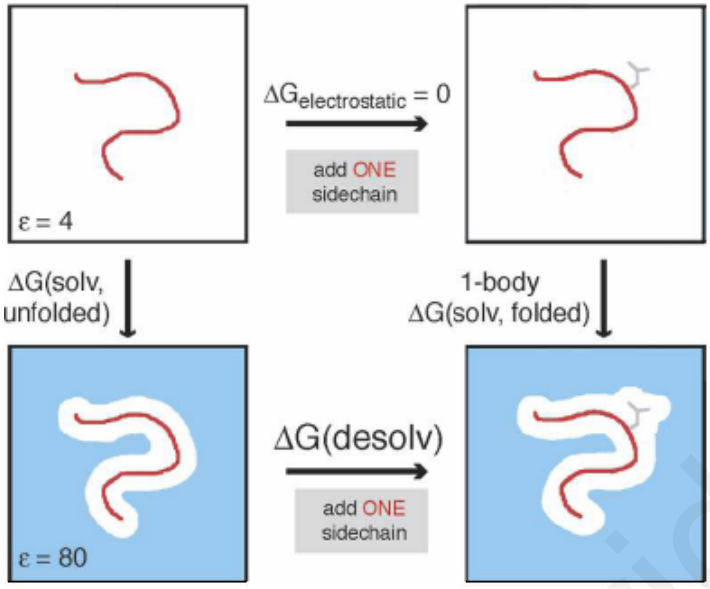


Figure 4.7: The solvation energy difference (vertical steps) of the isolated backbone (unfolded state) and the backbone with the side chain of interest attached (folded state) represents the one body backbone desolvation (bottom horizontal step). The figure is taken from Ref. [70].

solvation energy difference of the side chain i at the folded and unfolded state accounts for the contribution of the backbone to the desolvation of the given side chain:

$$\Delta G^{i}_{\text{desolv}} = \frac{1}{2}\sum_{u}^{sc} q_u(\phi_i^{i,bb} - \phi_i^{ib}) \tag{4.14}$$

where u and q_u are, respectively, the atoms and partial charges of side chain i.

To compute the side chain-backbone electrostatic interactions, the electrostatic potential on backbone atoms, due to the atomic charges at side chain i, is determined. A calculation at a homogeneous dielectric medium ($\epsilon_p = \epsilon_w$) with all charges present yields the backbone-side chain Coulombic interactions. A second calculation in solution ($\epsilon_p \neq \epsilon_w$) accounts for the screening effect due to the surrounding, high-dielectric medium. The total electrostatic energy is given by the equation:

$$\Delta G^{i/bb}_{\text{screenedCou.}} = \sum_{t}^{bb} q_t(\phi_i^{i,bb}(\epsilon_p \neq \epsilon_s) + \phi_i^{i,bb}(\epsilon_p = \epsilon_s)) \tag{4.15}$$

Approximating the solvent accessible surface area of a protein conformation as a sum of contributions from single side chains overestimates the side chain solvation energies, especially for buried residues. To improve the calculation, a second-order ("two-body") approximation is introduced. In this approximation, the protein backbone and two side chains are used to define the boundary dielectric surface of the folded protein.

The "two-body" approximation improves the calculation by including contributions

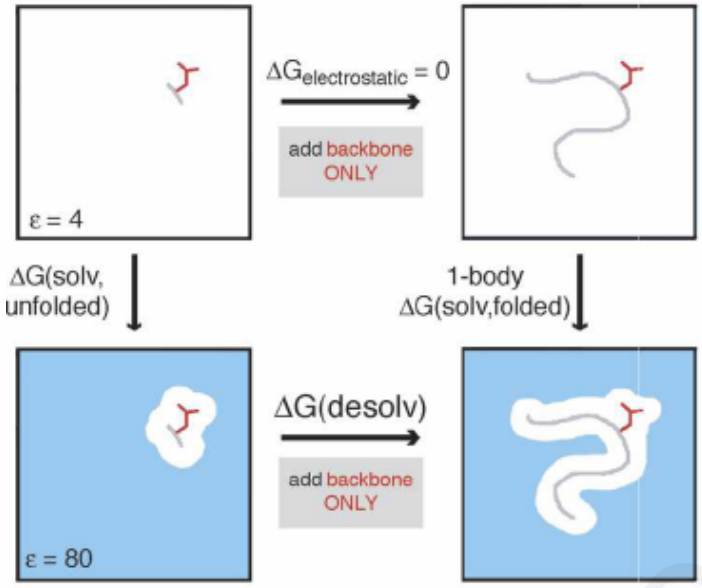


Figure 4.8: The side chain desolvation (bottom horizontal step) is given by the difference of the two vertical steps of the thermodynamic cycle, the side chain solvation energy in the presence of the entire backbone (folded state) and the local backbone atoms (unfolded state).The figure is taken from Ref. [70].

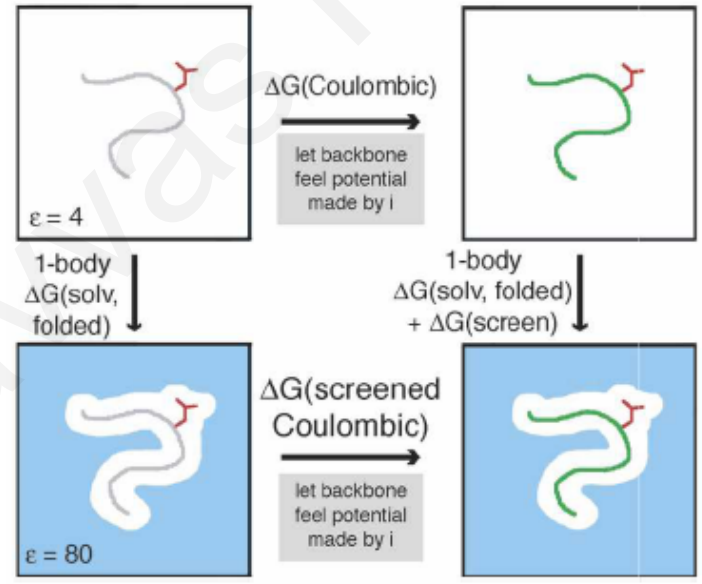
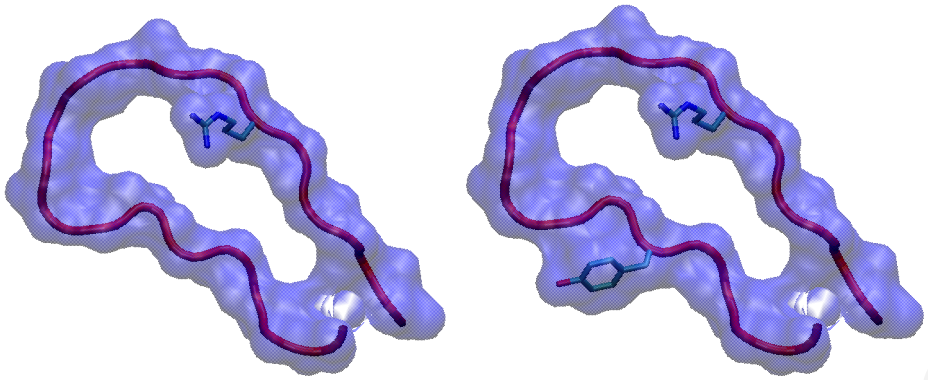


Figure 4.9: The top horizontal step yields the Coulombic interactions. The difference between the two vertical steps yields the screening of Coulombic interactions. The sum of the Coulomb and screening contributions yields the total electrostatic interactions in solution (bottom horizontal step).The figure is taken from Ref. [70].



(a) One-body PB (b) Two-body PB

Figure 4.10: (a) The "one-body" PB approximation employs the backbone and a single side chain to define the low dielectric cavity (b) The "two-body" approximation employs the backbone and a pair of side chains to define the low dielectric cavity.

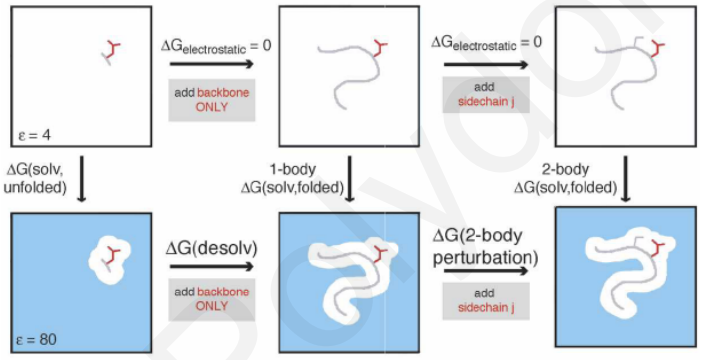


Figure 4.11: The thermodynamic cycle 4.8 is expanded by increasing the dielectric boundary with an additional side chain. The correction to the one - body side chain desolvation energy is computed by the difference of the two outer vertical steps minus the left horizontal step (one-body desolvation). The figure is taken from Ref. [70].

from all other side chains j to the desolvation and side chain-backbone energy of a side chain i.

Eq. (4.14) is now written as:

$$\Delta G^{i,2^{nd}}_{desolv.} = \Delta G^{i,1^{st}}_{desolv.} + \sum_{j \neq i} [\frac{1}{2} \sum_{u}^{sc} q_u (\phi^{i,j,bb}_i - \phi^{ib}_i) - \Delta G^{i,1^{st}}_{desolv.}] \tag{4.16}$$

The quantity $\phi^{i,j,bb}_i$ is the electrostatic potential of side chain i due to atomic charges of side chain i inside a low dielectric cavity consisting of the backbone, the side chain of interest i and another side chain j. The sum is over all possible side chain - side chain pairs j,i involving the side chain of interest i. Inside the parenthesis is the desolvation energy change of the side chain i, due to the presence of a second side chain j.

Similarly we can include the contribution of all other side chains to the screening of the side chain i- backbone Coulombic interactions, by writing Eq. (4.15) as:

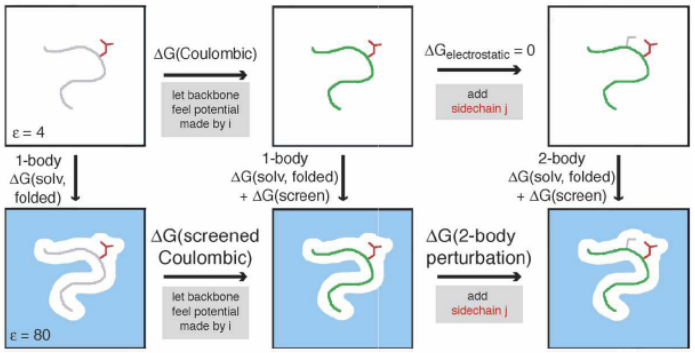


Figure 4.12: The thermodynamic cycle 4.9 is expanded by increasing the dielectric boundary with an additional side chain. The correction to the one - body side chain - backbone interaction energy is computed by the difference of the two outer vertical steps minus the left horizontal step (one-body screened coulombic interaction). The figure is taken from Ref. [70].

$$\Delta G^{i/bb,2^{nd}}_{\text{screenedCoul}} = \Delta G^{i/bb,1^{st}}_{\text{screenedCoul}} + \sum_{j\neq i}[\sum_{t}^{bb} q_t(\phi_i^{i,j,bb}(\epsilon_p \neq \epsilon_s) - \phi_i^{i,bb}(\epsilon_p \neq \epsilon_s))] \quad (4.17)$$

In the "two-body" approximation, the side chain - side chain interactions are also computable. Atomic charges of side chain i interact with atomic charges of side chain j according to Coulomb's law. At the same time, the atomic charges of side chain i polarize the surrounding high dielectric solvent (defined by the backbone and the side chain pair i,j), producing a reaction field. This reaction field produced by charges of side chain i interacts with atomic charges of side chain j:

$$\Delta G^{i/j,2^{nd}}_{\text{screenedCoul}} = \sum_{v} q_v(\phi_i^{i,j,bb}(\epsilon_p \neq \epsilon_s) + \phi_i^{i,j,bb}(\epsilon_p = \epsilon_s)) \quad (4.18)$$

where v and q_v are, respectively, the atoms and partial charges of side chain j.

Gathering the above energy contributions (backbone desolvation, side chain desolvation, side chain - backbone and side chain - side chain interactions), as expressed by the "one-body" and "two-body" approximations, a total electrostatic energy is obtained:

$$\Delta G^{prot}_{elec} = \Delta G^{bb,1^{st}}_{\text{desolv}} + \sum_i (\Delta G^{i,2^{nd}}_{\text{desolv}} + \Delta G^{i/bb,2^{nd}}_{\text{screenedCoul}}) + \frac{1}{2}\sum_i \sum_{i\neq j} \Delta G^{i/j,2^{nd}}_{\text{screenedCoul}} \quad (4.19)$$

This function is pairwise-decomposable (since at most two side chains are simultaneously present in the electrostatic free energy evaluations); as such, this approximation is applicable in high-throughput CPD problems.

The "two-body" approximation yields reasonable agreement with the exact FDPB

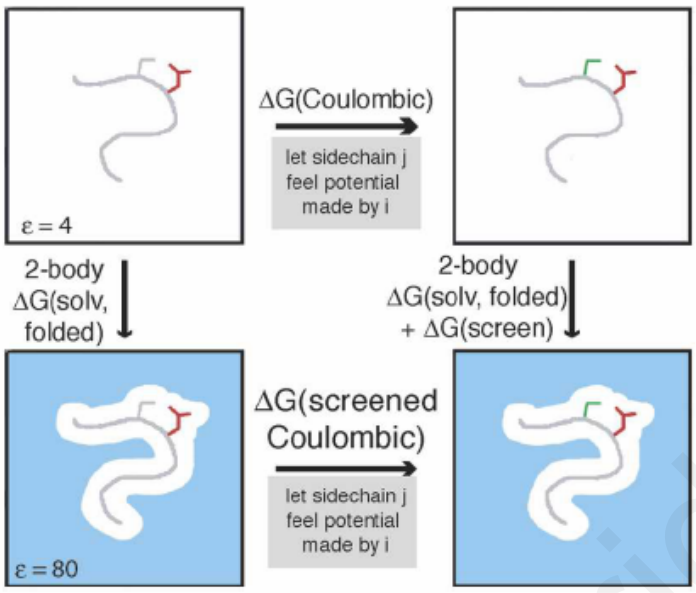


Figure 4.13: The total side chain - side chain electrostatic interactions are computed from the above thermodynamic cycle. The Coulombic interaction is obtained by the top horizontal step. The screening of this Colombic interaction is obtained as the difference between the two vertical steps. Adding these contributions together results in the total electrostatic interaction between the two side chains in solution (bottom horizontal step).The figure is taken from Ref. [70].

method. To advance the method further, Mayo and coworkers developed an alternative pairwise-decomposable FDPB approximation [177], in which the solvent-protein dielectric boundary surface is defined by the protein backbone, a pair of side chains and three overlapping spheres in the place of each other side chain Fig. 4.14. All the calculations are performed in a similar way. The results of this method are compared to our "residue-GB" pairwise approximation in Section 5.3.

The Modified Tanford - Kirkwood Model

The original Tanford-Kirkwood (TK) model approximated a protein by a sphere, a simple geometry, for which the PB equations could be solved analytically. The atomic charges of the protein were mapped onto the sphere as shown in Fig. 4.15. Each charge was mapped by preserving its solvent exposed surface area. The modified Tanford - Kirkwood model [178] (MTK) improved the original model by (i) replacing the atomic point charges with charge distributions, (ii) by using a Coulomb-field approximation in the charge-mapping procedure, and (iii) by substituting the original solution of the PB equation with a more precise solution, which employed image charges.

The electrostatic free energy is given by Eq. (4.20).

$$W^{\text{elec}} = \sum_i W_i^{\text{self}} + \sum_{i<j} I_{ij}^{\text{inte}} = \frac{1}{2} \sum_{i,j} q_i \Phi_j \tag{4.20}$$

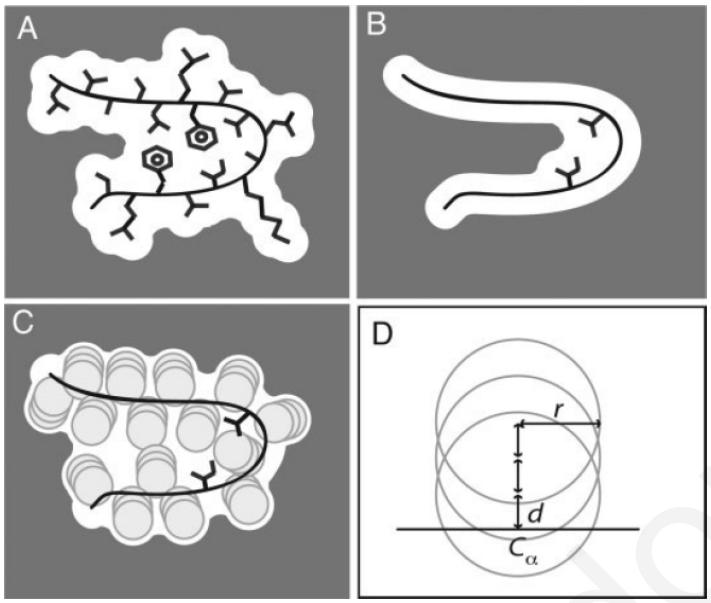


Figure 4.14: Schematic representation of the dielectric boundary surface used in exact FDPB (A), the "two-body" approximation (B), and the improved FDPB approximation of Ref. [177] (C). In the improved method, the missing side chains are replaced by three overlapping spheres (D). The figure is taken from Ref. [177].

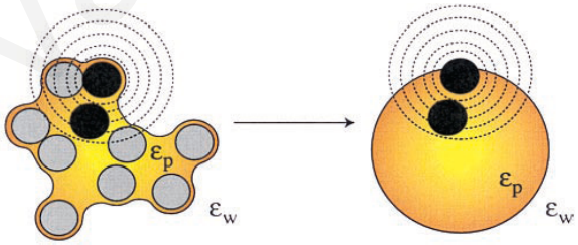


Figure 4.15: The TK model approximates the arbitrary shape of a protein molecule by a sphere. The atomic charge positions are matched inside the sphere. The dashed lines represent increasing probe radii, used to compute a series of estimates for the accessible surface area of each atom. Atomic charges are positioned in the spherical geometry, while preserving the average of the ASA series of each atom. The figure is taken from Ref. [178].

It consists of two terms, describing atomic self-energies (interactions of charges with their own reaction field) and pair-interactions (interactions of a charge with the reaction field induced by a different charge). For a single atomic charge present, the calculated electrostatic energy corresponds to the atomic self energy W_i^{self}. The interaction energy of two charges is computed as the difference in the electrostatic field due to the pair of co-existing charges, and the individual charges, present one at a time: $I_{ij}^{inte} = W_{ij} - W_i - W_j$. This follows from the linearity of the PB equation.

Atomic charges q_i are distributed onto spherical shells of radius equal to the atomic Born radius. Moving from the arbitrary shape of the protein to a low dielectric sphere representing the protein, shell charges are mapped into positions of the sphere so that they maintain their electrostatic self energy. Calculating the self energy of q_i by an energy density integral, the Coulomb Mean Field approximation is employed to assume a uniform dielectric field all over the integration space. Finally, the reaction potentials Φ_j of Eq. (4.20) can be obtained analytically using image charge solutions.

Generalized Born Approximation

The Generalized Born model (GB) is a continuum electrostatics approximation, in which the atoms of the protein are associated with a charge q_i and a radius b_i, that is indicative of the distance separating atom i from the protein-solvent boundary. Atoms inside the core of the protein are associated with large Born radii, while atoms situated close to the solute-solvent boundary surface have smaller radii. The GB energy function depends on the atomic coordinates.

The original Born model applies to an isolated solvated spherical ion of radius R with a charge q_i at its center (or distributed with spherical symmetry). The atomic charge polarizes the surrounding solvent, inducing a reaction field (ϕ_{reac}). The interaction between the ionic charge and its reaction field results in the Born solvation free-energy, given by the following formula [179].

$$\Delta G = \frac{1}{2} q_i \phi_{reac} = \frac{1}{2} q_i (\phi_{sol} - \phi_{gas}) = \frac{q_i}{2} (\frac{q_i}{\epsilon_w b} - \frac{q_i}{b}) = (\frac{1}{\epsilon_w} - 1) \frac{q_i^2}{2b} \qquad (4.21)$$

This formula expresses the change in free-energy upon transfering the ion from gas phase ($\epsilon_{in} = \epsilon_{ex} = 1$) to the solvent ($\epsilon_{in} = 1, \epsilon_{ex} = \epsilon_w$).

For molecules with an arbitrary shape and an arbitrary charge distribution (like proteins), the above formula is replaced by the generalized Born model; this model corresponds to an interpolating formula, which bridges two limits: in one limit the atoms of a molecule are treated as atomic spheres, completely immersed in water and largely separated from each other. In the other limit, the atoms are superposed on top of each other.

The first limit corresponds to a collection of spherical ions with charges q_i and radii

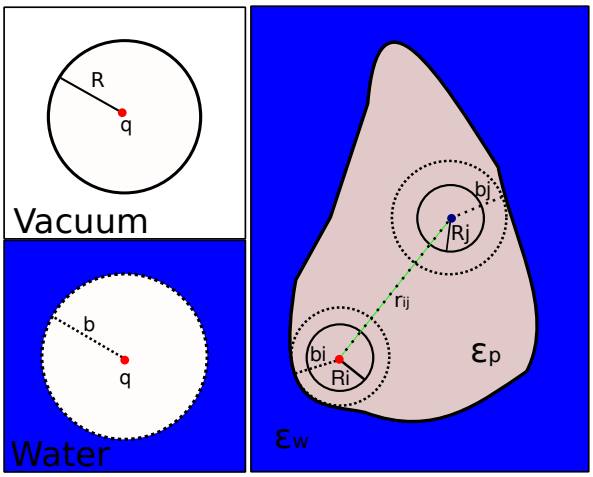


Figure 4.16: Left: A spherical ion in vacuum is immersed inside water. The charge at the center of the ion polarizes the solvent which creates a reaction field interacting with the charge. The solvation energy corresponds to a charge inside a sphere of radius b the Born radius. Right: The generalized Born model is employed in cases with many ions inside an arbitrary shaped cavity of low dielectric medium.

b_i (e.g. the van der Waals radii of the individual atoms), which are separated by very large distances ($r_{ij} >> b_i + b_j$). Imagine for simplicity that the system consists of a pair of two charges q_1 and q_2, with identical van der Waals radii b. The total solvation free energy of the system is the sum of the Born energies [Eq. (4.21)] and an interaction term, which has a screened-Coulomb functional form:

$$\Delta G = \tau \sum_{i=1}^{2} \frac{q_i^2}{2b} + \tau \frac{1}{2} \sum_{i \neq, j=1}^{2} \frac{q_i q_j}{r_{ij}} \tag{4.22}$$

where $\tau \equiv (1/\epsilon_w - 1)$.

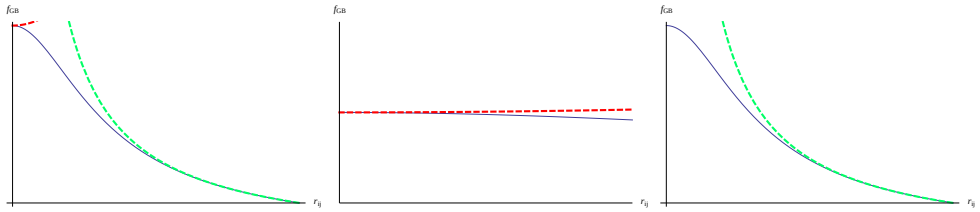
In the opposite limit ($r_{ij} \longrightarrow 0$) the two charges are superposed on each other, yielding a spherical ion of radius b and total charge $q_1 + q_2$. The solvation free energy becomes:

$$\Delta G = \tau \sum_{i,j=1}^{2} \frac{q_i q_j}{2b} = \tau \frac{1}{2b}(q_1 + q_2)^2 \tag{4.23}$$

The GB model [149] is an interpolation formula between these two limits:

$$\Delta G = \tau \frac{1}{2} \sum_{i,j=1}^{2} \frac{q_i q_j}{f_{GB}} = \tau \frac{1}{2} \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + b^2 e^{-r_{ij}^2/(4b^2)}}} \tag{4.24}$$

It is easy to verify that the above equation has the correct limiting behavior for $r_{ij} \to 0$ and $r_{ij} \to \infty$. (Fig. 4.17). In general, for a collection of charges $\{q_i\}$ centered on different nuclei of a biomolecular system, the GB expression for the solvation free energy becomes:



(a) Still interpolation function (b) Short distance limit (c) Long distance limit

Figure 4.17: (a) The interpolation function proposed by Still satisfies the two extreme limits. (b) At very short distances $r_{ij} \to 0$ the exponential term survives reproducing the born self energies (c) When atomic charges are isolated (long distance limit) the exponential term diminishes preserving only the coulombic interactions.

$$\Delta G = \tau \frac{1}{2} \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i\, b_j e^{-r_{ij}^2/(4 b_i\, b_j)}}} \tag{4.25}$$

The parameters b_i, b_j are referred to as "atomic Born radii". These parameters are related to the distance of the atomic charges from the solute-solvent boundary surface, which depends on the atomic coordinates of the entire molecule. Due to this dependence, the GB solvation energy is not pairwise decomposable, even though it appears to be expressed as a sum of atom pairs (this point is discussed in detail below).

The Born radii can be exactly computed, if we consider that the biomolecule contains only one charge (q_i) at a time. In this case, the solvation free energy is given by the solution of the PE:

$$\Delta G_i^{\text{solv}} \equiv \big(\frac{1}{\epsilon_w} - 1\big) \frac{q_i^2}{2\, b_i} \Rightarrow b_i = \big(\frac{1}{\epsilon_w} - 1\big) \frac{q_i^2}{2\, \Delta G_i^{\text{solv}}} \tag{4.26}$$

Of course, following this approach is impractical, as it would require N PE calculations (one for each charge), for a *single* GB energy evaluation! Thus, for GB to be practical, it is necessary to devise a fast and accurate way to compute the Born radii. From classical electrostatics it is known that the work needed to insert a charge q_i inside a linear dielectric medium ($\vec{D} = \epsilon \vec{E}$) is given by the formula [150]

$$W_i = \frac{1}{8\pi} \int d^3 r \vec{E}_i \cdot \vec{D}_i = \frac{1}{8\pi} \int d^3 r \frac{\vec{D}_i}{\epsilon} \cdot \vec{D}_i \tag{4.27}$$

where $\vec{E}_i$ is the electric field and $\vec{D}_i$ the corresponding electric displacement vector of charge q_i. The GB model makes another approximation at this point, assuming that the electric displacement of charge q_i is parallel to the radial direction emanating from charge q_i [$\vec{D}_i = \frac{q_i \vec{r}}{r^3}$] in the entire space (inside and outside the solute), despite the inhomogeneity of the solvent/solute medium. This is known as the "Coulomb Field Approximation" (CFA) [151]. With this approximation, the above integral simplifies:

$$W_i = \frac{1}{8\pi} q_i^2 \int d^3r \frac{1}{\epsilon(r) r^4} \tag{4.28}$$

The atomic charges q_i are assumed to be located on small spheres of radius a_i (instead of point charges) to avoid singularities. The integration in the above formula is over all space external to the biomolecule (where the dielectric constant is ϵ_{ex}) and over the interior of the biomolecule (where the dielectic constant is ϵ_{in}), with the exception of the interior of the charge-containing sphere of radius α_i. The change in solvation free energy due to transfering the charge-containing sphere from gas phase ($\epsilon_{in} = \epsilon_{ex} = 1$) at site i of the solvated biomolecule is then given by the expression:

$$\begin{aligned}
\Delta G_i &= W_i^{\epsilon_{ex}=\epsilon_w} - W_i^{\epsilon_{ex}=\epsilon_{in}=1} \\
&= \Big(\frac{1}{8\pi}\int_{in, r>a_i} \frac{q_i^2}{r^4} d^3r + \frac{1}{8\pi}\int_{ex} \frac{q_i^2}{\epsilon_w r^4} d^3r\Big) - \Big(\frac{1}{8\pi}\int_{in, r>a_i} \frac{q_i^2}{r^4} d^3r + \frac{1}{8\pi}\int_{ex} \frac{q_i^2}{r^4} d^3r\Big) \\
&= -\tau \frac{1}{8\pi}\int_{ex} \frac{q_i^2}{r^4} d^3r = -\tau \frac{q_i^2}{8\pi}\int_{ex} \frac{r^2 \sin\theta \cos\phi}{r^4} dr d\theta d\phi \equiv -\tau \frac{q_i^2}{2b_i}
\end{aligned} \tag{4.29}$$

The last line corresponds to the definition of the Born radius b_i. This Born radius can be computed from the following integrals:

$$\begin{aligned}
b_i^{-1} &\equiv \frac{1}{4\pi}\int_{ex} \frac{d^3r}{r^4} = \frac{1}{4\pi}\int_{ex} \frac{d^3r}{r^4} + \frac{1}{4\pi}\int_{in, r>a_i} \frac{d^3r}{r^4} - \frac{1}{4\pi}\int_{in, r>a_i} \frac{d^3r}{r^4} \\
&= a_i^{-1} - \frac{1}{4\pi}\int_{in, r>a} \frac{d^3r}{r^4}
\end{aligned} \tag{4.30}$$

The second integral is performed over the interior of the solute (with the exception of the charge-containing sphere α_i). In practice, this integral is calculated numerically; the approximation used generates several different GB models. In what follows, we present GB models, which express the integral of Eq. (4.30) as a sum of pairwise terms. For these models, the GB *self energy* term $[\sum_i q_i^2/(2b_i)]$ is pairwise-decomposable and can be precomputed, as required in high-throughput protein design calculations.

The Hawkins Crammer Truhlar Generalized Born (GB-HCT) Approximation

The Hawkins-Cramer-Truhlar (HCT) approximation to the GB model [90], approximates the integrals of the GB radii as sums of atomic contributions. We can rewrite Eq. (4.30) extending the integral all over the space, by a suitable step function $W(r_i, ..., r_N)$(unity for the interior and zero outside the solute), that depends on the atomic coordinates of the system; b_i is the Born radius of atom i, r is the radius of the integration sphere centered at atom i, $r_{ii'}$ the distance between the atoms i, i' and a_i the intrinsic radius of atom i. Performing the integration over the azimuthal and

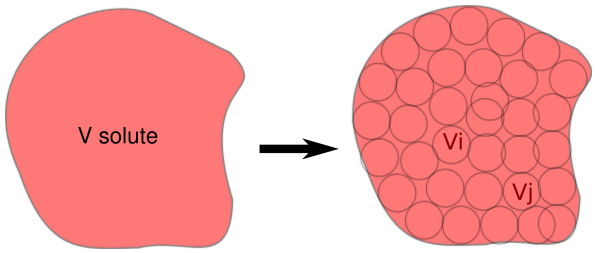


Figure 4.18: The integral over the solute volume (right) is approximated by the sum of atomic integrals over the atomic volumes (left).

inclination angles, the three dimensional step function W is converted to the solvent-exposed surface area of the solute, $A\{r_i, ..., r_N\}$. The fraction of the exposed surface area to the total atomic surface ($A(r)/4\pi r^2$) is expressed by the function $F(r_i, ..., r_N)$, while $H(r_i, ..., r_N)$ corresponds to the solvent excluded-surface-area fraction of the solute ($1 - F(r)$). All the above functions are defined by the geometry of the system specified by all atomic coordinates $\{r_{ii'}, a_{i'}\}$.

$$\begin{aligned} b_i^{-1} &= \frac{1}{4\pi}\int_{a_i}^{\infty} dr\, \frac{1}{r^4}\int_0^{\pi} d\theta\, \sin\theta \int_0^{2\pi} d\phi\, W(r,\theta,\phi,\{r_{ii'},a_{i'}\}) \\ &= \int_{a_i}^{\infty} \frac{dr}{r^2}\frac{A(r,\{r_{ii'},a_{i'}\})}{4\pi r^2} \\ &= \int_{a_i}^{\infty} \frac{dr}{r^2} F_i(r,\{r_{ii'},a_{i'}\}) \\ &= \int_{a_i}^{\infty} \frac{dr}{r^2}(1 - H_i(r,\{r_{ii'},a_{i'}\})) \\ &= a_i^{-1} - \int_{a_i}^{\infty} \frac{1}{r^2} H_i(r,\{r_{ii'},a_{i'}\}) \end{aligned} \tag{4.31}$$

The pairwise-approximation [90] assumes that the integral over the solute space can be expressed as a sum of individual atomic contributions (atomic integrals), as shown schematically in Fig. 4.18 and Eq. (4.32).

$$b_i^{-1} = a_i^{-1} - \sum_{i'} \int_{a_i}^{\infty} \frac{1}{r^2} H_{ii'}(r_{ii'}, r_{i'}) \tag{4.32}$$

The intersection surface area of two overlapping spheres i and i' is calculated analytically by $H_{ii'} = \frac{1}{2} - \frac{r_{ii'}^2 + r^2 - a_{i'}^2}{4 r_{ii'} r}$. The Born radius takes the following form :

$$
\begin{aligned}
b_i^{-1} &= a_i^{-1} - \frac{1}{2}\sum_{i'}\int_{U_{ii'}}^{L_{ii'}} dr\,(\frac{1}{r^2} - \frac{r_{ii'}}{2r^3} - \frac{1}{2r_{ii'}r} + \frac{a_{i'}^2}{2r_{ii'}r^3}),\\
&= a_i^{-1} - \frac{1}{2}\sum_{i'}\big[\frac{1}{L_{ii'}} - \frac{1}{U_{ii'}} + \frac{r_{ii'}}{4}(\frac{1}{L_{ii'}^2} - \frac{1}{L_{ii'}^2}) + \frac{1}{2r_{ii'}}\ln\frac{L_{ii'}}{U_{ii'}}\\
&\quad + \frac{a_{i'}^2}{4r_{ii'}}(\frac{1}{L_{ii'}^2} - \frac{1}{U_{ii'}^2})\big],
\end{aligned}
\tag{4.33}
$$

$$
\begin{aligned}
L_{ii'} &= 1 \text{ if } r_{ii'} + a_{i'} \le a_i,\\
L_{ii'} &= a_i \text{ if } r_{ii'} - a_{i'} \le a_i < r_{ii'} + a_i,\\
L_{ii'} &= r_{ii'} - a_{i'} \text{ if } a_i \le r_{ii'} - a_{i'},\\
U_{ii'} &= 1 \text{ if } r_{ii'} + a_{i'} \le a_i,\\
U_{ii'} &= 1 \text{ if } a_i < r_{ii'} + a_{i'}
\end{aligned}
\tag{4.34}
$$

The Analytical Continuum Electrostatics Generalized Born (GB/ACE) Approximation

The analytical continuum electrostatics method (ACE) [154] of Schaefer and Karplus is a different GB variant. It approximates the electrostatic contribution to solvation energy by an analytic potential energy function with continuous derivatives. Its differences from other GB variants are (a) the description of charge densities by gaussian distributions ρ instead of point (or surface) charges, and (b) the use of gaussian density functions P instead of step functions to describe the solute volume:

$$
\rho_i(\overrightarrow{x}) = q_i\pi^{-3/2}a_i^3\exp(-a_i^2(\overrightarrow{x} - \overrightarrow{x_i})^2) \quad a_i = \frac{\sqrt{\pi/2}}{R_i}, \tag{4.35}
$$

$$
P_k(\overrightarrow{x}) = \frac{4}{3\sqrt{\pi/2}a^3}\exp(-(\overrightarrow{x} - \overrightarrow{x_i})^2/(a\widetilde{R})^2) \tag{4.36}
$$

The parameter a_i is inserted to reproduce the Born self-energy in the case of single ion: $\Delta G_i^{\text{self}} = (1/2)\int \rho_i\phi_i dV = q_i^2/(2\epsilon R_i)$. The solute volume is divided into atomic volume contributions, weighted by atomic density functions $P_k((\overrightarrow{x}))$:

$$
\sum_k P_k(\overrightarrow{x}) = P_{solute}(\overrightarrow{x}) = \begin{cases} 1 & \text{if } \overrightarrow{x} \text{ inside the solute} \\ 0 & \text{elsewhere} \end{cases} \tag{4.37}
$$

These functions are normalizable and continuously differentiable, as shown above. One can express these atomic volumes as Voronoi polyedra [180], avoiding overlapping and voids between atoms when spheres are considered, but it is shown that, in general the atomic volumes fluctuate slightly around a mean value $\widetilde{V_k}$, which is independent of the

geometry employed, and constant for each atom type.

The Asymptotic GB Variant To determine the atomic Born radius as given by Eq. (4.30) one needs to evaluate the integral over all space occupied by the solute, by decomposing it into individual contributions of atomic volumes. In the long-range limit $r_{ij} \longrightarrow \infty$, the atomic integrals $\int r_{ij}^{-4} dV$ are just the product, $r_{ij}^{-4} V_j$ where V_j is the volume of atom j. Qiu et. al. proposed that the atomic solvation self-energy (and therefore the Born radius of the atom) is proportional to $r_{ij}^{-4} V_j$ [151], but the assumption fails to describe the short-range limit $r_{ij} \longrightarrow 0$. To overcome this, the self-energy can be written as a sum of $r_{ij}^{-4} V_j$ terms, multiplied by empirical scaling factors P_i which change with the interatomic distance:

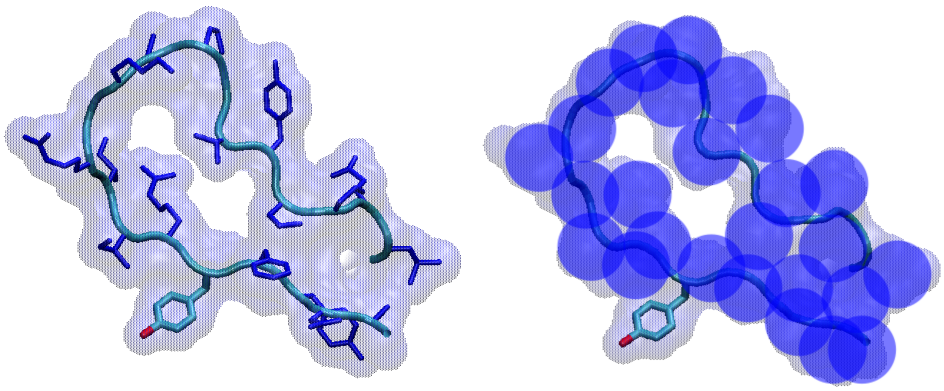
$$\int \frac{dV}{r^4} = P_1 + \sum_j P_2 \frac{V_j}{r^4} + \sum_j P_3 \frac{V_j}{r^4} + \sum_j P_4 \frac{V_j}{r^4} \tag{4.38}$$

The factors P_1, P_2, P_3 describe, respectively, the contribution of atomic pairs that overlap with each other, are directly bonded, are separated by two covalent bonds. Parameter P_4 describes atoms, which are separated by at least three bonds or are nonbonded. Optimized values of the P_i parameters are obtained by fitting the above equation to FDPB results, to achieve the maximum accuracy of the approach with respect to PB.

The Residue-Pairwise GB Approximation of Handel

The main criterion judging the usefulness of these models is their accuracy, that is usually tested by comparison with the PB approximation. The CASA approximation has been extensively used in biomolecular calculations, including CPD. GB models have been extensively used in solvation calculations, but not in CPD design due to the fact that they are many-body functions. Handel and coworkers developed a pairwise-decomposable GB approximation [71], based on the asymptotic variant. For the method to be fully pairwise it is essential that the total GB interaction energy of a residue pair can be expressed solely in terms of the atomic coordinates of the pair. This implies that the atomic Born radii entering in the GB expression need also to depend only on the coordinates of the particular pair. In Handel's approach, the atomic GB radii are computed from the GB self energy as $b_i = \frac{-\tau q_i^2}{2\Delta G_i^{self}}$. When calculating the self energy of atom i, the dielectric boundary surface is defined by the side chain of interest, and the backbone. All the rest side chains are replaced by dummy atoms represented by spheres filling up the solute space of the omitted side chains (Fig. 4.19). A similar approach was taken by Mayo and his group [70] to developed a pairwise-decomposable PB model; this was described in detail in Section 4.1.2.

Following the asymptotic-GB approximation, the self energy is written in Handel's approach as a sum over atomic contributions:



(a) Exact structure (b) Approximate structure

Figure 4.19: (a) The atomic Born radii of the side chain of interest are computed using the exact structure of the protein in the presence of all the other side chains (blue lines) described in atomic detail. (b) In the approximate structure the blue lines are replaced by dummy atoms represented by spheres (blue circles) filling the missing side chains space.

$$\Delta G_i^{self} = \underbrace{P_1 + \sum_j P_2 \frac{V_j}{r^4} + \sum_j P_3 \frac{V_j}{r^4} + \sum_j P_4 \frac{V_j}{r^4}}_{\text{side chain of interest}} \underbrace{+k_{BB} \sum_j P_4 \frac{V_j}{r^4}}_{\text{backbone and pseudo side chains}} \tag{4.39}$$

The first four terms correspond to sums over neighbouring atoms inside the same side chain of interest, and the last term sums contributions from the rest environment: the backbone and the pseudoside chains. In the exact calculation of the atomic Born radii, the environment of the side chain of interest is represented by the atomic coordinates of the backbone and the remaining side chains. The scaling factor k_{BB} as well as the radius of the employed dummy atoms employed are optimized by fitting approximate GB self-energies to PB values.

Chapter 5

The Residue GB Approximation

High-throughput CPD methods require a compromise between speed and accuracy. Solvent effects are usually taken into account by approximate models, such as the ones described in Chapter 4. As explained in Section 3.1.4, the residue-pairwise additivity property of the scoring function is essential for efficient algorithms employed in high-throughput CPD calculations. Since the environment around a particular residue pair is not fixed (surrounding residues usually undergo conformational or chemical changes during the design), the pair interaction energy needs to be independent of its surrounding residues and depend solely on the atomic coordinates of the specific residue pair. This condition is satisfied by simple solvation free energies, such as the CASA approximation of Section 4.1.1.

More accurate representations based on continuum electrostatics, like the PB or GB models, are not pairwise-decomposable, since the solute-solvent dielectric boundary depends on all the atomic coordinates of the protein (Section 4.1.2). Consequently, these models need further development, in order to be applicable in CPD problems. Some residue-pairwise PB and GB approximations were presented in Section 4.1.2. These models handle the problem of the fluctuating (due to the design) dielectric boundary with a "mean-field" assumption; e.g. in Mayo's approach, the biomolecular volume is filled with spheres, centered along the fixed backbone (Section 4.1.2) [70].

In 2005, Archontis and Simonson derived a residue-pairwise generalized Born approximation [91], which is not based on such assumptions. This approximation is presented in the remaining of this chapter.

5.1 Theoretical Formulation

In "atomic" GB solvation models, detailed in Section 4.1.2, the total solvation free energy is given by the Still formula [149]:

$$\Delta G_{\mathrm{solv}}^{\mathrm{GB}} = \frac{\tau}{2} \sum_{i,j} q_i \, q_j \, g_{ij} = \frac{\tau}{2} \sum_{i,j} \frac{q_i \, q_j}{[r_{ij}^2 + b_i b_j exp[-r_{ij}^2/(4 b_i b_j)]]^{1/2}} \tag{5.1}$$

with $\tau \equiv 1/\epsilon_w - 1/\epsilon_p$, and ϵ_w, ϵ_p, respectively, the solvent and protein dielectric constants.

Each of the diagonal terms is a "self-energy" of interaction between a charge q_i and the electrostatic potential it induces, by polarizing the surrounding medium. It is derived by the above formula, by setting $i = j$ and $r_{ij} = 0$:

$$\Delta G_i^{\text{self}} = \tau \frac{q_i^2}{2\, b_i}. \tag{5.2}$$

As explained in Section 4.1.2, the atomic Born radii (b_i) are *defined* by the above equation; that is, if the self-energy of a charge q_i can be computed by some method, the corresponding atomic radius b_i is

$$b_i \equiv \tau \frac{q_i^2}{2\, \Delta G_i^{\text{self}}} \tag{5.3}$$

A "perfect" method to compute ΔG_i^{self} would imply solving the Poisson (PE) [or Poisson-Boltzmann (PB)] equation for a single charge q_i at its corresponding location in the solvated biomolecule, while setting all other charges to zero. This approach is impractical, as it would require N PE evaluations (for N atomic charges). More practical approximations replace this PE-based self-energy evaluation by integrals over the biomolecular volume, as explained in Section 4.1.2. Different approximations to these integrals yield distinct GB variants. Two such variants, the GB/ACE [154] and GB/HCT [90] were presented in Section 4.1.2.

The starting point of the residue-pairwise GB approximation is that the self-energy of atom i [Eq. (5.2)] can be written in many GB variants, including GB/ACE and GB/HCT, as a sum of residue-pair terms:

$$\Delta G_i^{\text{self}} = \sum_j f_{ij}, \tag{5.4}$$

where f_{ij} is a quantity that can be thought of as the free energy to replace solvent by the low dielectric biomolecular solute in the volume of atom j, when q_i is the only charge present; f_{ij} is related to the integral of the electrostatic energy density over the volume associated with atom j [see Eq. (4.27) in Section 4.1.2].

As explained in Section 4.1.2, this integral is computed by (i) making the "Coulomb-field approximation" (which describes the electrostatic field of charge q_i by a radial function $\propto r^{-2}$, a hypothesis that is true in a homogeneous medium), and (ii) by approximating the volume around atom j by a suitable analytic function (e.g. a gaussian density distribution in the GB/ACE approximation - see Section 4.1.2). After these approximations, the function f_{ij} acquires an analytic form that depends on the atomic coordinates of the atom pair i-j (usually the magnitude of the separating vector $\vec{r}_{ij}$); this function is differentiable with respect to the coordinates $\vec{r}_i$, $\vec{r}_j$, permitting also the evaluation of atomic forces (in addition to the energies), and the use of the GB

approximation in Molecular Dynamics algorithms. Examples of functional forms of f_{ij} for the GB/ACE and GB/HCT models are given in Section 4.1.2.

Eq. (5.4) is a key property of these GB models. It implies that the GB *self-energy* is strictly pairwise-decomposable. Indeed, the function f_{ij} depends only on the coordinates of i and j and can be computed without knowledge of the surrounding medium. Note that this pairwise-decomposability does not hold for the non-diagonal terms ($i \neq j$) of Eq. (5.1). Indeed, from the definition of the Born radius for atom i, it follows:

$$b_i \equiv \tau \frac{q_i^2}{2\,\Delta G_i^{\text{self}}} = \tau \frac{q_i^2}{2\,\sum_j f_{ij}(\vec{r}_i,\,\vec{r}_j)}. \tag{5.5}$$

From Eq. (5.5) it is clear that the Born radius b_i of atom i depends on the coordinates of all atoms.

The pairwise "residue-GB" approximation [91] exploits the pairwise-decomposability property of the GB self-energy [Eq. (5.4)], to derive a residue-pairwise form for the non-diagonal terms. First, by summing over all atoms in residues R and R$'$, a self-energy contribution can be defined, that is attributed to the residue pair (R, R$'$):

$$\Delta G_{RR'}^{\text{self}} \equiv \sum_{i\in R}\sum_{j\in R'} f_{ij} \tag{5.6}$$

This contribution depends only on the atomic coordinates of the residue pair ($\{\vec{r_i}...\} \in R, \{\vec{r_j}...\} \in R'$).

By going through all compatible side chain chemical types and rotamers for a residue at all positions $\{R\}$, all possible residue-pair terms $\Delta G_{RR'}^{\text{self}}$ can be computed and stored into an energy-matrix *prior* to the design, exactly in the same way as with other pairwise-decomposable quantities (e.g. the residue-pair Coulomb or van der Waals interaction energy). Then, for a particular combination of chemical types/orientations, the total GB self energy of residue R can be reconstructed from this matrix and the equation:

$$\Delta G_R^{\text{self}} = \sum_{R'} \Delta G_{RR'}^{\text{self}} \tag{5.7}$$

The total self energy of a given sequence/structure combination can be computed directly by adding together the corresponding residue self-energies:

$$\Delta G^{\text{self}} = \sum_{R} \Delta G_R^{\text{self}} \tag{5.8}$$

If the design algorithm modifies a residue (chemical type or orientation) at a particular position R, the self-energy of this residue can be updated by use of Eq. (5.7); similarly, the self-energies of all other residues (R$'$ $\neq$ R) can be also updated rapidly by the same

equation.

To derive an efficient and pairwise form for the non-diagonal terms of the GB approximation, a *residue* solvation radius B_R is first defined as:

$$\Delta G_R^{\rm self} \equiv \tau \sum_{i\in R} \frac{q_i^2}{2\,B_R} \Longrightarrow B_R = \tau \sum_{i\in R} \frac{q_i^2}{2\,\Delta G_R^{\rm self}} \tag{5.9}$$

This common radius is assigned to all atoms in residue R. The above definition permits the reduction of Born radii from N (the number of atoms) to $N_{\rm res}$ (the number of residues, $N_{\rm res} \approx N/10 - N/20$).

The residue Born radius B_R can be thought of as the average burial depth of residue R from the dielectric boundary surface. Actually, from the definition of the self energy of residue R, it follows:

$$\begin{aligned} \Delta G_R^{\rm self} &= \frac{\tau}{2} \sum_{i\in R} \frac{q_i^2}{b_i} = \frac{\tau}{2\,B_R} \sum_{i\in R} q_i^2 \\ &\Rightarrow \frac{1}{B_R} = \frac{1}{\sum_{i\in R} q_i^2} \sum_{i\in R} \frac{q_i^2}{b_i} \end{aligned} \tag{5.10}$$

Thus, the quantity B_R is a harmonic average of the atomic radii, weighted by the squares of the atomic charges.

The next step is to define the interaction between a pair of residues. An *ansatz* formula analogous with Eq. (5.1) is used:

$$g_{RR'} \equiv \tau \sum_{i\in R, j\in R'} \frac{q_i q_j}{[r_{ij}^2 + B_R B_{R'} \exp[-r_{ij}^2/(4\,B_R B_{R'})]]^{1/2}} \tag{5.11}$$

In the above definition, a factor 1/2 is included if R≡R$'$. This new interpolation scheme of the interaction function is reasonable: it behaves correctly at the two opposite limits of infinitely separated ($r_{ij} \to \infty$) and coinciding atoms ($r_{ij} \to 0$), while for intermediate distances it estimates the energy with a comparable accuracy as its atomic counterpart (see below). Eq. (5.11) is more convenient computationally than its atomic analog [Eq. (5.1)], because it contains fewer parameters ($N_{\rm res}$ radii B_R, instead of $N_{\rm atom}$ radii b_i); however, it is still not pairwise-decomposable, as the parameters B_R still depend on the coordinates of all atoms.

To derive a fully pairwise scheme, it can be noted that the interaction energy [Eq. (5.11)] of a given residue pair (R, R$'$) in a *fixed* rotamer orientation is a simple function of the product $B \equiv B_R\,B_{R'}$:

$$\tau \sum_{i\in R} \sum_{j\in R'} \frac{q_i q_j}{[r_{ij}^2 + B \exp[-r_{ij}^2/(4B)]]^{1/2}} \approx c_1^{RR'} + c_2^{RR'} B + c_3^{RR'} B^2 + c_4^{RR'} B^{-1/2} + c_5^{RR'} B^{-3/2} \tag{5.12}$$

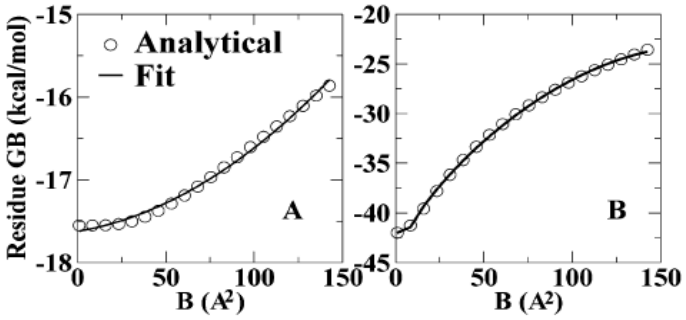


Figure 5.1: Representative fit of the residue GB interaction energy to a parabolic (A) or a five point (B) function of $B = B_R B_{R'}$. A and B are two different rotamer combinations of the same residue pair. The figure is taken from Ref. [91].

The behaviour of the left-hand-side is described very well by this five-parameter function (Fig. 5.1) at very large or very small values of B. For intermediate values of B, a simple quadratic function, containing the first three polynomial terms of the above expression, is sufficient.

The coefficients $\{c_i^{RR'}\}$ depend on the given pair (R, R$'$) through the atomic charges and (fixed) interatomic distances r_{ij} entering in Eq. (5.12). For each pair of residues (in all possible chemical types and orientations compatible with the pair), the coefficients $\{c_i\}$ can be determined by fitting of Eq. (5.12) in a reasonable range of B values (e.g. 1 - 200 Å) (Fig. 5.1). The fitted coefficients can be stored into a matrix and used to compute the interaction energy for specific values of B.

The self-energy matrix elements $\Delta G^{\rm self}_{RR'}$ [Eq. (5.6)] and the fitting coefficients [Eq. (5.12)] are computed for all possible sidechain chemical types and rotamer combinations compatible with a given protein fold. This calculation is performed prior to the actual design calculation. During protein design, the total self energy of a particular residue ($\Delta G^{\rm self}_R$) is constructed from Eq. (5.7), using the pretabulated matrix of pair self-energies $\Delta G^{\rm self}_{RR'}$. This self energy is inverted via Eq. (5.9), to compute the corresponding radius B_R. Then, for each pair (R, R$'$) the product $B \equiv B_R\, B_{R'}$ is computed and used in Eq. (5.12) to determine the interaction energy of the pair. The total GB energy is computed by adding all self- and interaction energies. In this way, the calculation of the GB energy is both efficient and fully pairwise-decomposable, both for the interaction and the self-energy contributions to the solvation energy.

5.2 Tests of the Residue GB Approximation

The "residue GB" approximation has (by construction) the same self-energy as its parent atomic-GB model [see Eq. (5.10)]. Thus, any differences between residue and atomic-GB energies result from the non-diagonal terms of Eq. (5.1) (the "interaction-energy" terms).

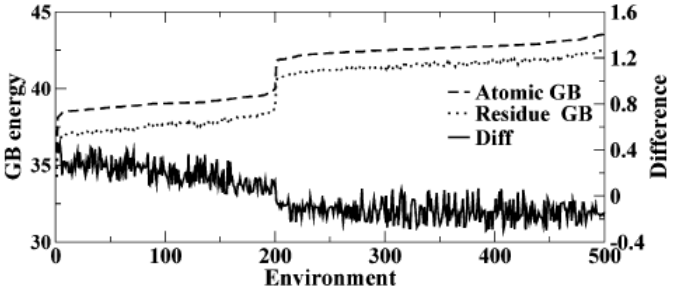


Figure 5.2: Comparison of the fluctuations in the atomic- and residue-GB interaction energies of the AspRS pair Lys198(in rotamer 22)-Glu235(in rotamer 10), due to changes in the surrounding environment. These changes are produced by considering 500 distinct structures of the AspRS active site, with randomized side chain conformations. For clarity, the individual atomic and residue-GB energies are displayed sorted in terms of increasing energy. The difference between the atomic and residue GB values is plotted as a solid line (with the scale on the right Oy axis). The difference of the average values has been subtracted. The figure is taken from Ref. [91].

5.2.1 Fluctuations in the Environment of a Fixed Residue Pair

For a residue pair in a *fixed* orientation, the Coulombic interaction is constant and independent of the environment. However, solvent effects are not constant, but fluctuate with the surrounding environment (e.g. the orientation of surrounding side chains and the shape of the dielectric boundary, separating the biomolecule from the solvent).

In the GB formulation, solvent effects are taken into account by the pair interaction-GB energy; the environmental dependence enters through the values of the solvation radii $\{b_i\}$. A first test of the sensitivity of the "residue-GB model" examined the question whether this approximation can reproduce changes in the interaction-GB energy of a residue pair, kept in a fixed orientation, due to fluctuations in its environment. To examine this question, the electrostatic interactions of five charged-residue pairs in the active site of the protein aspartyl-tRNA synthetase were computed. For each pair, 30-50 rotamer combinations were produced; for each combination, 500 random rotameric structures of the surrounding environment were created. A typical behavior is shown in Fig. 5.2, which displays atomic- and residue-GB/ACE interaction energies of the pair Lys198(in rotamer 22)-Glu235(in rotamer 10), for 500 randomized conformations of the AspRS active site (rotamers were taken from the Tuffery rotamer library [135]). As can be seen from the figure, the residue-GB approximation follows the fluctuations in the atomic-GB model, i.e. it is as able as its atomic counterpart in capturing effects of environment variations to the residue-pair interaction energies. The average standard deviation between the residue and atomic GB interaction energies for the given pair is 0.11 kcal/mol.

5.2.2 Dependence of Total Solvation Energies on Rotamer Conformations

In a second test, we checked the ability of the residue GB approximation to reproduce solvation energies of conformations with different side-chain rotamer orientations [181].

For each of the four proteins BPTI, ubiquitin, thioredoxin and lysozyme, we produced several hundred random conformations by randomizing the sidechain rotamer orientations, while keeping the protein backbone (and proline, cysteine, glycine, alanine side chains) fixed. We computed total solvation energies with the atomic- and residue GB/HCT approximation and compared with corresponding values from the Poisson equation (PE). The PE calculations employed the finite-difference UHBD program [182]. We used a cubic grid with an initial 0.8 Å spacing, to map the protein structure, followed by a focussing procedure with a final grid spacing of 0.4 Å. The dielectric constants for the protein and solvent were, respectively, 1 and 80. The protein/solvent dielectric boundary surface coincided with the molecular surface of the protein; this molecular surface was constructed by rolling a probe sphere with a 2Å-radius on the van der Waals surface of the protein; 2000 points per atom were used to describe the resulting molecular surface. The dielectric discontinuity at the interface was alleviated using a smoothing function. Atomic charges and radii were adopted from the AMBER all-atom force field [92]. The conformational changes introduced by the rotamer randomization can produce small voids in the protein interior. In a Poisson calculation these voids are assumed to be part of the surrounding solvent and are automatically assigned a high dielectric constant. To avoid this artifact, in our Poisson calculations the voids were filled with dummy atoms, which were assigned the protein dielectric constant.

Atomic- and residue GB/HCT solvation energies are plotted against the corresponding PE solvation energies in Fig. 5.3. The values vary in a range of 500 kcal/-mol. The atomic GB/HCT model (optimized in earlier work [76]), shows a very good agreement with the benchmark PE values across the whole energy range. The residue-GB/HCT model is at least as good as the atomic model, and is even somewhat better for lysozyme.

RMS differences (rmsd) between the GB/HCT and PE electrostatic solvation free energies are included in Table 5.1. The rmsd values vary between 21-45 kcal/mol for the atomic and 21-40 kcal/mol for the residue-GB approximation. In calculations with a less optimized GB/ACE model [76], the residue approximation was found to outperform the atomic model. Here, the residue GB/HCT is as good as the optimized atomic GB/HCT.

A similar calculation was done with the GB/HCT approximation, using several hundred randomized structures of the proteins BPTI, ubiquitin, thioredoxin and lysozyme. Atomic and residue-GB/HCT values are plotted against PE energies in Fig. 5.4. For all

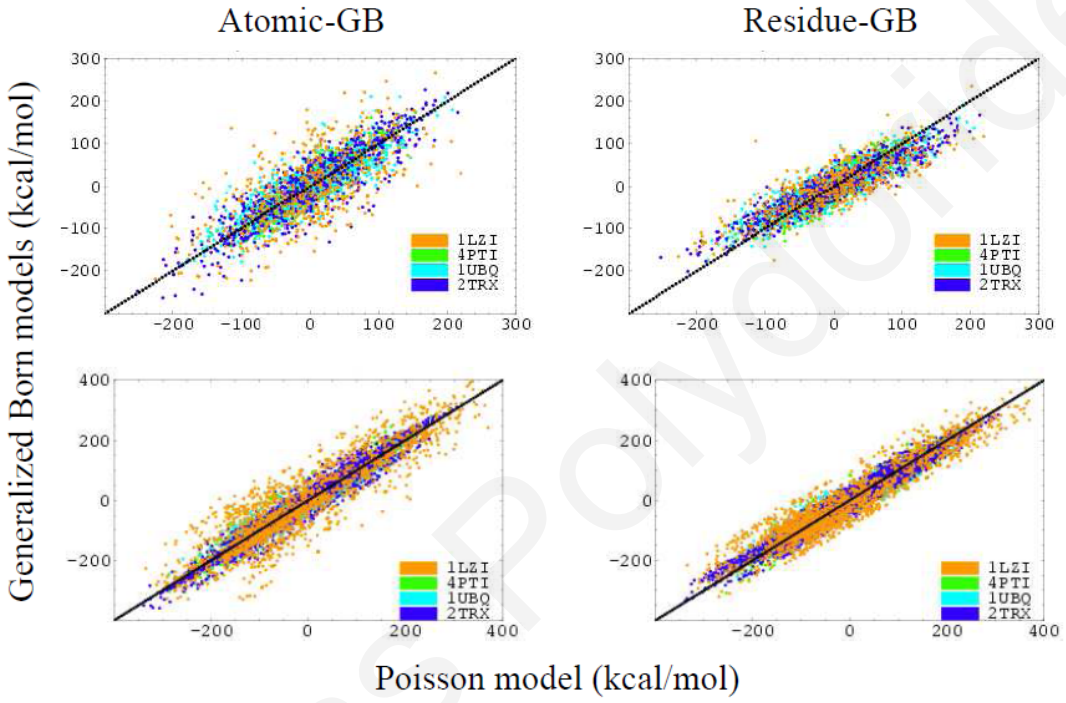


Figure 5.3: Solvation free energies of four proteins for multiple rotamer combinations (upper panels) and multiple protonation states (lower panels). Atomic-GB/HCT (left, vertical axis) and residue-GB/HCT values (right, vertical axis) are compared to the PE values (horizontal axes). The proteins are identified by their PDB codes and a color scheme(1LZ1 = lysozyme; 4PTI = BPTI; 1UBQ = ubiquitin; 2TRX = thioredoxin). The figure is taken from Ref. [181].

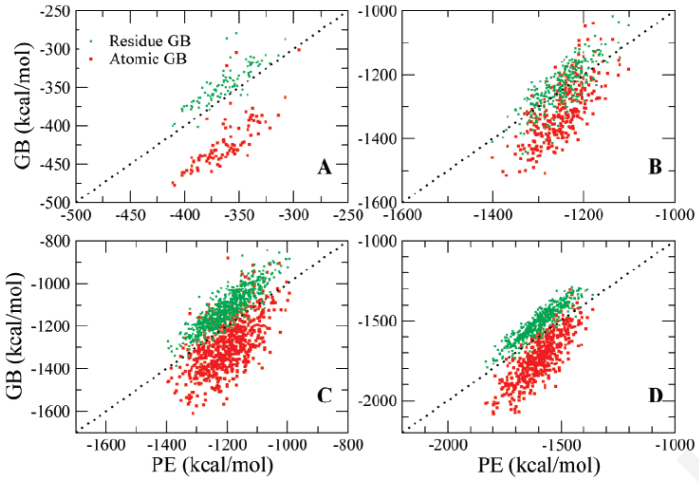


Figure 5.4: Residue- and atomic-GB/ACE solvation energies for several hundred random structures of trpcage (A), BPTI (B), ubiquitin (C) and thioredoxin (D), plotted against the corresponding PE energies. The figure is taken from Ref. [91].

proteins, the residue GB solvation energies (green dots) are better correlated with the PE values. The atomic GB solvation energies (red dots) are in general more negative than residue GB energies, due to more negative residue-pair interaction energies. This originates from the tendency of the atomic GB model to yield smaller solvation radii to solvent-exposed side chains; in contrast, the residue-GB approximation assigns slightly larger average radii in those residues.

5.2.3 Dependence of Total Solvation Energies on Changes in the Protonation State of Ionizable Residues

The model was also tested for predicting the solvation free energy change due to mutations in the protonation state of ionizable residues (Asp, Gly, His, Cys and Lys) of the four proteins BPTI, lysozyme, thioredoxin and ubiquitin. In each protein, the ionizable residues were initially assigned their corresponding dominant charge state at physiological pH. Subsequently, the charge state of one ionizable residue at a time was modified, and the resulting solvation free energy difference was computed for a 100 rotameric structures of the protein involved, created by randomizing the rotameric state of all sidechains. The resulting GB/HCT solvation energies are plotted against the corresponding PE energies in Fig. 5.3. The corresponding rms differences are included in Table 5.1. The GB/HCT values have a good correlation with the PE values, showing that the implicit-solvent model is able to reproduce this mutation. The rmsd values range between 22 - 52 kcal/mol for the atomic GB and 23 - 43 kcal/mol for the residue approximation. The residue approximation has a somewhat better agreement with PE for lysozyme and is comparable to the atomic model for the other three proteins.

Table 5.1: Comparing atomic-GB and residue-GB to the Poisson model

	Rotamers		Protonation states	
Protein	Atomic-GB	Residue-GB	Atomic-GB	Residue-GB
1LZ1	45.3	39.4	52.0	42.8
2TRX	30.2	28.9	24.5	24.9
1UBQ	26.0	27.7	25.4	24.9
4PTI	20.7	20.9	22.2	22.7

RMS deviation (kcal/mol) between the solvation free energies from GB and the Poisson model (PE) for random rotamer combinations and protonation states of titratable sidechains in four proteins (indicated by their PDB code). The table is taken from Ref. [181].

5.2.4 Dependence of Total Solvation Energies on Changes in the Residue Chemical Type

For each of the above four proteins, we generated 11 rotameric structures of the native sequence. For each structure, we performed all possible single-point mutations involving the twelve amino acid types listed in Table 5.2, and evaluated the corresponding GB/HCT and PE solvation free energies. The atomic or residue GB/HCT solvation free energies are plotted against the corresponding PE free energies in Fig. 5.5, and the resulting GB-PE rms differences (rmsd) are listed in Table 5.3. The strong correlation between the GB and PE reflects the good quality of our optimized GB/HCT model. Even though the mutations studied here are more complicated than the titrations considered above, the residue-GB approximation is still comparable to the atomic GB model. The agreement between the residue-GB and PE model is somewhat better for the two smallest proteins ubiquitin and BPTI, and somewhat worse for the two larger proteins. In Fig. 5.6 we decompose the residue GB/PE plot of Fig. 5.5 according to the following mutation types: cc, cn, cp, pp, pn, nn, and nn. In order to focus on the mutation effects, we consider separately the results of the different rotamer conformations. In Fig. 5.6 we show the data of Fig. 5.6, except that now we center the residue GB and PE free energy of each mutant sequence around the average (respectively, GB or PE) value of all mutants in the same category and in the same rotamer; thus, each value corresponds to a free-energy change with respect to the corresponding mean. Fig. 5.7 shows the corresponding atomic GB/PE plots. The range of free-energy changes is smaller, especially for the mutations which do not change the net residue charge. The correlation between the GB and PE energies is strong for mutations of charged residues and weaker for polar and neutral-residue transformations. Similarly, we compute separately the GB-PE rmsd for all mutants with the same initial rotamer conformation and average over the rmsd of the 11 rotamers. The resulting average rmsd values are included in Table 5.3; they are further decomposed according to the nine possible mutation types in Table 5.4.

Table 5.2: Aminoacid types considered in the chemical mutations

Group	Amino acids			
Charged (c)	Arg	Lys	Asp	Glu
Polar (p)	Ser	Thr	Gln	Tyr
Hydrophobic (n)	Ala	Phe	Val	Leu

The table is taken from Ref. [181].

Table 5.3: RMS difference between the generalized Born and corresponding Poisson solvation energies for the four proteins considered. All quantities in kcal/mol.

Protein	Chemical Mutations	
	Atomic	Residue
1LZ1	11.6	9.3
2TRX	8.2	7.9
1UBQ	6.5	5.4
4PTI	6.2	6.3

The table is taken from Ref. [181].

Table 5.4: RMS differences of Table 5.3, decomposed in terms of mutation types.

Mutation	GB Approximation	Proteins			
		1LZ1	2TRX	1UBQ	4PTI
CC	Atomic	22.6	11.6	10.2	12.9
	Residue	17.9	9.8	8.9	9.8
CP	Atomic	15.2	7.5	8.2	6.4
	Residue	11.5	8.5	8.6	10.6
CN	Atomic	14.5	7.6	8.0	7.7
	Residue	11.3	8.5	8.5	9.2
PP	Atomic	8.3	5.9	5.2	5.1
	Residue	7.6	5.3	3.7	5.2
PN	Atomic	7.4	5.1	4.4	4.1
	Residue	6.9	4.8	3.0	4.3
NN	Atomic	6.4	8.4	4.3	3.0
	Residue	4.7	8.2	2.5	2.3
NP	Atomic	6.9	11.2	5.1	4.2
	Residue	4.9	10.4	2.9	2.9

The table is taken from Ref. [181].

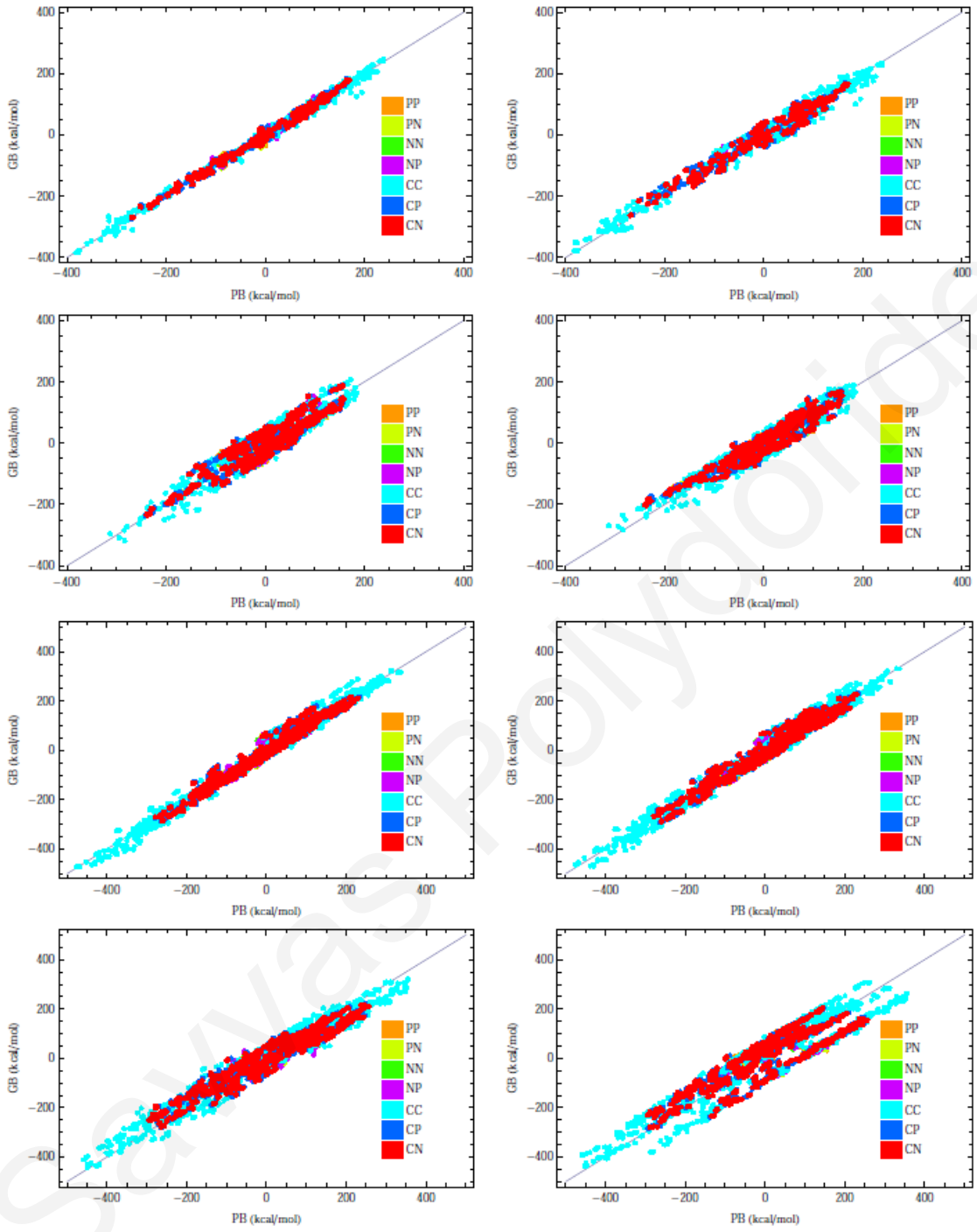


Figure 5.5: Atomic (left column) and residue (right column) GB/HCT electrostatic solvation free energies, plotted against the corresponding Poisson free energies for the proteins BPTI, ubiquitin, thioredoxin and lysozyme (from top to bottom). In each protein, single-point mutations involving 12 charged (c), polar (p) or hydrophobic (n) residues have been introduced into 11 rotameric structures. Points are colored according to the mutation type. The figure is taken from Ref. [181].

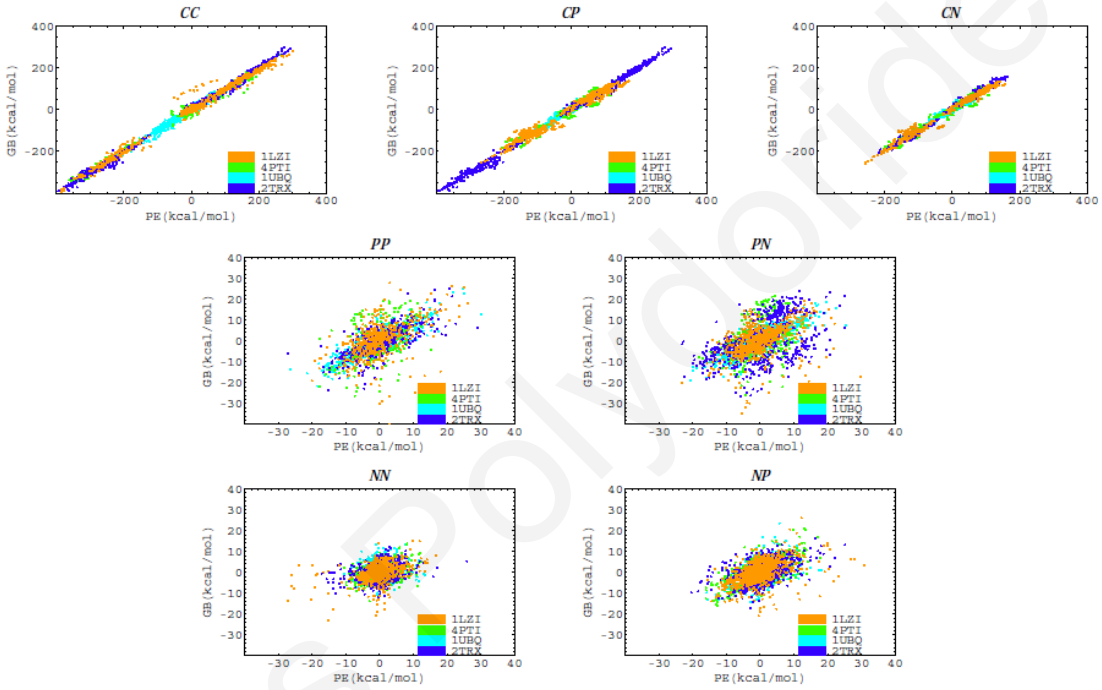


Figure 5.6: Decomposition of residue GB/HCT and PE electrostatic solvation free energies according to one of the seven mutation types (see text). The residue GB and PE free-energies of each mutant are centered, respectively, around the average residue GB and PE value for all mutants of the same category and in the same rotamer. Points are colored according to the proteins, listed by their pdb codes (1LZ1, lysozyme; 1UBQ, ubiqutin; 2TRX, thioredoxin; 4PTI, bovine pancreatic trypsin inhibitor). The figure is taken from Ref. [181].

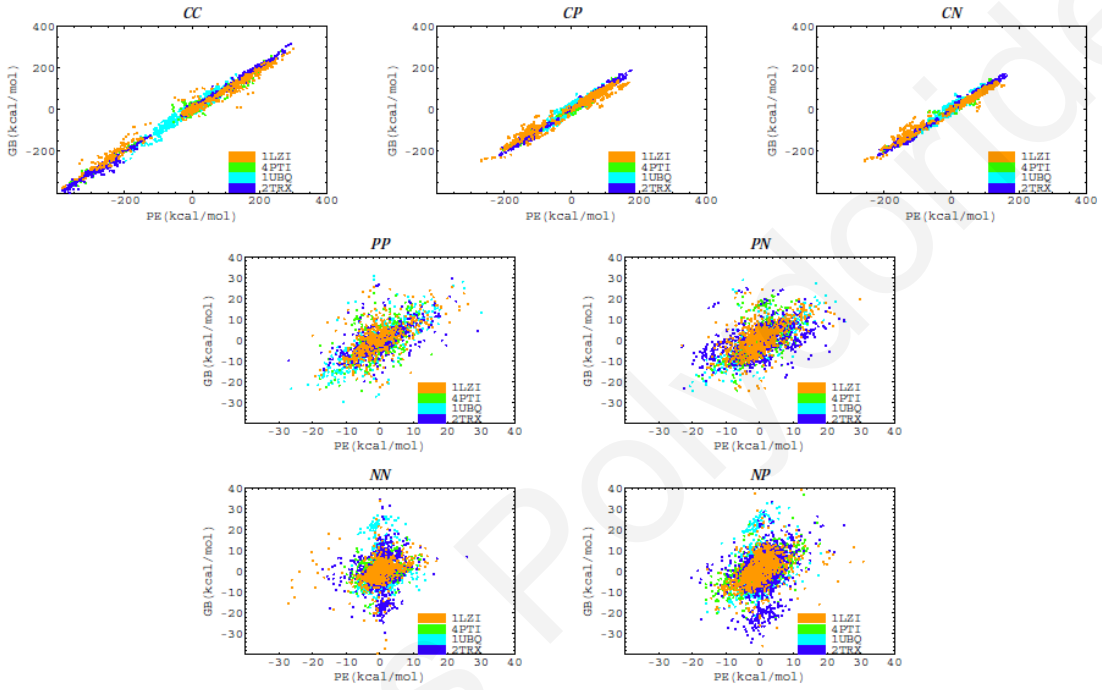


Figure 5.7: Decomposition of atomic GB/HCT and PE electrostatic solvation free energies according to one of the seven mutation types (see text). The atomic GB and PE free-energies of each mutant are centered, respectively, around the average atomic GB and PE value for all mutants of the same category and in the same rotamer. Points are colored according to the proteins, listed by their pdb codes (1LZ1, lysozyme; 1UBQ, ubiqutin; 2TRX, thioredoxin; 4PTI, bovine pancreatic trypsin inhibitor). The figure is taken from Ref. [181].

5.3 A Simplified Residue-GB Approximation Implemented in CPD

The residue-GB approximation analyzed in Section 5.1 is fully pairwise-decomposable; Nevertheless, it is still relatively laborious: to update the GB energy following a structure or sequence change, it is necessary to reevaluate GB solvation radii [coefficients B_R in Eq. (5.10)] for all residues and use them to compute GB interaction energies. Therefore, it is of interest to test a computationally simpler version of residue-GB, where the solvation radius of each sidechain is computed by assuming that the rest of the protein has its native sequence and conformation (see Fig. 5.8). This approximation is in the spirit of the residue-pairwise GB model of Handel, presented in Chapter 4. It permits the computation of the residue-GB solvation radii once, prior to the design.

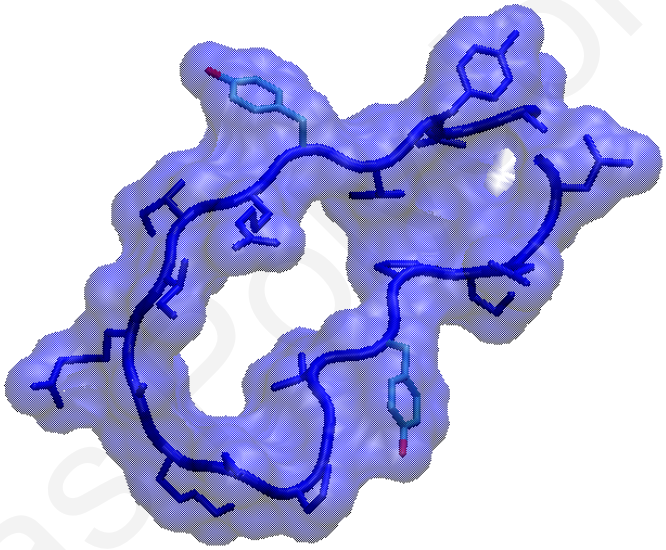


Figure 5.8: The simplified residue GB model, assumes for each residue pair the same environment, which corresponds to the native sequence / structure.

Prior to the design, atomic solvation radii are computed for each position R , by a direct evaluation of the residue self energy ΔG_R^{self}. Then, the corresponding residue-parameter (B_R) is computed, using Eq. (5.10). For each position R, all compatible sidechain / rotamers combinations (denoted as "chemical states" $\{k\}$) are considered. The resulting $\{B_R(k)\}$ values are stored for later use. All calculations are performed with the program XPLOR [183], using appropriate in-house scripts.

From this matrix, the GB interaction term between two residue segments at positions R and R$'$ and chemical states k and l, respectively, can be computed as follows: The appropriate residue-solvation radii $B_R(k)$, $B_{R'}(l)$ are recalled and assigned to each atom $i \in R$ and $j \in R'$ of the two residues respectively. The corresponding energy term is evaluated by Eq. (5.11) and stored.

Finally, the residue pair self and interaction energy terms are combined and saved in a matrix form, where each element represents the solvation energy contribution of a given pair of residues. The energy matrix is upper triangular and includes all possible sidechain rotamer combinations. For a protein with 10 residues, all allowed to take the twenty natural amino acid types, and five rotamer orientations per amino acid type, the number of possible states per residue is one hundred ($aa \times rotamers = 100$), producing 1.0×10^{20} sidechain rotamer combinations.

To compare the "exact" residue-GB and approximate residue-GB models, we plot the corresponding interaction energies against the value of atomic-GB, for all possible sidechain-total backbone and sidechain-sidechain pairs of the protein Asparaginyl - tRNA synthetase (AsnRS) in Figs. 5.9A-B. The protein is kept in its native AsnRS sequence and conformation. In this case, the approximate and exact residue self-energies are identical but the interaction terms are slightly different, resulting in RMS differences (residue-atomic) of 0.15 kcal/mol and 0.06 kcal/mol, respectively for the approximate and exact residue-GB approximations.

Mayo and coworkers developed a residue-pairwise approximation to the Poisson - Boltzmann model for CPD calculations [177] (Section 4.1.2). They use simplified molecular structures defined by the backbone and one or two sidechains, while all other missing sidechains are replaced by three generic spheres. For a set of 12 proteins in their native sequence and conformation, they reported similar RMS differences from exact PB (0.18 kcal/mol for sidechain-total backbone and 0.05 kcal/mol for sidechain - sidechain pairs; Fig. 5.10).

Fig. 5.9 C-D compare energies of sidechain-backbone pairs for 179 AsnRS sequences with random mutations in the five active positions (187, 190, 225, 227, 366). Each sequence is in a generally different, low-energy conformer. The RMSD between approximate or exact residue-GB and atomic-GB is 0.17 kcal/mol, slightly higher with respect to the native-AsnRS value (0.15 kcal/mol).

Fig. 5.9 E-F compare sidechain-sidechain interaction energies for a single, low-energy conformer of a designed AsnRS sequence, with substitutions Y187, S190, Q225, K227, S366 in the five active positions. The RMS difference from atomic GB is 0.07 kcal/mol for the exact-residue GB model and 0.27 kcal/mol for the approximate-residue GB model. Thus, the introduction of mutations increases somewhat the deviation from the atomic-GB values.

Pokala and Handel also use a simplified structure when calculating the atomic Born radii for a given sidechain. The self energy of each atom is computed in the presence of the residue of interest and the fixed backbone, following the pairwise stragety, but the missing sidechains are substituted with several dummy atoms which mimic the absent volume [71]. They use a large set of 26 protein structures to test the approximate GB model. Residue self energies and sidechain - sidechain interactions are in agreement to FDPB values, as shown in Fig. 5.11.

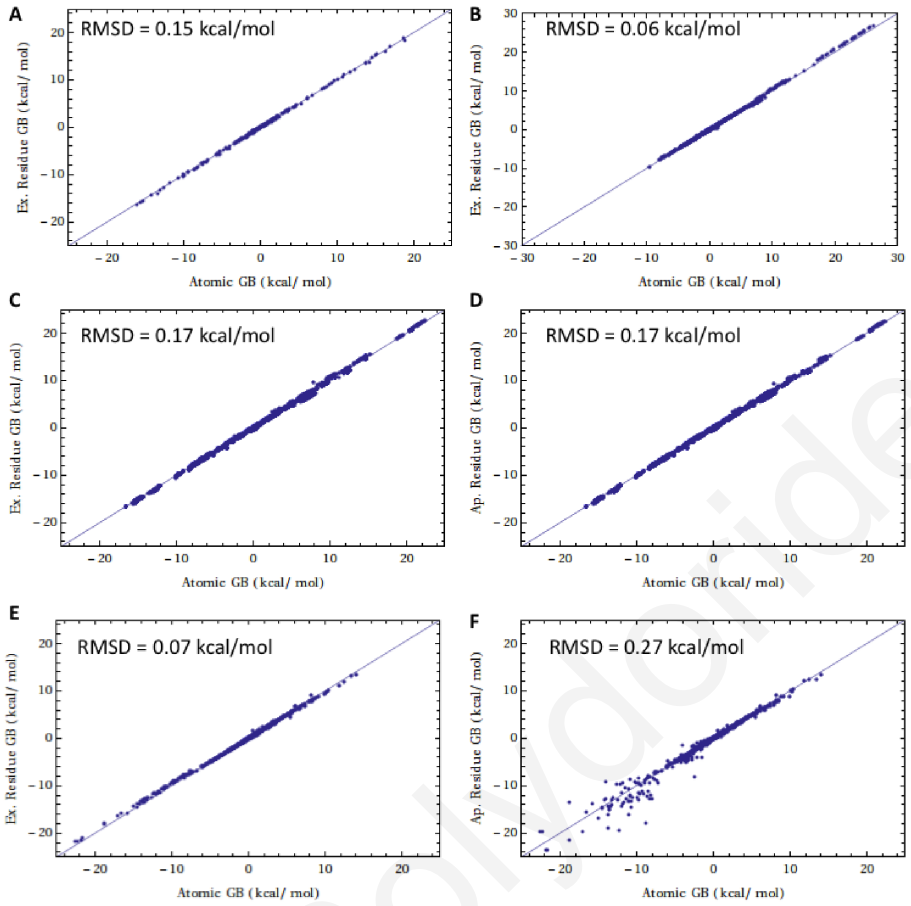


Figure 5.9: Accuracy of exact and approximate residue-GB, with respect to atomic-GB. (A) Residue-GB sidechain-total backbone and (B) sidechain-sidechain interaction energies, for the native AsnRS sequence and conformation; (C) Exact residue-GB and (D) approximate residue-GB sidechain-total backbone energies, for a set of 179 mutant AsnRS sequences; (E) Exact residue-GB and (F) approximate residue-GB sidechain-sidechain energies, for a designed AsnRS sequence.

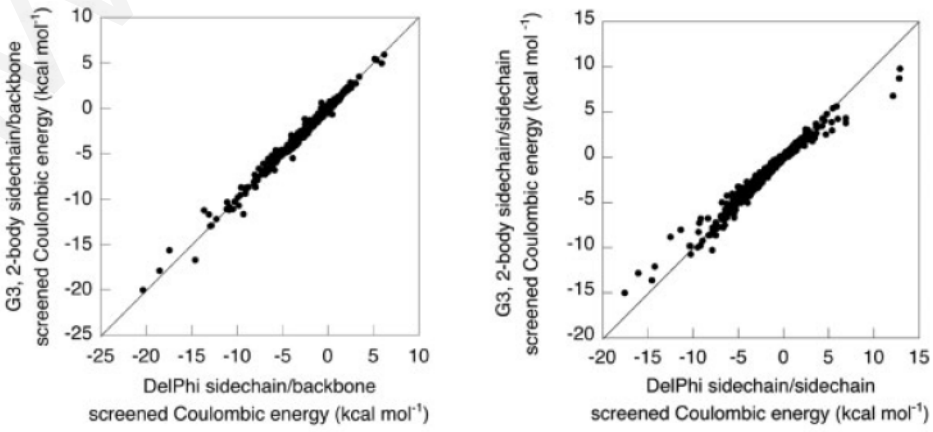


Figure 5.10: Sidechain - backbone (left) and sidechain - sidechain (right) interactions energies calculated by exact FDPB and the pairwise PB model (G3,2-body). Figure taken from Ref. [177].

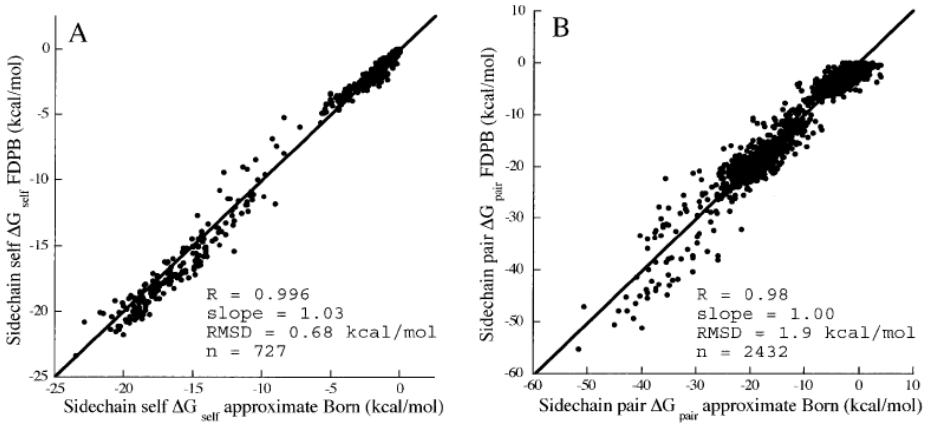


Figure 5.11: Sidechain self energy (A) and sidechain - sidechain interaction energy (B) calculated by FDPB and the approximate GB model proposed by Handel. Figure taken from Ref. [71].

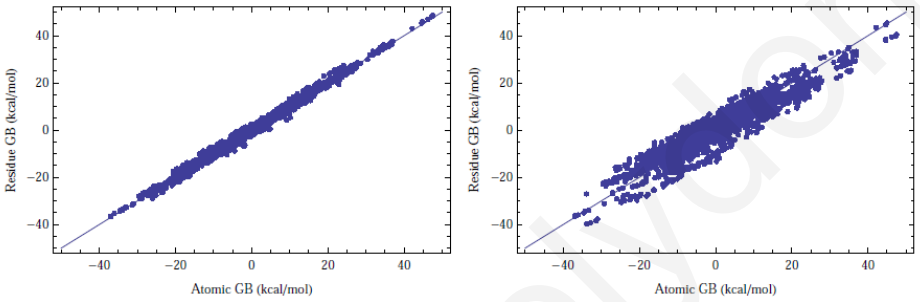


Figure 5.12: Residue-GB solvation energies for 1,800 AsnRS mutant sequences/conformations generated by Proteus, plotted against the corresponding atomic-GB values. Left: Original residue-GB approximation. Right: Simplified residue-GB approximation, used in the present design work.

Fig. 5.12 shows a comparison of the exact (original model) residue GB and the approximate (simplified model) residue GB for calculating the total solvation free-energies for a set of 1800 mutant AsnRS sequences and conformations. The RMSD between exact residue-GB and atomic-GB energies is 1.3 kcal/mol. When we switch to the simplified residue-GB implementation, the RMSD increases to 4.2 kcal/mol, but the good correlation between residue-GB and atomic-GB energies is retained.

Chapter 6

Comparison of CASA and GB Implicit Solvent Models

In this chapter we briefly outline the results of two recent CPD studies, which are directly relevant to the scope and subject of this thesis. The first study [174] compared the ability of CASA and GB models to reproduce free energy changes due to the introduction of conformational or chemical modifications in several proteins. From the results, it was concluded that an optimized GB model can yield better agreement with benchmark Poisson-Boltzmann calculations. The second study [77] employed a CASA implicit solvent model, to alter the amino acid specificity of the protein Asparaginyl-tRNA synthetase, the main system also studied in the present thesis. As explained below, the performance of the CASA treatment in this design problem was not satisfactory.

These studies supported the conclusion that it is worth to implement and test a GB model in CPD studies, despite its more intensive computational requirements. To achieve this, it is necessary to derive a pairwise-additive approximation to GB; this point is further elaborated in Chapter 5.

6.1 Computational Sidechain Placement and Protein Mutagenesis

Modifications in side chain rotamer conformations and chemical types are the two fundamental changes employed in CPD studies. Thus, parameterization and assessment studies of implicit solvent models for protein design need to handle both types of modifications accurately. The performance of CASA and GB for rotamer placement and side chain mutations was evaluated in Ref. [76]. Two GB models were employed; the "analytical continuum electrostatics" (GB/ACE) model of Schaefer and Karplus [154] and the "Hawkins-Crammer-Truhlar" (GB/HCT) model of Ref. [90]. The GB/ACE free energy is combined with the CHARMM19 polar-hydrogen energy function [62],

Table 6.1: RMS deviation between the CASA and GB energies and the PB benchmark solvation energies, for different rotamer conformations.

	CASA					
	dielectric constant ϵ					
Protein	1.0	1.5	2.0	2.5	20.0	GB/HCT
4PTI	46.9	32.5	32.8	35.7	55.6	20.7
2TRX	87.7	64.8	59.1	58.3	71.0	30.8
1LZ1	70.1	52.0	49.8	51.3	69.6	26.4

All values in kcal/mol. 4PTI = bovine trypsin inhibitor; 2TRX = thioredoxin; 1LZ1 = lysozyme. For each protein, the RMS values are computed over 750-1000 random rotameric structures (fixed backbone). The table is taken from Ref. [76].

whereas GB/HCT is combined with the AMBER all-atom energy function [138]. We only report results for the GB/HCT model, which has better performance [76], and is used in the main design studies of this thesis.

The authors computed solvation energies with CASA, GB and Poisson-Boltzmann (PB) calculations for four proteins: trypsin inhibitor (4PTI), thioredoxin (2TRX), lysozyme (1LZ1), and ubiquitin (1UBQ). To test the CASA and GB performance in reproducing solvation free energies of different rotamer conformations, they generated 750-1000 different structures for each protein, by randomizing side chain rotamers. Table 6.1 reports PB-CASA root mean square differences (RMSD) for a range of $\epsilon = 1 - 20$ values and PB-GB/HCT RMSD for $\epsilon_p = 20$, $\epsilon_w = 80$.

In the case of CASA, the single dielectric constant represents a uniform dielectric medium, with the average behavior of the protein and surrounding solvent; the PB-CASA RMSD ranges from 33-58 kcal/mol, for optimal values $\epsilon = 1.5 - 2.5$. Notice that the optimum ϵ varies with the protein. This represents a difficulty, as it would be desirable to employ a common (optimum) dielectric constant in routine CPD studies of several proteins. A consensus value $\epsilon = 20$ (optimal in charged mutations, as explained below) yields much larger RMSD values. The RMS deviation between GB/HCT and PB (last column) is 21−45 kcal/mol, significantly lower than that of CASA.

The CASA and GB/HCT total solution energies of the various rotameric structures are compared with PB in the top panel of Fig. 6.1. For a perfect agreement with the benchmark PB calculations, the values should fall on the diagonal. A much better correlation is obtained with GB/HCT, as can be seen from the figure.

An additional set of calculations compared the predictions of CASA and GB against PB for free-energy changes due to "artificial" charge mutations, in which (i) a net charge ± 1 was removed from Arg, Lys, Asp, Glu side chains, (ii) a net charge ± 1 was added on Ala, Ile, Leu, Val, Met, Pro, Thr or Tyr side chains, or (iii) the polarity of a side chain was modified (Asn, Gln or singly-protonated His side chains were made apolar, or a dipole was introduced on the Cys side chain).

The RMSD PE-CASA values are listed in Table 6.2. With optimum dielectric

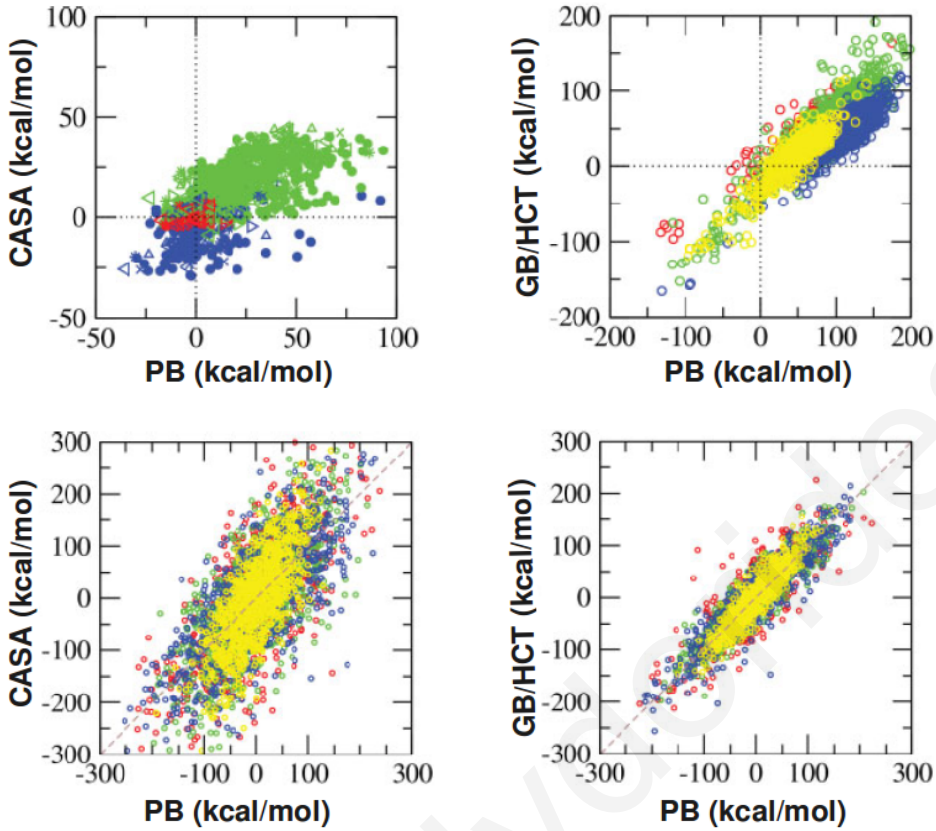


Figure 6.1: Top panel: Folding free energy changes due to charge mutations with the CASA (left) and GB/HCT implicit solvent model (right). Bottom panel: Solvation energies of randomized rotameric structures with the CASA (left) and GB/HCT implicit solvent model (right). Results for different proteins are included with different colors and symbols. ASPRS = Aspartyl tRNA synthetase; 1LZ1 = lysozyme; 1UBQ = ubiqutin; 2TRX = thioredoxin; 3NR3 = RNAse A; 4PTI = bovine trypsin inhibitor; 5CYT = cytochrome C. The plot is taken from Ref. [76].

constants 16-20, the RMSD ranges from 17.5 kcal/mol for the largest protein (AspRS) to 10.0−11.0 kcal/mol for the smallest proteins (ubiquitin and BPTI). The RMSD between GB/HCT and PE is 8−16 kcal/mol, smaller or comparable to CASA. Notably the success rate (defined in the footnote to this table) is much higher with GB. The CASA and GB/HCT total solution energies are compared with PB in Fig. 6.1. The correlation between GB/HCT and PE is excellent, far superior to CASA.

From these results, it follows that the GB/HCT model of Ref. [76] yields in general better predictions with respect to CASA, for rotamer transitions and charge mutations.

Table 6.2: RMS deviation between the CASA and GB energies and the PB benchmark solvation energies, for charge mutations.

Protein	No of mutations	PE-CASA	Success rate	PE-GB/HCT	Success rate
AspRS	518	17.5	80.1		
3RN3	106	13.1	79.0		
4PTI	51	11.0	74.4	8.3	99.2
5CYT	84	13.5	79.8		
2TRX	94	12.3	80.8	12.1	98.5
1LZ1	108	12.2	83.4	16.5	98.2
1UBQ	67	10.4	86.5	10.8	92.3

All values in kcal/mol. For each protein, 1,000 charge mutations are introduced. The optimum ϵ values for CASA are 16-20. The success rate reflects the percentage of mutations predicted to have a positive or negative stability change by both PB and CASA or GB/HCT. The table is taken from Ref. [76].

6.2 Modification of Asparaginyl - tRNA Synthetase Specificity for its Natural Amino Acid: CPD with a CASA Implicit Solvent Model.

The second study outlined in this chapter was reported in Ref. [77]. It employs a CASA model to alter the amino acid specificity of the protein Asparaginyl-tRNA synthetase, that is the main system also studied in the present thesis. As explained in Chapter 8, the protein asparaginyl-tRNA synthetase (AsnRS) has specificity for the polar but neutral amino acid asparagine (Asn). In Ref. [77], Simonson and coworkers attempted to introduce specificity for the negatively charged amino acid aspartic acid (Asp), using a physical free energy function combined with the CASA approximation for solvent effects. In what follows we overview the employed methodology and results, and discuss the influence of the energy function and specifically of the solvation term on the quality of the designed sequences. A similar design was attempted by us with the "residue GB" model, and is presented in the next chapter. Comparison of the predictions by the two studies show that the GB model improves considerably the results.

6.2.1 Methodology

Simonson and coworkers employed a stability criterion (Section 7.2.1) in conjuction with the Wernisch algorithm (Section 7.1.3); their algorithm searched for sequences which minimized the folding free energy of the complex between AsnRS and the non-cognate ligand Aspartyl-adenylate (AspAMP). Protein and ligand interactions were described by the CHARMM19 polar-hydrogen energy function [62], which has the

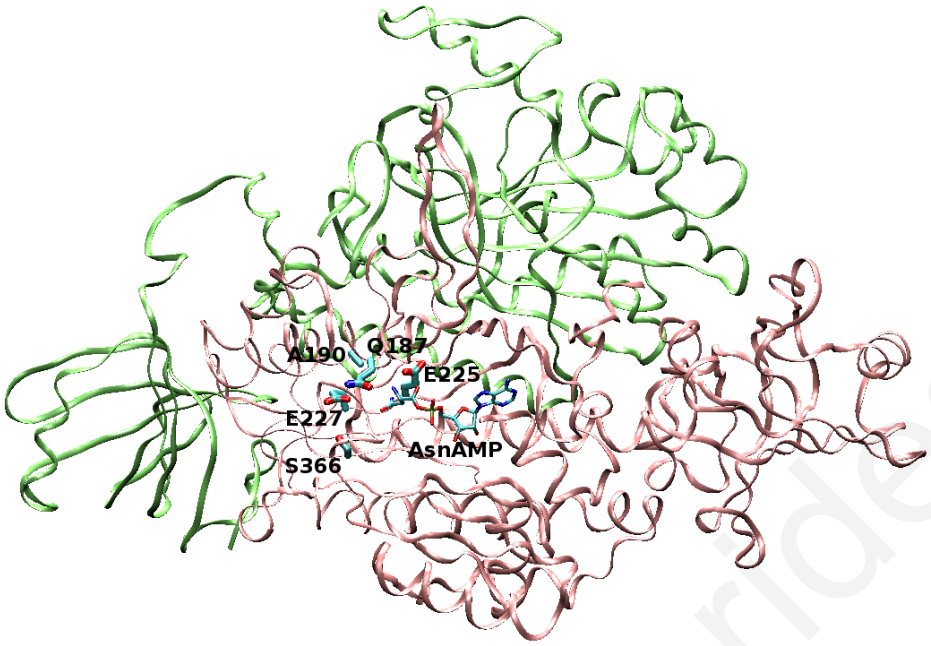


Figure 6.2: The Asparaginyl-tRNA synthetase, in complex with the AsnAMP ligand. The five active positions are indicated with labels.

functional form of Eq. (3.1). The CHARMM19 contains an explicit representation of all atoms, with the exception of hydrogens covalently bonded to non-polar carbons, which are united with their adjoining carbons [62]. Solvent effects were treated with the CASA approximation:

$$E_{\text{solv}} = E_{polar} + E_{\text{nonpolar}} = (\frac{1}{\epsilon} - 1) \sum_{i<j} \frac{q_i q_j}{r_{ij}} + \alpha \sum_i \sigma_i A_i, \tag{6.1}$$

where ϵ is the dielectric constant (of the homogeneous protein/solvent medium), A_i is the solvent-accessible surface area of atom i and σ_i are atomic solvation parameters, which describe the tendencies of various atom types to be buried (hydrophobic) or exposed to water (hydrophilic). The optimized values $\sigma = -0.08$ kcal/mol/Å^2 for polar atoms, -0.10 kcal/mol/Å^2 for ionic atoms, -0.04kcal/mol/Å^2 for aromatic atoms, -0.005kcal/mol/Å^2 for non-polar atoms, and 0.0kcal/mol/Å^2 for hydrogen atoms, were employed [174]. The non-polar term entering in Eq. (6.1) was precomputed and assigned to residue pairs with the pairwise-additivity approximation of Eq. (4.3)[160]. A global scaling factor $\alpha = 0.5$ was employed, to correct overcounting of the computed buried surface areas, due to this assumption. [160].

Five positions in the vicinity of the AsnAMP binding site (residues 187, 190, 225, 227 and 366 in the numbering of *Thermus thermophilus* AsnRS) were designated as "active", i.e. were allowed to change both chemical type and side chain orientation (Fig. 6.2). In the original AsnRS sequence, these positions had, respectively, the chemical types glutamine (Gln187), alanine (Ala190), glutamic acid (Glu225), glutamic acid (Glu227) and serine (Ser366). Four of the five positions are highly conserved in AsnRS

sequences; position 190 is more variable, being occupied by a variety of hydrophobic residues in natural AsnRSs. In AspRS, Ala190 is occupied by a conserved lysine (Lys198) .

Other positions were designated as "inactive" (i.e. retained their chemical type but could explore different orientations from the Tuffery rotamer library [135]), with the exception of glycines, prolines and histidines, which were kept "fixed" at the chemical type and orientation of the native structure. The protein backbone was also kept fixed. In the case of the AspAMP ligand, the aspartic acid side chain moiety was allowed to explore rotamers from the same rotamer library; the remaining part of the ligand was kept fixed at the crystallographic position.

6.2.2 Results

The five mutatable positions have a net charge of -2 in the native protein, due to the two negatively charged residues Glu225 and Glu227. Table 6.3 lists the most promising sequences resulting from the design. Since the goal is to identify sequences with improved affinity for the negatively charged aspartic acid, it would be reasonable to expect that the design would favor the removal of negatively charged residues and/or the insertion of positively charged residues near the target ligand. This was partly accomplished by the design: The negatively charged residues Glu225 and Glu227 were frequently mutated to polar residues (serine, glutamine, asparagine), neutral (alanine) and sometimes positive (lysine). However, a negatively charged residue (aspartic acid or glutamic acid) was also most of the time introduced at position 187 (where it interacted with the positively charged ammonium group of the ligand) and/or position 366 (where it interacted with the positively charged side chain of Arg368). As a result, several of the designed sequences retained the net negative charge (-2) of the original sequence (table 3).

In the active site of the protein AspRS, the five positions homologous to the "active" AsnRS positions considered here, are, respectively, Gln187, Lys198, Asp233, Glu235 and Ser487 (in *E. coli* numbering) Fig. 8.16 and 8.18. Even though AspRS is specific for a negative amino acid (aspartic acid), it tolerates two conserved negative residues in the vicinity of the ligand (Asp233 and Glu235). This can be explained in terms of the AspRS active site: Asp233 makes a direct electrostatic interaction ("salt bridge") with the positively charged Lys198 and Glu235 makes a similar direct interaction with the positively charged residue Arg489. Both residues Arg489 and Ly198 make key interactions with the Asp ligand side chain; the positioning of the oppositely charged residues Glu235 and Asp233 near them serve to reduce these interactions, as shown by continuum electrostatics calculations [142].

The elimination of a negative charge at position 225 of AsnRS (homologous to AspRS 233), that is often predicted by the CASA design (Table 6.3), can be explained by

Table 6.3: Binding free energies for designed AsnRS mutant sequences

	Sequence						AspAMP binding		AsnAMP binding	Asp-Asn difference
	187	190	225	227	366	Charge	CASA	PB	PB	PB
AsnRS	Q	A	E	E	S	-2	-79	-104.8	-116.4	+11.6
	D	A	A	N	D	-2	-106	-100.4	-100.9	+0.5
	L	A	A	A	D	-1	-108	-98.9		
	E	A	S	A	E	-2	-108	-107.8	-81.1	-26.7
	E	A	A	K	D	-1	-108	-100.8	-91.1	-9.7
	D	A	A	A	D	-2	-110	70.9		
	E	A	S	A	A	-1	-110	102.0	-100.8	-1.2
	D	A	S	M	D	-2	-111	-97.1		
	E	A	A	A	A	-1	-113	-82.0		
	D	K	M	M	D	-1	-108	-110.9	-99.1	-11.8
AspRS	Q	K	D	E	S	-1				-18.0

"QAEES" is the native sequence of Asparaginyl-tRNA synthetase (AsnRS) in the active positions 187, 190, 225, 227, 366; "QKDES" is the corresponding sequence for Aspartyl-tRNA synthetase (AspRS). The "Charge" column reports the net total charge of the five active positions. The columns "AspAMP binding" and "AsnAMP binding" refer to association free energies of the corresponding protein-ligand complexes, evaluated with CASA or with a Poisson-Boltzmann (PB)model. Values are averaged over snapshots, extracted from MD trajectories of the corresponding complexes. These snapshots are post-processed by PB, as explained in section 2.1.2. All values in kcal/mol. The table is adapted from Ref. [77].

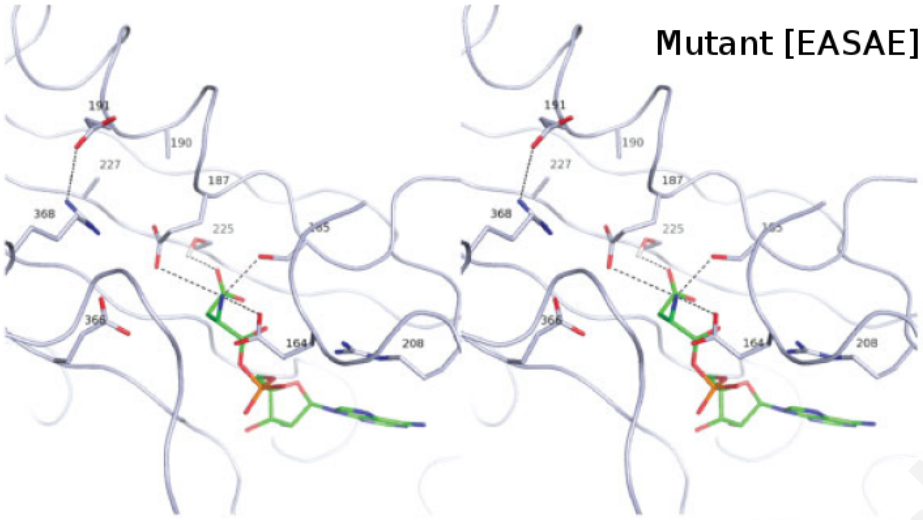


Figure 6.3: Stereoscopic view of the AsnRS active site with the most important protein-ligand interactions. The native sequence Gln187/ Ala190/ Glu225/ Glu227/ Ser366 (QAEES) is mutated to Glu187/ Ala190/ Ser225/ Ala227/ Glu366 (EASAE). The structures are extracted at the end of 2-ns all-atom MD simulations of the AsnRS:AspAMP complex. The figure is taken from Ref. [77].

the absence of a positively charged lysine at position 190 (homologous to AspRS 198). However, the negative-charge elimination at position 227 (homologous to AspRS 235) perturbs the stability of the Arg368 side chain and affects the electrostatic interaction balance in the vicinity of the Asp ligand. In fact, the negative charge at position 227 (glutamine) is in most designed sequences moved to position 366 (aspartic or glutamic acid). Structural analysis of the designed mutants showed that this charge displacement causes the Arg368 side chain to rotate and increase its interaction with Asp or Glu366. Furthermore, in native AsnRS the side chain of Gln187 interacts with the positively charged ammonium group of the ligand (which also interacts with conserved residues Glu164 and Ser185). In the designed sequences, it is usually mutated to Asp or Glu, as this increases even more the interaction with ammonium.

Molecular dynamics simulations with several of the designed AsnRS:AsnAMP complexes showed that these negative-charge insertions at positions 187 and 366 destabilized the AsnRS:AspAMP complex, causing significant structural distortions in the binding site and the loss of important protein-protein and protein-ligand interactions. Some of the designed sequences manifested preferential binding towards AspAMP over AsnAMP, but their binding affinities were much weaker, with respect to the affinity of native AsnRS for Asn. Fig. 6.3 shows the binding site of the complex between designed sequence Glu187/ Ala190/ Ser225/ Ala227/ Glu366 (sequence "EASAE" in fourth row in Table 6.3) and AspAMP, at the end of a 2-ns MD simulation. The ligand side-chain has rotated too far away, and the interaction with Arg368 is eliminated; Instead, Arg368 forms a new interaction with Asp191.

In all designed sequences, position 190 contained an alanine residue. The homol-

ogous position in the active site of the protein Aspartyl-tRNA synthetase (which is specific for aspartic acid) contains a positively charged residue (Lys198, in the numbering of *E. coli* AspRS), which interacts directly with the ligand side aspartic acid chain and the key residue Arg489 (homologous to AsnRS Arg368). The design introduced invariably an alanine residue at the same position. To improve the resulting sequences, additional CPD calculations were conducted in Ref. [77], with an invariant ("active") lysine at position 190. Furthermore, the orientation of the ligand side chain (AspAMP) was forced into a conformation that was analogous with the one seen in the AspRS:AspAMP complex. The most promising sequences from this design contained Asp187/ Lys190/ Meth225/ Mett227/ Asp366, Asp187/ Lys190/ Gln225/ Ala227/ Asp366, or Glu187/ Lys190/ Glu225/ Ala227/ Ala366 chemical types at the five positions.

MD simulations were performed on selected sequences and the association free energies of the AspAMP complexes with native and mutant AsnRS were computed by the end-point MM/PBSA method (Section 2.1.2). Two of the mutant sequences (Glu187/ Ala190/ Ser225/ Ala227/ Glu366 and Asp187/ Lys190/ Met225/ Met227/ Asp366) had stronger association energies with the AspAMP ligand (respectively, -107.8 kcal/-mol and -110.9 kcal/mol), compared to the native protein (-104.8 kcal/mol). For other three sequences (Asp187/ Ala190/ Ala225/ Asn227/ Asp366, Glu187/ Ala190/ Ala225/ Lys227/ Asp366 and Glu187/ Ala190/ Ser225/ Ala227/ Ala366), the AspAMP binding affinity was weaker by 2 - 4 kcal/mol and for the rest it was worse.

For the best five sequences, the binding affinity for the native ligand (AsnAMP) were also computed. The resulting relative affinity differences (AspAMP - AsnAMP) were negative. Nevertheless, all designed sequences had weaker binding affinities for AspAMP, compared to the affinity of the original AsnRS protein for its native ligand AsnAMP (-116.4 kcal/mol). Thus, even though the design was able to destabilize the binding of the original ligand (AsnAMP) with respect to the target ligand (AspAMP), it was not able to identify sequences with AspAMP affinity that was as strong, as the affinity of native AsnRS for its own ligand.

Preserving the binding location and orientation of the ligand side chain is also important, in order to maintain the functionality of the protein. Only the mutant Asp187 / Lys190 / Met225 / Met227 / Asp366 was predicted to have a similar active-site conformation as in the AspRS:Asp complex, and only after fixing the Lys198 orientation (Fig. 6.4).

6.2.3 Conclusions

The above results suggest that the CASA approximation does not give very satisfactory design predictions, even with an optimized parameterization of the solvation coefficients. The CASA model is computationally efficient, but has limited accuracy

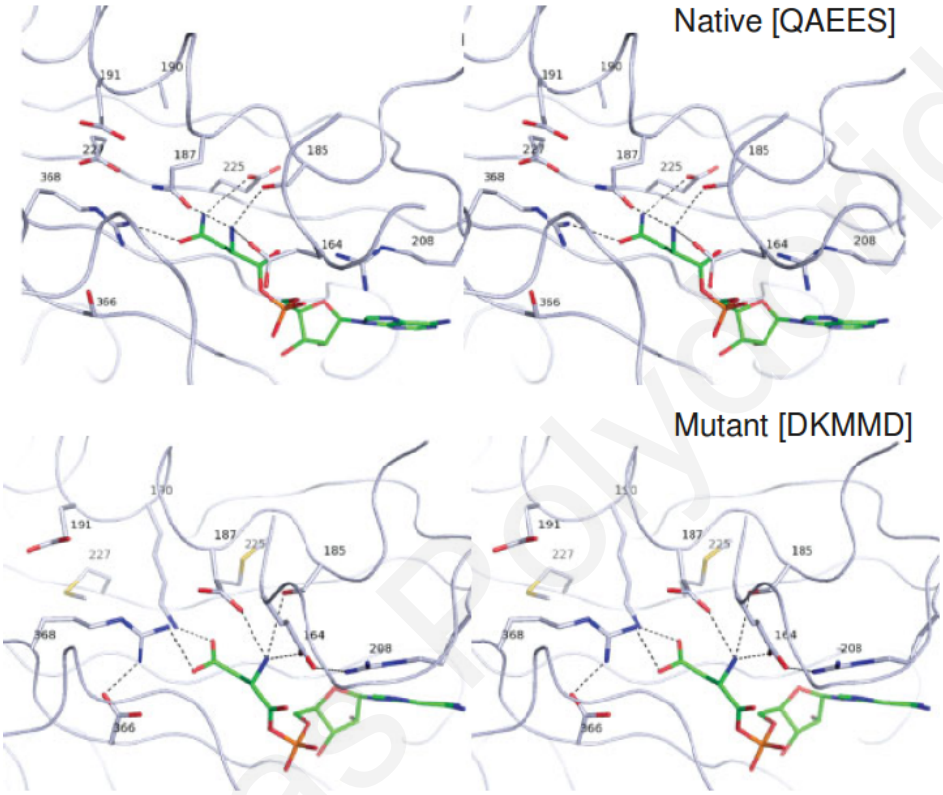


Figure 6.4: Stereoscopic view of the AsnRS active site with the most important protein-ligand interactions. The native complex AsnRS:AsnAMP (top panel) is compared to the most promising designed mutant (bottom panel). The native sequence Gln187/ Ala190/ Glu225/ Glu227/ Ser366 (QAEES) is mutated to Asp187/ Lys190/ Met225/ Met227/ Asp366 (DKMMD). The structures are extracted at the end of 2-ns all-atom MD simulations of the AsnRS:AspAMP complex. The figure is taken from Ref. [77].

for buried mutants. This is partly because CASA represents the protein/solvent as an average uniform dielectric medium, with a common constant ϵ, and ignores interactions between charges and their own reaction potentials (i.e. it does not have the "self-energy" terms of the GB approximation [Eq. (4.5)], [76]. As showed above (Section 6.1), CASA is not as accurate as GB when calculating solvation energies of protein structures composed by random rotamers or stability changes caused by charge mutations. To distinguish among sequences with affinity differences of a few kcal/mol, the energy function must be relatively accurate. A better solvation model is very likely to improve this predictions.

Apart from the quality of the polar-hydrogen/CASA energy function, another concern is the use of a simple stability criterion to improve the AsnRS affinity for a new ligand. The reason is twofold: (i) the stability design is based on the minimization of the difference in free energy between the folded and unfolded state (Section 7.2.1). The folded state employs a reasonable structural model (the experimental structure of the native complex); on the other hand, the unfolded state is modeled as a superposition of non-interacting peptides, associated with specific unfolded-state free energies (Section 9.2.2). The accuracy of the estimated folding free energies (and hence the quality of the designed sequences) depends crucially on these reference energies. (ii) Even if the designed sequences are correctly predicted to have very negative folding free energies, it is not guaranteed that they will also have very negative association free energies (Section 9.2.6). Therefore, a better criterion is needed, that will select sequences with strong affinity. The main goal of the work presented in this thesis is to deal with these issues. In the next chapter, we implement a residue-pairwise treatment of the generalized Born model, and use it to redesign the amino acid specificity of AsnRS. We employ three different criteria (stability, absolute and relative affinity) and show that the design is significantly improved, compared to the CASA predictions of Ref. [77].

Chapter 7

Implementation of Design Criteria into the CPD Program Proteus

As explained in Chapter 3, high-throughput CPD methods employ deterministic or heuristic algorithms to search the large conformation/sequence space of a protein or protein-ligand complex. In this Chapter, we outline the CPD program Proteus [184], which has been used in the design work of the present thesis. Proteus is based on a heuristic algorithm introduced by Wernisch and coworkers [73]. It is written in C/C++ code. The original program identified sequence/conformation combinations of high stability (low free energy). In our work, we have implemented and employed additional searching criteria, which identify sequences/conformations that maximize the affinity of a protein for a specific molecule, or the *relative* affinity between two molecules. In what follows, we outline the organization of Proteus, explain the various searching criteria and our novel implementation, and present design results with the protein Asparaginyl-tRNA synthetase.

7.1 The Proteus Program

In the CPD methodology implemented in Proteus, a protein (or protein complex) is subdivided into three groups: a "fixed" part, which maintains its conformation and chemical identity during the design, an "inactive" part, which is allowed to change conformation, and an "active" part, which changes both conformation and chemical identity. In standard applications, the "fixed" part corresponds to the protein main chain and a subset of side chains (e.g. prolines, glycines and cysteines in the native sequence). The "active" part consists of selected side chains, chosen by their location in the structure and by the purpose of the design. For example, if the design aims to increase the binding free energy of a specific molecule onto the protein, active residues can be chosen near the binding site. The "inactive" part consists of the remaining side chains.

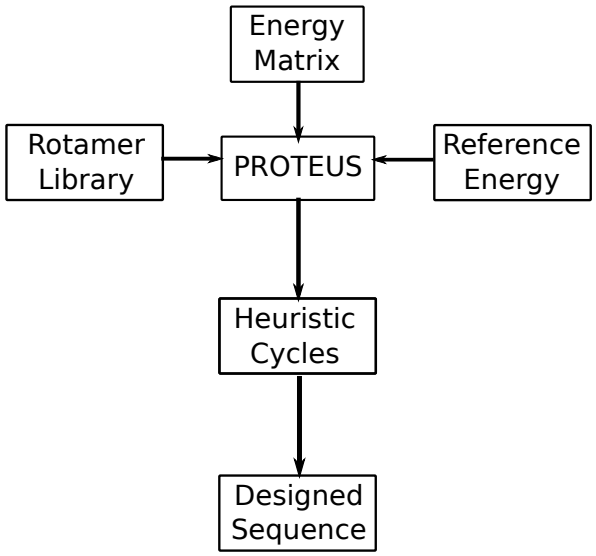


Figure 7.1: Flowchart of the design procedure implemented in Proteus.

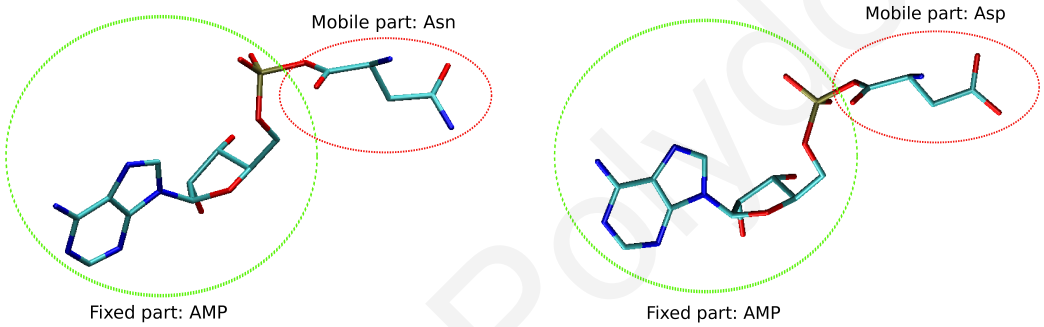


Figure 7.2: The two ligands AsnAMP (left) and AspAMP (right) consist of a fixed part (enclosed in the green elipse) and a variable ("inactive") part, corresponding to the amino acid moiety. The variable part has two main chain and two side chain torsional dihedral angles, assuming a total of 161 distinct rotameric conformations [135]).

The flowchart of Fig. 7.1 illustrates the standard sequence-optimization procedure implemented in Proteus. In the beginning, the program reads a list with the chemical types accessible to active aminoacids and the number of corresponding rotamer conformations. This list is flexible and depends on the particular design problem. The chemical types / rotamers employed in our work is shown in Table 7.1. Prolines, glycines and cysteines are not included because they remain invariant during the design (they belong to the "fixed" part, defined above); a histidine side chain is represented by two states [neutral (HIE) or protonated (HIP)]. An active residue is accessible to a total number of 216 "chemical states" (chemical types/rotamers).

The ligand AspAMP (last row in Table 7.1) is accessible to 161 rotameric conformations. The structure of AspAMP and the cognate ligand AsnAMP is shown in Fig. 7.2. The rotamers result from variation of the main-chain and side-chain torsional angles in the amino acid moiety.

Table 7.1: List of amino acids / rotamers employed in the CPD study of asparaginyl-tRNA synthetase with Proteus.

Amino acid name	One-letter code	Number of rotamers
ALA	A	1
ASP	D	5
ASN	N	11
ARG	R	39
GLU	E	12
GLN	Q	19
HIE	H	9
HIP	H	9
ILE	I	7
LEU	L	9
LYS	K	49
MET	M	17
PHE	F	4
SER	S	3
TYR	Y	8
THR	T	3
TRP	W	8
VAL	V	3
LIG	-	161

The listed number of rotamers corresponds to the Tuffery rotamer library [135]. HIE and HIP denote, respectively, a neutral and protonated histidine side chain. "LIG" is the Aspartyl-adenylate (AspAMP) ligand.

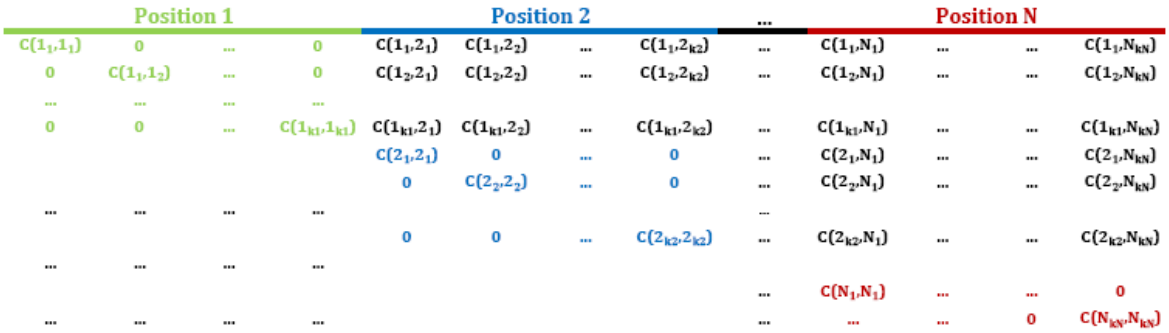


Figure 7.3: The form of the interaction-energy matrix. In the example shown the system has N "active" or "inactive" positions. Each position $l \in \{1, N\}$ is compatible with l_{N_l} chemical states (chemical types/conformations); if a position is "inactive", l_{N_l} (the number of rotamers compatible with the chemical type of l); finally, if it is "active", $l_{N_l} = 216$ (Table 7.1). Colored fonts indicate "diagonal" elements (see text).

7.1.1 The Interaction Energy Matrix

The interaction-energy matrix has the form shown in Fig. 7.3. It contains N positions, classified as "active", or "inactive" (contributions from the "fixed" part are absorbed into the diagonal elements, as explained below). Each position l is compatible with a total of l_{N_l} chemical states (chemical types / conformations); l_{N_l} is equal to 216 for "active" positions (Table 7.1) and $N_{\rm rot}[j(l)]$ for "inactive" (the number of rotamers compatible with chemical type j of position l). The colored fonts indicate "diagonal" elements of interactions between atoms at the same position. For example, element $C(1_1,\ 1_1)$ contains the interaction energy terms of atoms within the side chain at position 1 and "chemical state" 1; if position 1 is "inactive" with chemical type j, chemical state "1" is the first rotamer of chemical type j; if it is "active", chemical state "1" is the first rotamer of the first chemical type in the list of all 216 possible chemical types. Obviously, atoms of position-1/chemical state-1 will never interact with atoms at the same position and a different chemical state k ($k \neq 1$). Thus, all "diagonal" elements $C(1_k,\ 1_l)$, $k \neq l$ are set to zero.

The "fixed"part of the protein acts like an external field; its contributions are included in the diagonal elements. For example, element $C(1_1,\ 1_1)$ contains also the interaction between atoms inside the side chain at position-1 / chemical state-1, with the "fixed" part of the system (e.g. the entire protein backbone and all "fixed" side chains).

In our work, the interaction-energy calculation is done by the program XPLOR [183], using in-house input files (Appendix A). The elements of the resulting interaction energy matrix are manipulated by in-house PERL scripts; subsequently, they are fed into the program Proteus, which performs the actual design calculation (the search in

sequence / conformational space), using the criteria analyzed in the previous sections. For a given sequence/conformation, corresponding to a series of chemical states $\{l(1), \ldots, l(N)\}$, the total energy of the system is computed by summing up the appropriate elements in the interaction energy matrix:

$$E[l(1), \ldots, l(N)] = \sum_{i=1}^{N} C(i_{l(i)}, i_{l(i)}) + \sum_{i<j} C(i_{l(i)}, j_{l(j)}) \tag{7.1}$$

During the design, if a mutation changes a chemical state at position k, $[l(k) \longrightarrow l'(k)]$, the energy is updated as follows:

$$E' = E + [C(k_{l'(k)}, k_{l'(k)}) - C(k_{l(k)}, k_{l(k)})] + \sum_{i \neq k} [C(k_{l'(k)}, i_{l(i)}) - C(k_{l(k)}, i_{l(i)})] \tag{7.2}$$

Inspection of Eq. (7.1-7.2) reveals the computational efficiency offered by the use of matrix C: The original all-atom energy has $N_{\text{atom}} \times (N_{\text{atom}} - 1)/2$ terms. The total energy from matrix C [Eq. (7.1)] has $N \times (N + 1)/2$ terms, with N the number of positions ($N \approx N_{\text{atom}}/20$). The energy *update* [Eq. (7.2)] has only $\propto N$ terms.

7.1.2 The Heuristic Search in Protein / Sequence Space

CPD protocols reduce the conformational space into a set of discrete, enumerable states, by fixing part of the protein (e.g. the backbone and selected sidechains), and allowing the remaining sidechains to take a discrete number of conformations from a rotamer library. The remaining discretized sequence / conformation space is explored efficiently by using suitable algorithms; one such method is the dead-end-elimination (DEE) algorithm. This algorithm, first described in Section 3.2.1, identifies and discards rotamers which increase the energy of the sequence with respect to others. This procedure is based on the observation that if a better rotamer r' than r can be found at position i, rotamer r cannot be part of the optimum solution. The algorithm proceeds by excluding these energetically unfavored rotamers from the search. At the end, only the optimum solution survives the elimination process. This scheme can be adapted to consider pairs of rotamers; a pair residue threshold energy is applied to identify and exclude disparate pairs. Here the algorithm can eliminate a pair of (simultaneously existing) rotamers, without neccessarily excluding each single rotamer. Nevertheless, a combination of the two approaches is usually employed. The DEE method decreases the computational demands of the space exploration and is oftenly preferred for providing the global minimum solution; it is still computationally (time demanding) inferior compared to other heuristic procedures [73; 185].

The branch-and-bound algorithm can be applied after the DEE method to accelerate the identification of the optimum sidechain rotamer at each position. Fig. 7.4

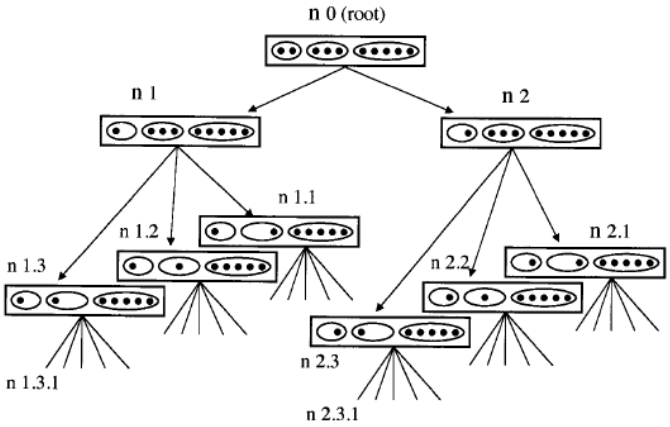


Figure 7.4: A tree diagram representation of the Branch & Bound algorithm. The figure is taken from Ref. [73].

illustrates the basic concept of this algorithm. The parent node of the tree n0 represents the initial state, with groups of all available rotamers at each position. The daughter nodes represent successive stages in the design process produced after selecting a rotamer at a given position. For example, if the sidechain at position i has two available rotamers, the daughter nodes n1, n2 adopt rotamers 1 and 2 respectively. Moving on to the next stage, the daughter nodes at the second level of the tree correspond to the rotamers of the sidechain at the next position i+1 given that sidechain at position i is positioned either in rotamer 1 or 2. The procedure is repeated until the leaves at the end of the tree are reached; these represent all possible rotameric states of the problem specifying a single rotamer for each position of the sequence. Searching the leaves of the tree to identify optimum cantidates is of no practical use, since the B&B algorithm is used to overcome the exausting search of a large combinatorial space. The idea is to prune branches of the tree with high energy rotamers, which prevent finding an optimum solution in the leaves of the given branches. To prune these undesirable branches, the algorithm defines a minimum cutoff value of the energy, just above the global minimum. Eliminating daughter nodes from the tree by identifying the worst rotamers at each position reduces the search space significantly.

An exact optimization procedure like DEE would always provide the unique optimum solution (if converged), along with a large number of compatible sequences identified within a small energy window above the best scoring sequence. However, most current CPD procedures prefer to use faster heuristic algorithms, which accelarate the procedure by 2–5 orders of magnitude [73]. The number of high scoring sequences produced by such heuristic methods depends strongly on the number of iterations performed; although missing the best sequence is possible, using hundred of thousands of iterations improves the reliability of these methods.

7.1.3 The Wernisch Algorithm

Wernish and coworkers [73] proposed a simplified heuristic algorithm, which identifies sequences/rotamer conformations of low folding free energy. This algorithm was briefly mentioned in Section 3.2.2. Here we present it in more detail.The algorithm initially places all "inactive" residues of the protein in random conformations and all "active" residues in random chemical states/conformations. Starting from a given position i, the algorithm loops over the subset of chemical states (chemical types and/or rotamers) consistent with position i. For each possible state k, all pair interactions of position i with the rest of the molecule are extracted from the interaction matrix and used to update the total free energy of the folded state [Eq. (7.1-7.2)]. The energy of the unfolded state is updated in a similar way (if the chemical composition of the sequence has been modified; see below), and subtracted from the energy of the folded state. After considering all states, the algorithm selects for position i the state (chemical type and/or rotamer) that minimizes the folding free energy; this is in accordance with the DEE procedure, described earlier. The algorithm examines the next position $(i + 1)$ and goes through the same procedure; an elementary cycle ends when all positions have been examined. The new sequence / rotamer combination replaces the initial random choice, and the cycle is repeated several times to identify possible improvements (the convergence is not immediate, because each time a position i is visited, its environment differs; thus it is probable that each time a different optimum sidechain and/or rotamer will be selected for i). Eventually, the heuristic search ends when the algorithm converges to a sequence/conformation or if a maximum number n of improvement cycles (typically, n=500) is reached. A set of sequence / structure optimization cycles starting from a specific initial random state is called "heuristic cycle".

For an adequate sampling of the sequence and/or conformation space, Proteus executes a large number [O(10^6)] of heuristic cycles, starting from different random states. The sampling is improved and the probability to identify the global minimum increases with the number of heuristic cycles; the optimum number depends on the protein size and the number of mutatable positions.

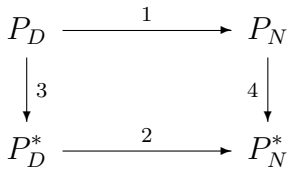
7.2 The Design Criteria

Optimization of a sequence/conformation can be done with a variety of criteria, depending on the goal of the design. The *stability criterion* optimizes the folding free energy of a protein that is constrained in a specific fold; i.e. it selects sidechains and rotamers that render as negative as possible the difference in the free energy between the folded and unfolded state, as described in the next section. This method can be used to create hyperstable proteins with a given structure, or to search for sequences that stabilize a novel structure. Variants of the method can be used to stabilize a

particular fold with respect to a second. In many applications, the aim of the design is to identify sequences and conformations that maximize the formation free energy of a particular complex (e.g. a protein and a small molecule, the "ligand"). This is the criterion of *absolute affinity*. Alternatively, the design may seek to introduce specificity by optimizing the preference of a protein for a given ligand, relative to a second ligand; this is the *relative affinity* criterion. These criteria are further explained below.

7.2.1 The Maximum Stability Criterion

The folding transitions of a native protein sequence P and a designed sequence P* can be linked by the following thermodynamic cycle:



In the above cycle, D and N denote, respectively, the denatured and folded states. The folding free energy of protein P (the difference between the free energies of the folded and denaturated state) is:

$$\Delta G_P = G_{P_N} - G_{P_D} \tag{7.3}$$

An analogous relation holds for sequence P^*. The difference in free energy between the native and a designed sequence is

$$\Delta\Delta G \equiv \Delta G_{P^*} - \Delta G_P = (G_{P_N^*} - G_{P_D^*}) - (G_{P_N} - G_{P_D}) \tag{7.4}$$

The free energy of the folded states (P_N, P_N^*) is computed by Proteus, assuming that the protein backbone is fixed into the specified fold (e.g. the crystallographic conformation of the native protein). To compute the free energy of the unfolded states (P_D, P_D^*), we need to employ a model for the conformation of the unfolded protein. A simple model assumes that the unfolded state is a sum of the individual monomers, which do not interact with each other (see Section 7.2.3).

Design calculations which optimize the folding free energy seek to minimize the quantity $\Delta\Delta G$ of the above equation. Since the second parenthesis in the right-hand side (rhs) of the above equation is constant (it refers to quantities of the native protein, which do not change during the design), it is sufficient to minimize the difference in the first parenthesis.

The method employed by Proteus in maximum-stability calculations utilizes the Wernisch algorithm, that was explained in Section 7.1.3. Each cycle begins with a

random assignment of chemical states in "active" and "inactive" positions. Each position is examined sequentially, by scanning over all compatible chemical states; the state with the maximum stability is selected, while the rest of the molecule remains unchanged. Then, the next position is examined. After all positions have been searched, the procedure is repeated several times until the sequence no longer changes or a predefined large number of passes is reached. The final sequence, rotamer set and energy are stored into a file, and a new cycle begins, starting from a new random assignment of chemical states. In this way, the algorithm explores several local regions of the sequence/conformation space from different initial positions, and avoid getting trapped into local minima. In a typical application, 100,000 cycles are conducted, each with 500 maximum number of passes.

7.2.2 Modifications Introduced into the Program Proteus

CPD calculations can also be employed to identify sequences and conformations that optimize the absolute affinity of the protein for a specific ligand (the association free-energy of the complex), or the affinity *relative* to a second ligand. In the work presented in this thesis, we implemented these two optimization criteria into Proteus. The optimization criteria and our implementation are described below.

Criterion of Absolute Binding Affinity

With this criterion, we design sequences that minimize the association free energy of a specific complex; e.g. a ligand (L) and a protein (P). The association of the ligand with two sequences P and P* is described by the following thermodynamic cycle

$$\begin{array}{ccc} P + L & \xrightarrow{1} & PL \\ \downarrow 3 & & \downarrow 4 \\ P^* + L & \xrightarrow{2} & P^* : L \end{array}$$

where P can be the native and P* the designed sequence. The formation of the native complex PL (step 1) is associated with the free energy difference $\Delta G_1 = G_{PL} - (G_P + G_L)$; for the mutant complex P*L (step 2), the corresponding free-energy difference is $\Delta G_2 \equiv G_{P^*L} - (G_{P^*} + G_L)$. The relative association free-energy of the mutant complex with respect to the native complex is

$$\begin{aligned} \Delta\Delta G \equiv \Delta G_2 - \Delta G_1 &= (G_{P^*L} - G_{P^*} - G_L) - (G_{PL} - G_P - G_L) \\ &= (G_{P^*L} - G_{P^*}) - (G_{PL} - G_P) \end{aligned} \quad (7.5)$$

Objective of the affinity optimization calculations is to design protein sequences that minimize the free energy difference $\Delta\Delta G$ in the above equation. Note that the free-ligand contribution cancels out and does not contribute to the above difference. In a design problem that optimizes a series of *ligands* bound to a given protein, the corresponding target free energy difference would be:

$$\Delta\Delta G \equiv \Delta G_2 - \Delta G_1 = (G_{PL^*} - G_{L^*}) - (G_{PL} - G_L). \tag{7.6}$$

The second parenthesis of the rightmost equality in Eq. (7.5) is constant. Thus, it is sufficient to identify mutations/conformations which minimize the term in the first parenthesis. This term depends on the free energies of the protein-ligand complex and the free protein. From a statistical-mechanical point of view, to determine the corresponding free energy difference $G_{P^*L} - G_{P^*}$, it would be necessary to insert the same mutation in both systems, and evaluate the corresponding partition functions $\mathcal{Z}_{P^*L}$ and $\mathcal{Z}_{P^*}$. Despite the fact that the protein has the same chemical structure in P*L and P*, the dominant contribution to each partition function could result from different protein/ligand conformations. In practice, this means that during the optimization of an active position i, the same mutation j should be inserted in the complex and the free protein; however, the algorithm should take into account that the inserted chemical type j can adopt different *optimized* rotamer orientations $j_{P^*L}(i,k)$ and $j_{P^*}(i,l)$, respectively, in the complex and the isolated protein. This is incorporated in the heuristic protocol of affinity optimization as follows:

(i) Initially, the program places all "active" and "inactive" residues to a random chemical state (chemical type/rotamer for active positions, rotamer for inactive positions), as in the maximum-stability protocol. Then a "pass" calculation starts, in which the program loops over active and inactive positions.

At each "active" position i, the program inserts the same chemical type j (among the 18 possible types of Table 7.1 in the complex and in the free protein and determines the rotamers $j_{P^*L}(i,k)$ and $j_{P^*}(i,l)$ that minimize, respectively, the free energy of the complex (G_{P^*L}) and the isolated protein (G_{P^*}). The program also stores the corresponding free energy difference $G_{P^*L} - G_{P^*}$ and proceeds to the next chemical type (sequence/conformation). After going through all possible chemical types, the program selects for position i the type j yielding the minimum value of $G_{P^*L} - G_{P^*}$. The side chain of type j is placed at the optimum rotamers $j_{P^*L}(i,k)$ (in the complex P^*L) and $j_{P^*L}(i,l)$ (in the free protein P^*); this choice reflects the fact that the complex and free protein are distinct states, that minimize separately their free-energy functions. After these steps, the complex and free protein have at position i a side chain with a common chemical type (j), in (possibly different) rotamers that minimize the binding free energy $G_{P^*L} - G_{P^*}$ and the folding free energies of the complex and free protein.

Similarly, at each "inactive" position the program chooses (possibly different) ro-

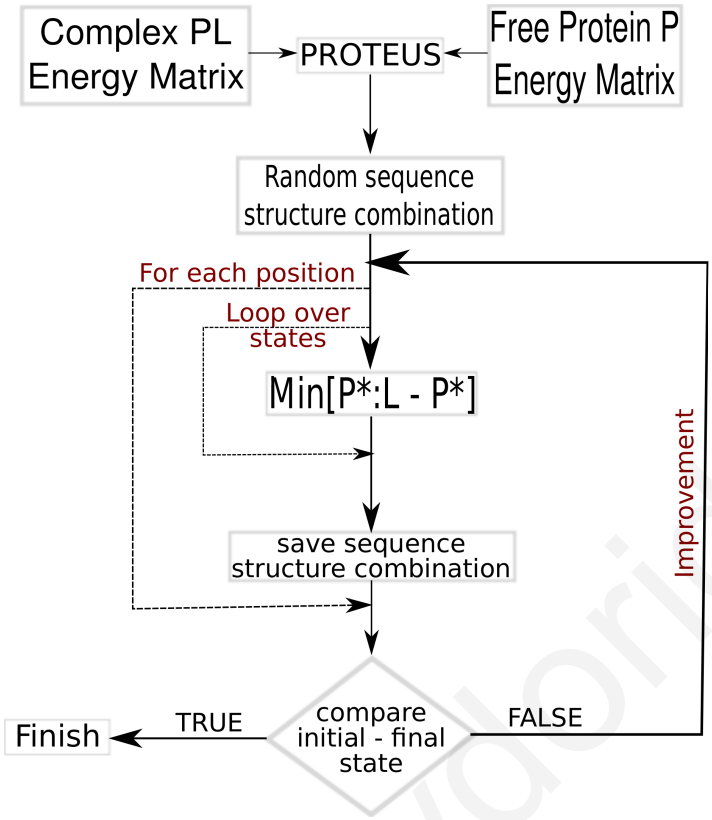


Figure 7.5: Flowchart describing the absolute-affinity algorithm employed in Proteus.

tamers, that minimize the free energies of the complex and free protein. A pass is completed after going through all "active" and "inactive" positions.

(ii) Several passes are executed, until the sequence/conformation does not change, or a maximum predefined number of passes is reached. Then, a new cycle is initiated, with all "active" and "inactive" residues placed in a new randomly chosen chemical state. In a typical design problem, we conduct 100,000 cycles with a maximum number 500 of passes per cycle.

The flowchart shown in Fig. 7.5 illustrates the absolute-affinity protocol implemented in Proteus. The modified code reads two interaction energy matrices, containing the residue pair interaction energies of the complex and the free protein. The new version can also store separate structures (rotamer states) for the complex and the free protein, which permit the independent structural optimization of the two systems.

The use of the absolute-affinity criterion requires some caution; indeed, Eq. (7.5) reveals a potential problem: very negative values of the free-energy difference $G_{P^*L} - G_{P^*}$ can be obtained not only by lowering the term G_{P^*L}, but also by raising the term G_{P^*}. This implies that a straightforward application of the absolute-affinity criterion may promote sequences which maximize affinity by destabilizing the isolated protein. This is undesirable, because such sequences might be unable to fold in the absence of the ligand. One way to obtain sequences with high affinities and high stabilities is to modify the criterion, so that it combines stability and affinity: For example, instead

of the binding free-energy $\Delta G \equiv G_{P^*L} - G_{P^*}$, one could minimize the weighted sum $w_a\,\Delta G + w_s\,G^*$, where $G^* \equiv (G_P + G_{PL})/2$ is the arithmetic mean of the free-protein and complex *folding* free energies. For the special cases ($w_a = 0$, $w_s = 1$) and ($w_a = 0.5$, $w_s = 0.5$), this criterion becomes identical to a stability optimization, respectively of the free protein and the complex. The combination ($w_a = 1$, $w_s = 0$) corresponds to an optimization of the complex with a pure affinity criterion. For other values of the stability/affinity weights, the above criterion takes into account both factors. In practice, to obtain sequences of near-native stability it is sufficient to use a small stability weight (0.1–0.3), as we show below.

Criterion of Relative Affinity (Specificity)

With this criterion, we design sequences that minimize the association free energy of a specific complex (PL_2), with respect to a second complex (PL_1). The relevant thermodynamic cycle is shown below:

$$\begin{array}{ccc} P + L_1 & \xrightarrow{\;1\;} & PL_1 \\ \Big\downarrow{\scriptstyle 3} & & \Big\downarrow{\scriptstyle 4} \\ P + L_2 & \xrightarrow{\;2\;} & PL_2 \end{array}$$

Steps 1 and 2 correspond to the formation of the two complexes PL_1 and PL_2. The relative affinity $\Delta\Delta G$ is

$$\Delta\Delta G = \Delta G_{PL_2} - \Delta G_{PL_1} = (G_{PL_2} - G_{PL_1}) - (G_{L_2} - G_{L_1}) \tag{7.7}$$

where $G_{PL_{1/2}}$ and $G_{L_{1/2}}$ are, respectively, the free energies of the complexes and isolated ligands. The free energy of the isolated protein cancels out and does not contribute to the difference.

We can write the analogous thermodynamic cycle for the complexes of a designed sequence P* and the same ligands L_1 or L_2. The corresponding relative affinity is

$$\Delta\Delta G^* = (G_{P^*L_2} - G_{P^*L_1}) - (G_{L_2} - G_{L_1}) \tag{7.8}$$

A design with the relative-affinity criterion seeks to identify protein sequences {P*}, whose relative affinity is much smaller than the corresponding affinity of the native protein P. That is, the desirable sequences should minimize the *change* in the protein relative affinity for ligands L_1 and L_2 due to the inserted protein mutation:

$$\Delta\Delta\Delta G = \Delta\Delta G^* - \Delta\Delta G = (G_{P^*L_2} - G_{P^*L_1}) - (G_{PL_2} - G_{PL_1}). \tag{7.9}$$

The second parenthesis on the rhs does not change during the design. Thus, to identify

mutations which optimize the relative affinity, it is sufficient to minimize the term in the first parenthesis.

As for the absolute-affinity protocol of the previous section, the relative-affinity calculation needs to compare complexes bearing the same protein mutation. The dominant contributions to the partition functions of these complexes may come from different conformations. This is accounted in our implemented relative-affinity cirterion as follows:

(i) In the beginning, the program reads separate interaction-energy matrixes for the two complexes PL_1 and PL_2.

(ii) When a design cycle is initiated, the algorithm assigns to all "active" and "inactive" residues a random chemical state (chemical type/rotamer for "active", rotamer for "inactive" residues), as for the stability and simple-affinity criteria.

(iii) During a pass, the algorithm examines all "active" and "inactive" positions. At each "active" position i, the algorithm introduces the same chemical type j in both complexes. It loops over all compatible rotamers of chemical type j and identifies the rotamers $j_{P^*L_2}(i,k)$ and $j_{P^*L_1}(i,l)$ that minimize, respectively, the free energies $G_{P^*L_2}$ and $G_{P^*L_1}$ of the two complexes; these free-energies are computed from the interaction energy matrices of the two complexes. The algorithm also computes and stores the corresponding free-energy difference $(G_{P^*L_2}-G_{P^*L_1})$ and proceeds to the next chemical type. After looping all chemical types, the algorithm selects for position i the chemical type j that minimizes $(G_{P^*L_2}-G_{P^*L_1})$, in rotamer conformations $j_{P^*L_2}(i,k)$ and $j_{P^*L_1}(i,l)$ that minimize, respectively, the free energies $G_{P^*L_2}$ and $G_{P^*L_1}$.

At each "inactive" position, the algorithm chooses rotamers that minimize the free energies of the two complexes, $G_{P^*L_2}$ and $G_{P^*L_1}$. These rotamers are possibly different.

(iv) A pass is repeated until the sequence no longer changes, or a predefined, maximum number of passes is reached. Then, a new cycle is initiated, by randomizing the chemical state of "active"/"inactive" residues. In a typical design, 100,000 cycles are executed with 500 maximum passes per cycle.

As for the absolute-affinity criterion, the use of the relative-affinity criterion should be done with caution. Eq. (7.9) shows that a low relative affinity can be obtained by destabilization of the first complex PL_1, possibly due to the selection of sequences with high folding free energy (low stability) of the protein P. In order to identify protein sequences with near-native stability and low relative affinity, a combined criterion can be used. In our calculations, we introduce this combined criterion as follows:

(a) In stage (i) (above), the modified program reads three interaction energy matrices, containing the interaction energies of the two complexes and the free protein. The modified program can also store separate structures (rotamer states) for the complexes and the free protein, which permit the independent structural optimization of the three systems.

(b) In stage (iii), the program also determines the rotamer $j_{P^*}(m)$ that minimizes

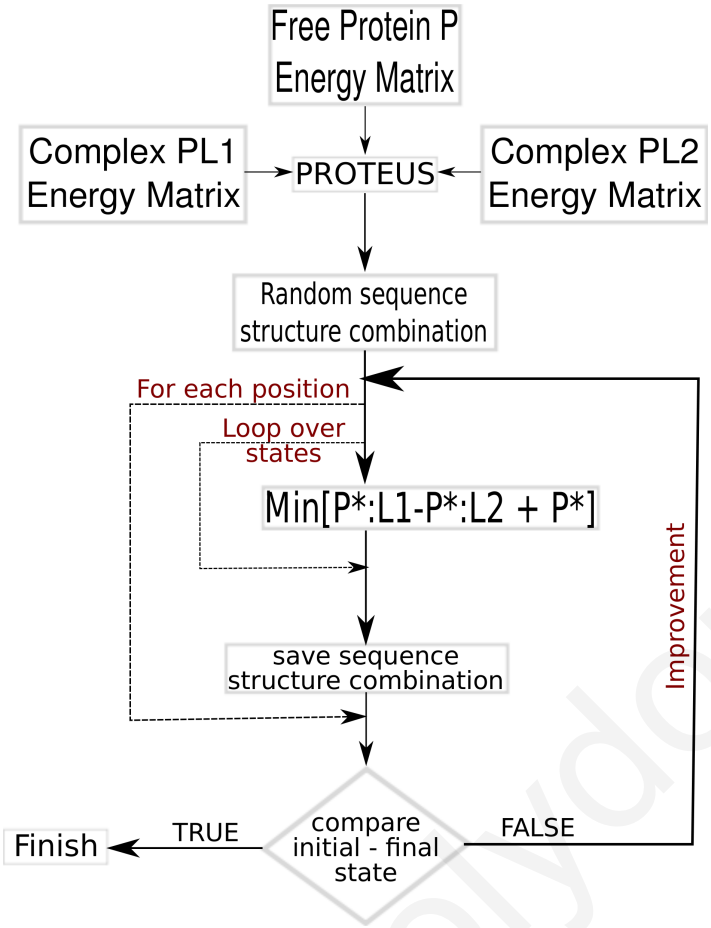


Figure 7.6: Diagram representation of the Proteus algorithm employed in design calculations with the criterion of relative affinity.

the folding free-energy of the free protein P*, and computes the free-energy combination $[G_{P^*L_2} - G_{P^*L_1}] + G_{P^*}$. The term in brackets $[\cdots]$ corresponds to the relative affinity of the protein for the two ligands L_1 and L_2; the last term is the *folding* free energy of the free protein. The minimization of this combination ensures that the algorithm searches for sequences, which differentiates between the two ligands without destabilizing the free protein. After looping over all chemical types, the program places into position i of the complexes P^*L_1, P^*L_2 and P^* the chemical type j that minimizes the above free-energy combination, respectively in rotamers $j_{P^*L_2}(i,k)$, $j_{P^*L_1}(i,l)$, and $j_{P^*}(i,m)$. At each "inactive" position, the program selects (possibly different) rotamers, that minimize the free energies of the two complexes and the free protein.

The flowchart shown in Fig. 7.6 illustrates the Proteus relative affinity protocol, employed for designing sequences which optimize the binding free energy for a ligand relative to another.

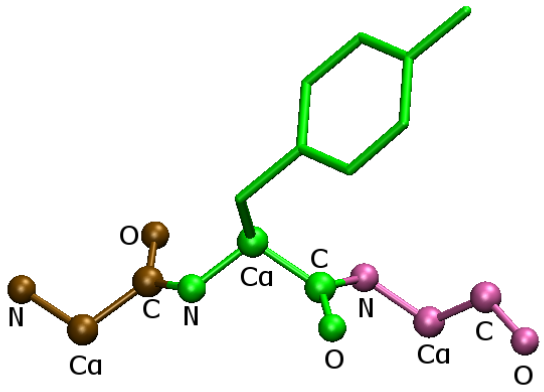


Figure 7.7: The tripeptide Ala-X-Ala model used in the reference-energy calculations of the unfolded state. The tripeptide consists of the backbone parts (atoms N, C_α, C, O) of three residues (brown, green, purple) and the sidechain X at the center (green).In the example shown, the side chain corresponds to a tyrosine.

7.2.3 The Unfolded State

The maximum-stability criterion (Section 7.2.1) or the combined stability/affinity criteria (Section 7.2.2) requires the computation of a folding free-energy, which is the difference between the free energies of the protein in its folded conformation and the unfolded state [Eq. (7.3)]. The free-energy of the folded state employs a well-defined structural model for the "fixed part" (e.g. the crystallographic conformation). For the unfolded state there is no satisfactory (or even unique) structural model. Thus, the corresponding free energy needs to be computed in an approximate manner. A simple, often used model treats the unfolded state as a set of non-interacting residues, with the chemical composition of the folded state [73]. Each of the possible chemical types (e.g. the 18 types in Table 7.1) j is associated with a "reference" free energy $G_j^{\rm ref}$ (see below). Then, for a particular sequence $\{j_1, j_2, \ldots, j_N\}$ the free energy of the unfolded state is:

$$G^D(j_1, j_2, \ldots, j_N) = \sum_{k=1}^{N} G_{j_k}^{\rm ref} = \sum_X N_X G_X^{\rm ref} \qquad (7.10)$$

In the last equality of Eq. (7.10), N_X is the number of of sidechains with chemical type X, and G_X^{ref} is the corresponding "reference" energy. The summation is over all possible chemical types.

To alleviate the independent-aminoacid assumption in the computation of the reference values $\{G_X^{\rm ref}\}$, the considered side chain X is usually placed in the context of a bigger molecule. In our case, it is part ot a tripeptide with sequence Ala-X-Ala (Fig. 7.7) [73].

Chapter 8

Aminoacyl-tRNA Synthetases

8.1 The Biological Role of Synthetases

Protein synthesis is the biological process in which proteins are produced inside the cell, through the successive mechanisms of transcription and translation (Fig. 8.1). In order to produce a protein sequence, a succession of transfer RNA (tRNA) molecules, charged with the correct amino acids (native complex aa:tRNA) are brought together and matched up with a messenger RNA (mRNA) molecule (Fig. 8.2). The "anticodon" part matches the appropriate "codon" part at the mRNA chain. The aminoacids at the 5' end of the tRNA are then linked together through peptide bonds to form the protein sequence, and release the tRNA molecule. This process is carried out by the ribosome, which is attached to the mRNA and moves along "reading" it (matching codons-anticodons), linking successive amino acids to each other, progressively assembling the protein sequence.

Transcription: In the transcription process, an mRNA chain is generated from one strand of the DNA double helix. The process takes place inside the cell nucleus, where DNA is located, with RNA Polymerase copying the template strand to a mRNA. Both DeoxyRibonucleic Acid (DNA) and Ribonucleic Acid (RNA) are nucleotide polymers, with monomers consisting of a base (Adenine-A, Guanine-G, Thymine-T, Cytosine-C, Fig. 8.3) attached to a deoxyribose sugar and phosphate groups joined by ester bonds. Codons correspond to nucleotide triplets that code for specific amino acids. There are 64 possible triplets which are mapped to amino acids or terminating signals, listed in Table 8.4. The genetic information stored in the DNA is transcripted into a sequence of codons, the mRNA molecule (in RNA the Thymine base is replaced by Uracile).

Translation: In the translation process, each successive codon of the mRNA sequence, is complemented by the anticodon of the appropriate charged tRNA transferring the appropriate aminoacids, in order to assemble a protein sequence. The

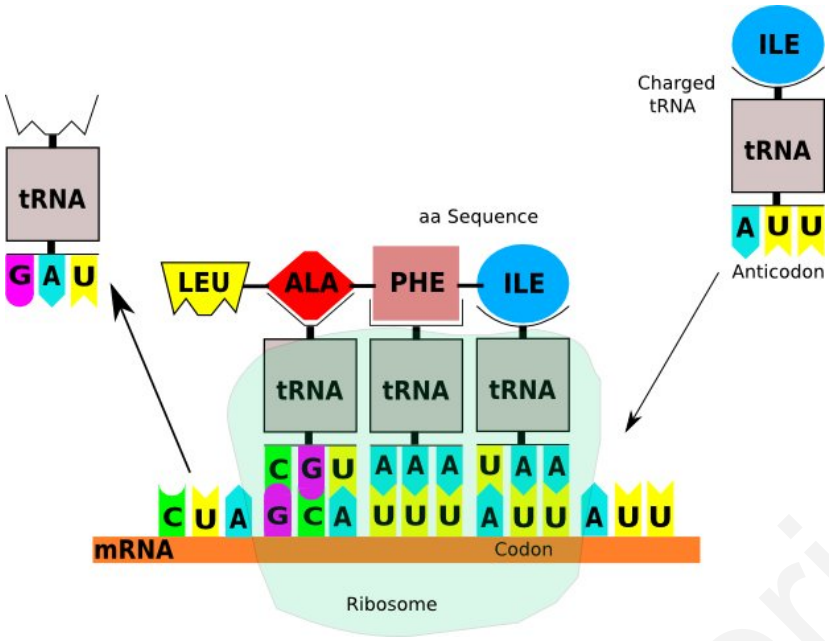


Figure 8.1: Schematic representation of the protein synthesis.

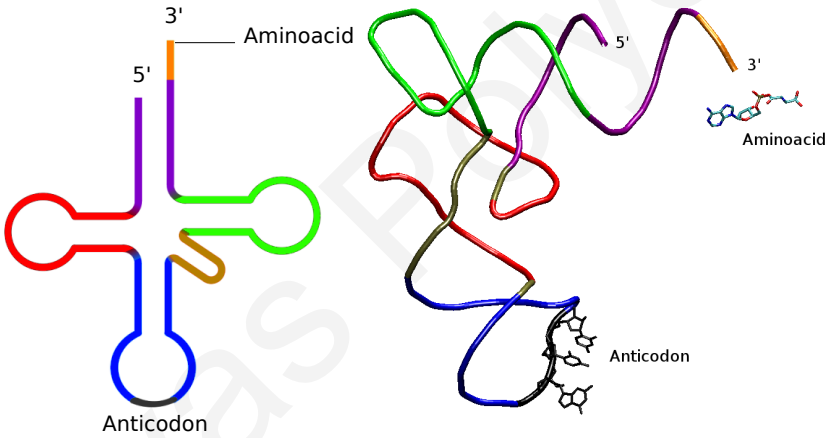
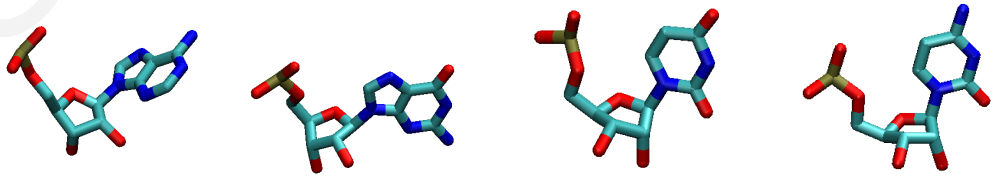


Figure 8.2: 2D (left) and 3D (right) representations of the tRNA molecule. The main secondary structure components of tRNA are the anticodon arm (blue), the variable arm (brown), the D arm (red), the TΨC arm (green), the acceptor stem (purple) and the CCA tail (orange).



(a) Adenine (b) Guanine (c) Uracile (d) Cytosine

Figure 8.3: Molecular structural of the RNA nucleotides. (a)-(b) Adenine and guanine are purines; (c)-(d) cytosine and uracil are pyrimidines.

	C	U	G	A	3rd base
A	ACU Thr	AUU Ile	AGU Ser	AAU Asn	U
	ACC Thr	AUC Ile	AGC Ser	AAC Asn	C
	ACA Thr	AUA Met	AGA Ser/Gly	AAA Lys	A
	ACG Thr	AUG Met	AGG Ser/Gly	AAG Lys	G
G	GCU Ala	GUU Val	GGU Gly	GAU Asp	U
	GCC Ala	GUC Val	GGC Gly	GAC Asp	C
	GCA Ala	GUA Val	GGA Gly	GAA Glu	A
	GCG Ala	GUG Val	GGG Gly	GAG Glu	G
C	CCU Pro	CUU Leu	CGU Arg	CAU His	U
	CCC Pro	CUC Leu	CGC Arg	CAC His	C
	CCA Pro	CUA Leu	CGA Arg	CAA Gln	A
	CCG Pro	CUG Leu	CGG Arg	CAG Gln	G
U	UCU Ser	UUU Phe	UGU Cys	UAU Tyr	U
	UCC Ser	UUC Phe	UGC Cys	UAC Tyr	C
	UCA Ser	UUA Leu	UGA Trp	UAA Ter	A
	UCG Ser	UUG Leu	UGG Trp	UAG Ter	G

Figure 8.4: The genetic code. 20 natural amino acid types are represented by 64 nucleotide triplets (including termination triplets (white)). Codons colored orange are deduced from aminoacylation of a tRNA activated by a class II synthetase; those colored light blue by class I. PheRS belongs to class II synthetases but it aminoacylates at 2'OH (characteristic of class I); while TyrRS belongs to class I and it aminoacylates at 3'OH (characteristic of class II). The figure is reproduced from Ref. [186].

translation mechanism involves four steps, as shown in Fig. 8.5

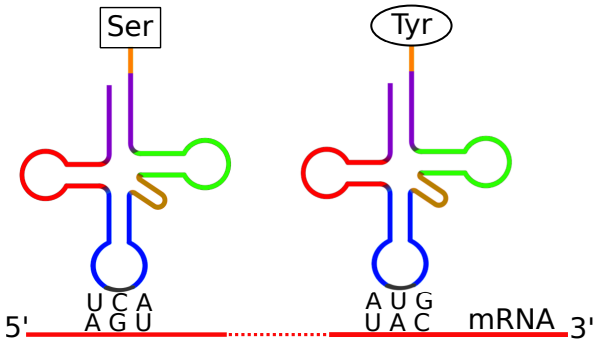


Figure 8.5: The translation mechanism maps the codon of the mRNA with the anticodon of the charged tRNA. The aminoacids transfered by tRNAs join the protein sequence.

I. Activation: During the activation step, specific enzymes known as aminoacyl-tRNA synthetases charge tRNA molecules with the correct amino acids (i.e. catalyze the formation of a covalent bond between a specific amino acid, and the tRNA with the appropriate anticodon). Aminoacyl-tRNA synthetases (aaRS) are proteins, whose biological role is to recognize the correct amino acid and tRNA, and form the covalent bond between them. This is achieved by a two-step reaction. First, the synthetase binds an adenosine 5' triphosphate (ATP) molecule and the cognate amino acid (aa), forms the complex aminoacyl-tRNA:aa-adenylate (aaAMP) and releases inorganic pyrophosphate [Eq. (8.1)]. The nucleotide ATP is responsible for the transport of energy inside the cell; it is transformed to an adenine monophosphate (AMP) by releasing a pyrophosphate molecule and energy ($ATP + H_2O \rightarrow AMP + PP_i$). The aspartyl-adenylate (AspAMP) shown in Fig. 8.6 consists of the AMP part (an adenine attached to a pentose sugar and a phosphate group), and the aminoacid moiety. The adenylate aaRS complex then binds the correct tRNA molecule and the aminoacid is transfered from the aa:AMP to the tRNA, which is now set to be charged [Eq. (8.2)].

$$aaRS + aa + ATP \longrightarrow aaRS : aaAMP + PPi \tag{8.1}$$

$$aaRS : aaAMP + tRNA \longrightarrow aa : tRNA + aaRS + AMP \tag{8.2}$$

The sum of these two steps is known as the aminoacylation process.

Transfer ribonucleic acids (tRNAs) are amino-acid carriers during the translation mechanism. Small in size, they adobt an L-shape 3D structure that helps them settling down at the protein synthesis site, inside the ribosome (Fig. 8.2). The nucleotide chain of tRNA consists of the 5'-terminal phosphate group initiating the acceptor stem, the D arm followed by the anticodon arm which "holds" the specific anticodon, the anticodon itself, the T arm and the CCA tail at the 3'-terminal, which is specific for recognition

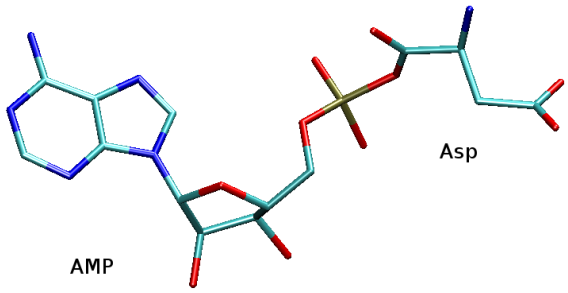


Figure 8.6: The aspartyl - adenylate (AspAMP) consists of the AMP and the aspartic acid.

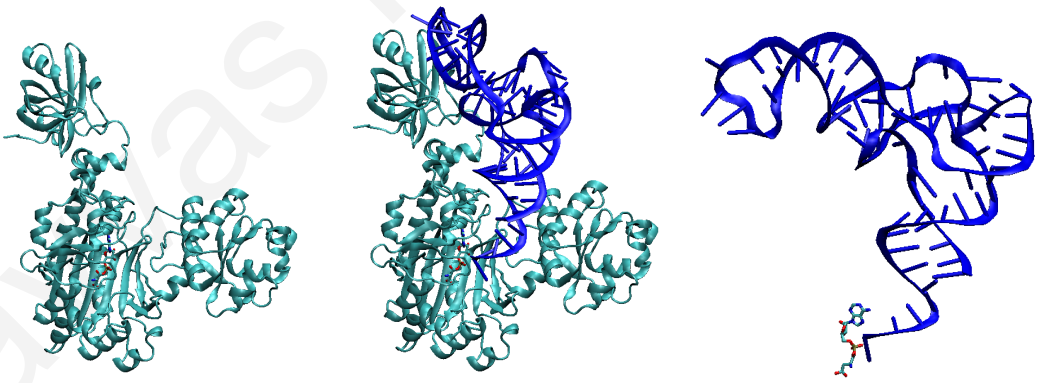


Figure 8.7: During aminoacylation, the AspRS synthetase binds an ATP molecule and an asprartic acid (Asp) to form the complex AspRS:AspAMP (step1). The complex then binds the tRNA molecule which captures the aminoacid, releasing AMP (step2). The synthetase is shown in a ribbon representation (cyan), the tRNA in cartoon (blue) and the AspAMP in licorice.

of the tRNA, by the appropriate aminoacyl-tRNA synthetase. Each tRNA is specific for a single aminoacid, but can bear different anticodons (multible codons describe the same amino acid: degeneracy).

II. Initiation: The ribosome provides the environment for the protein synthesis. At the initiation of the procedure, the ribosome is attached to the 5' end of the mRNA, moving along successive codons as the sequence assemples.

III. Elongation: During the elongation stage, the charged tRNA in turn binds its anticodon to the corresponding codon of the mRNA, and becomes a member of the protein sequence by forming a peptide bond to the preceding amino acid.

IV. Termination: The termination process is activated when the A site of the ribosome faces a stop codon and terminates the translation mechanism.

From the above, it follows that the high specificity of the aaRS proteins for both the amino acid and the tRNA molecule are absolutely necessary, to maintain the integrity of the translation mechanism. Manipulation of the synthetase amino acid (or tRNA) specificity may lead to a reduced genetic code (the incorporation of all amino acids by a smaller number of synthetases), or assist the addition of artificial aminoacids into a protein. The modification of synthetase specificity and binding affinity by high-throughput CPD methods are the main aspect of this work and will be discussed in detail later on.

8.2 The Structure of Aminoacyl-tRNA Synthetases

Aminoacyl-tRNA synthetases constitute a family of twenty proteins, responsible for the correct charging of tRNAs with their cognate aminoacids. This is achieved by the aminoacylation reaction described above. Although all proteins bind the same ATP molecule and attach the amino acid to the 3' terminal of the tRNA, their structures present a significant diversity, both in size and secondary structure. The shortest sequence member (TrpRS, responsible for the recognition of tryptophane) consists of 334 residues, while the longest (PheRS, recognizing phenylalanine) has 1112 residues. The common protein structure motif called "Rossmann fold" ($\beta-\alpha-\beta-\alpha-\beta$), that was observed at the ATP binding pockets of the first aaRS x-ray crystal structures, led to the assumption that all synthetases proteins had similar structural and chemical properties. The crystallographic analysis of new aaRS members rejected the previous hypothesis, and suggested the classification of the synthetase family into two classes, distinguished by their active site structural differences upon ATP-binding.

The synthetases have two active sites, one for the amino acid and one for the tRNA. During aminoacylation the synthetase binds the amino acid, the ATP molecule and the

Table 8.1: The two classes of aminoacyl-tRNA synthetases

Class I	Class II
2' OH	3' OH
Leu	His
Ile	Pro
Val	Ser
Cys	Thr
Met	
Glu	Asp
Gln	Asn
Arg	Lys
Tyr	Gly
Trp	Ala
	Phe

tRNA, which binds the aminoacyl adenylate at the adenine ribose of the 3' terminal (Fig. 8.2). Class I synthetases aminoacylate on the 2' OH of the ribose and class II on the 3' OH.

8.2.1 Class I Synthetases

Class I synthetases (left column of Table 8.1) are usually monomeric or dimeric (TyrRS, TrpRS) multidomain proteins. All members of the class share a similar active site domain topology, known as "Rossmann fold", which consists of a parallel five $\beta-$ strand sheet, flanked by $\alpha-$ helices on the sides, following a $\beta-\alpha-\beta-\alpha-\beta$ pattern. The motif is shown in Fig. 8.8. Multible sequence alignment of class I aaRS (Fig. 8.10), revealed two highly conserved oligopeptide motifs, His-Ile-Gly-High (HIGH) and Lys-Met-Ser-Lys-Ser (KMSKS), which constitute parts of the ATP-binding pocket [187]. During the aminoacylation reaction [Eq. (8.1)], (i) class I aaRSs bind an extended conformation of the ATP to activate the amino acid, (ii) dock the tRNA molecule in an orientation where its variable loop (Fig. 8.9) faces the solvent, and (iii) bring together the 2' OH of the tRNA's 3' terminal adenine ribose and the carbonyl of the aminoacyl-adenylate. Class I can be further grouped into aliphatic (Ia), charged (Ib), and aromatic (Ic) subclasses.

8.2.2 Class II Synthetases

Class II synthetases (right column of Table 8.1) are usually dimeric or tetrameric. The active site domain of these proteins is formed by an antiparallel $\beta-$ sheet, surrounded by $\alpha-$ helices (Fig. 8.11). They contain three motifs (different from class I), which participate either in the dimer interface (motif 1), or in the active site (motifs 2 and

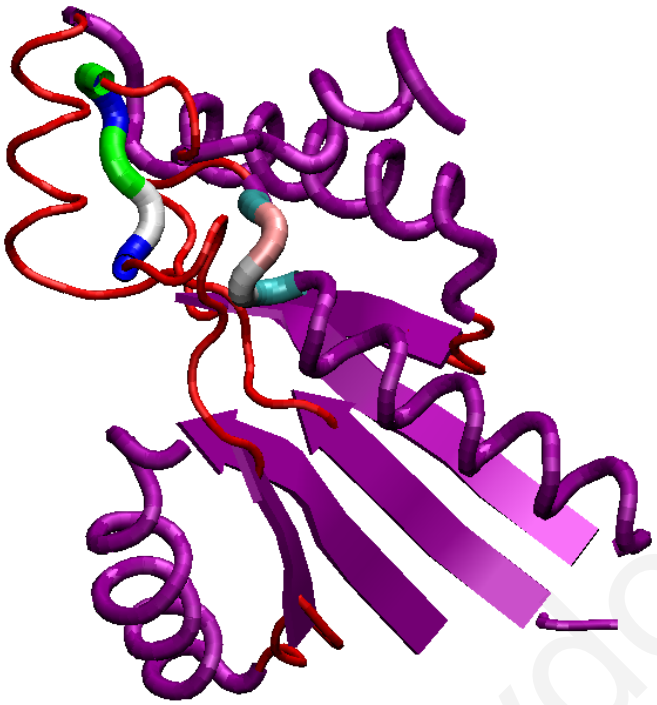


Figure 8.8: A 3D representation of the active site Rossmann fold, characteristic structural motif of the class I synthetases. The conserved peptides HIGH and KMSKS, associated with ATP binding, are shown in tubes colored by the residue names.

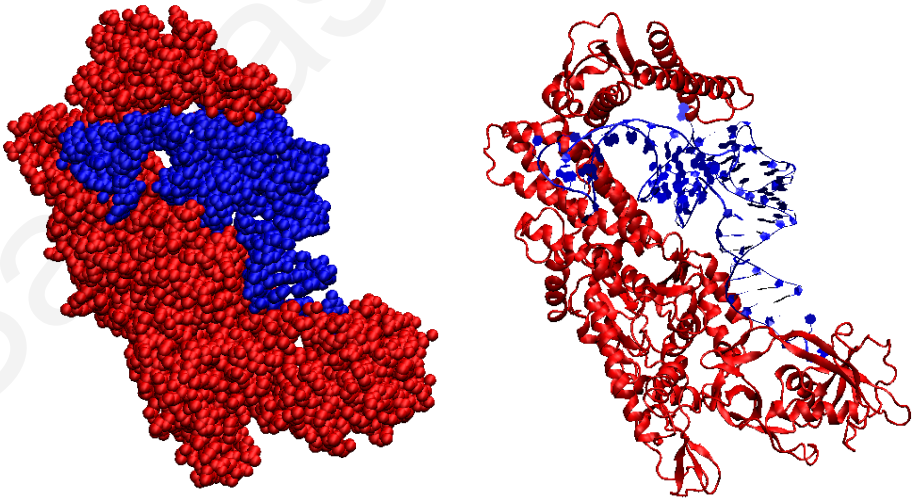


Figure 8.9: Class I synthetases bind an extended conformation of the ATP, while the tRNA binds its acceptor stem to the minor groove side and the variable loop faces the solvent.

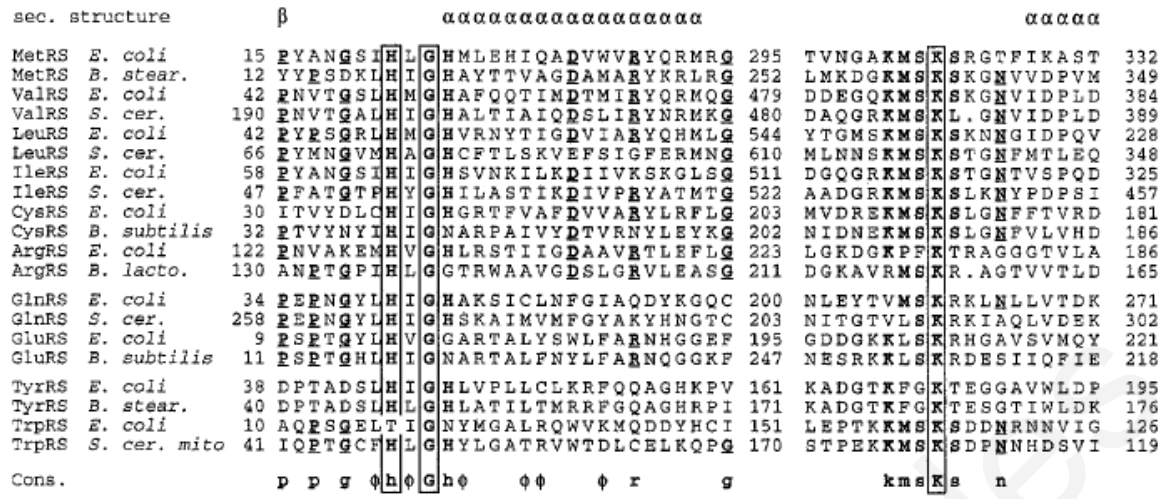


Figure 8.10: Sequence alignment of class I synthetases. Invariant residues are boxed and highly conserved residues are shown in bold face and underlined. The secondary structure of the sequence portion is shown at the top lines ($\alpha-$helix and $\beta-$sheet). The bottom line shows consensus residues using small letters for residues A, G, P, S, and T; ϕ for hydrophobic residues like C, V, L, I, M, P, T and W); + for positively charged and polar residues (H, K, R); and - for negatively charged and polar residues (D, N, E, Q). The figure is taken from Ref. [187].

3) [188]. The ATP adopts a more compact conformation upon binding compared to class I, while tRNA is positioned so that its variable loop faces the protein instead of the solvent. Class II synthetases choose the 3' OH of the tRNA's 3' terminal ribose to approach the carbonyl of the aminoacyl-adenylate, with the exception of PheRS, which aminoacylates at 2' OH.

Additional common features divide class II into subclasses with increased homology. HisRS, ProRS, ThrRS and SerRS exhibit a significant sequence homology at their carboxy terminal and constitute subclass IIa. Similarly AspRS, AsnRS and LysRS exhibit significant sequence homology at their amino terminal and form subclass IIb. The remaining AlaRS, GlyRS and the exception PheRS can be grouped into subclass IIc.

Class II synthetases AspRS, AsnRS bind the corresponding amino acids by an almost identical active site (except three positions), composed of highly conserved residues. The amino acid similarity (Asp / Asn) and the two alike active sites of these synthetases, was the reason of this work. In the next chapter a thorough analysis is performed on AsnRS to identify mutants that increase their stability or alter its sprecificity.

8.3 The AsnRS and AspRS Synthetases

AsnRS and AspRS are two closely homologous synthetases, with significant similarities in their amino acid sequences, and their tertiary structures. Their cognate ligands have

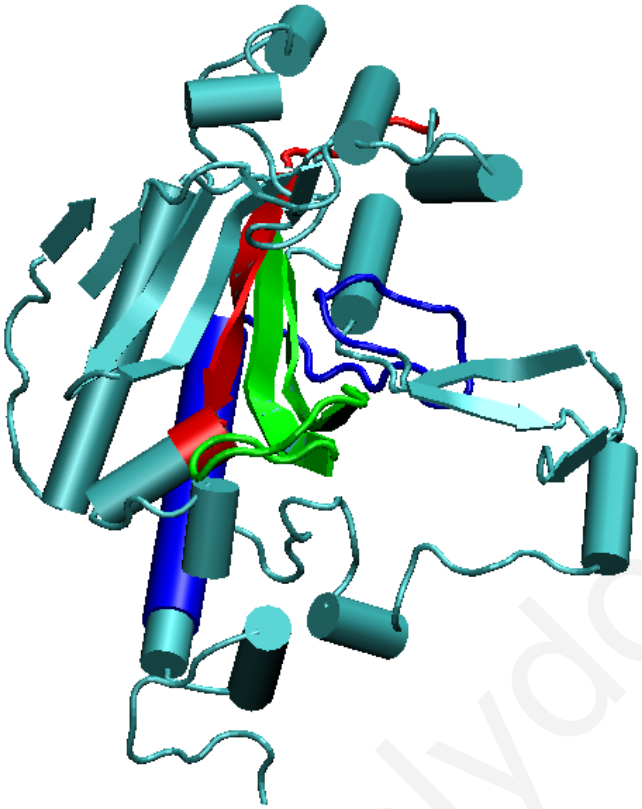


Figure 8.11: A 3D representation of the class II AspRS synthetase active site. The consensus sequence motifs **1** (residues 135-174) in *E. coli* numbering, **2** (208-236) and **3** (523-537) are colored blue, green and red, respectively.

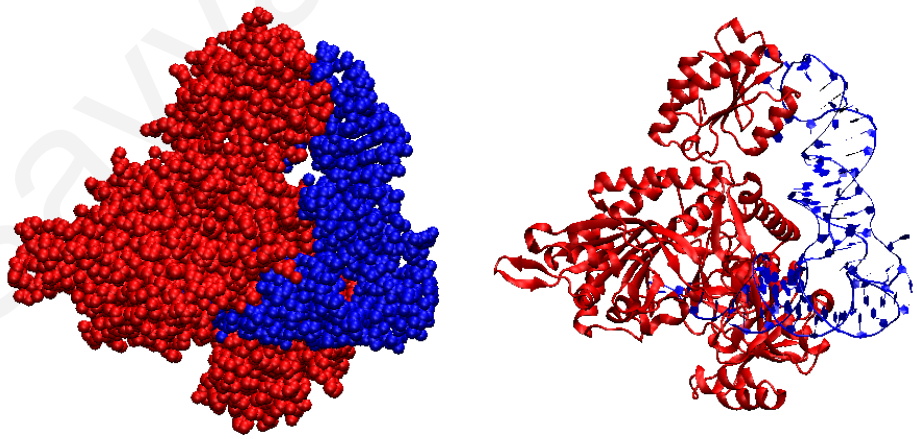


Figure 8.12: Class II synthetases bind a bent conformation of the ATP, while the tRNA binds with its acceptor stem to the side of the major groove and the variable loop facing the protein.

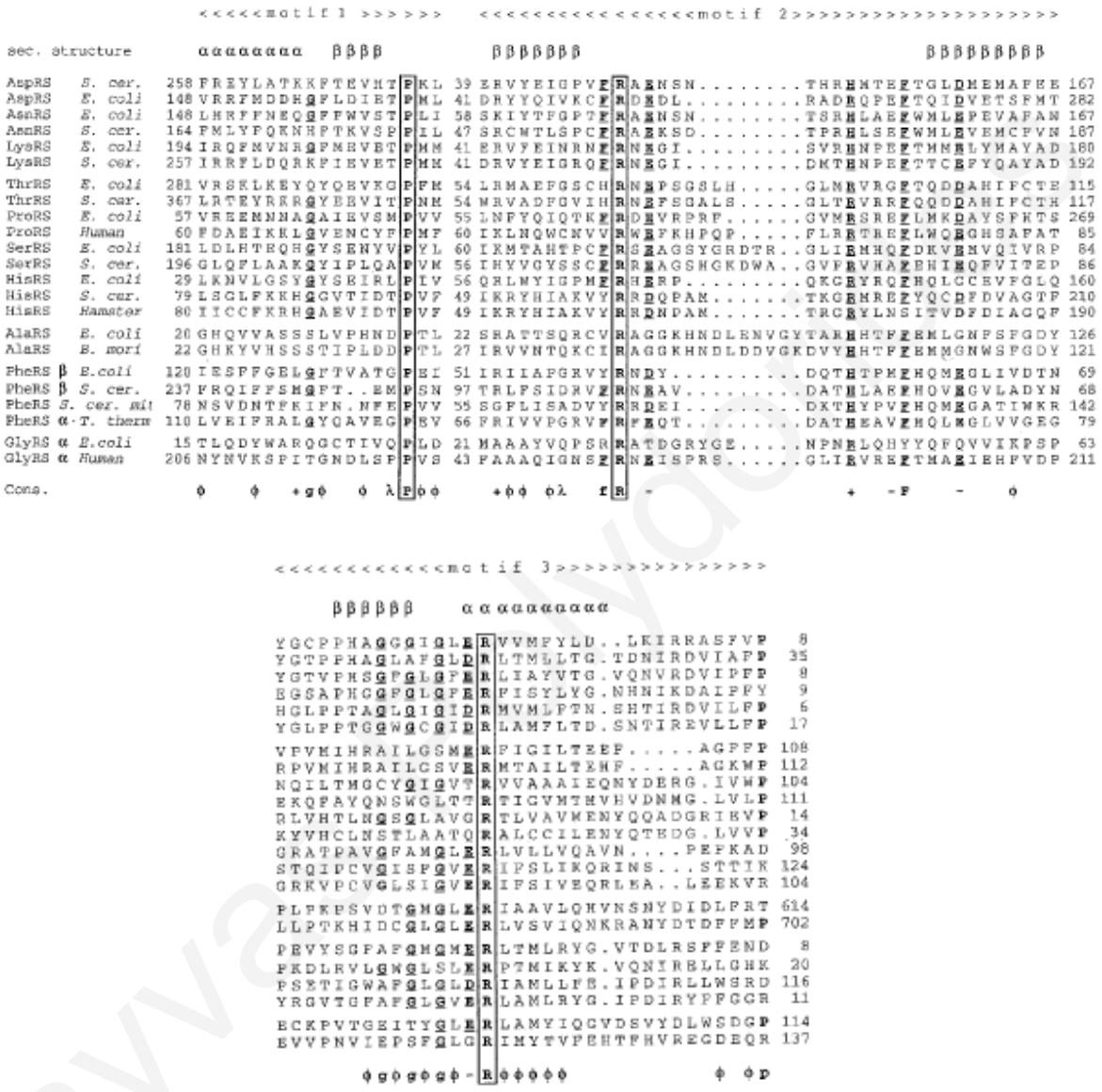


Figure 8.13: Sequence alignment of class II synthetases. Invariant residues are boxed; highly conserved residues are in bold face and underlined. The secondary structure of the sequence portion is shown at the top lines ($\alpha-$helix and $\beta-$sheet). The bottom line shows consensus residues using small letters for residues A, G, P, S, and T; ϕ for hydrophobic residues like C, V, L, I, M, P, T and W); + for positively charged and polar residues (H, K, R); and - for negatively charged and polar residues (D, N, E, Q). The figure is taken from Ref. [187].

Table 8.2: Active site residues in AsnRS and AspRS

AsnRS (Tt)	AspRS (Ec)
R208	R217
E210	E219
L218	Q226
F221	F229
D352	D475
E361	E482
M223	Q231
E164	E171
S185	S193
Q187	Q195
A190	K198
E191	Q199
E225	D233
E227	E235
S366	S487
R368	R489

also similar in size side chains, since asparagine corresponds to the neutral analogue of the aspartic acid. This fact supports the impressive similarity of the two active sites, especially in the vicinity of the common part of the ligand, the adenylate. Comparing the active sites of AsnRS from *Thermus themophilus* (Tt) and AspRS from *Escherichia coli* (Ec), AsnRS residues Arg208, Glu210, Phe221 stabilizing the adenylate are maintained as Arg217, Glu219, Phe229 in AspRS, while Leu218 is replaced by Gln226. The magnesium ion interacting with residues Asp352 and Glu361 in the binding pocket of AsnRS, remains in place by Asp475 and Glu482 in AspRS. The residue pairs Glu361 (Glu482) and Asp352 (Asp475), which held intact the essential Mg^{2+} ion, are invariant in all class II synthetases [140].

8.3.1 The Asparaginyl-tRNA Synthetase (AsnRS)

Asparaginyl-tRNA synthetase of *Thermus thermophilus* is a homodimeric protein with 438 residues for each monomer. The N-terminal domain which recognizes the tRNA anticodon, consists of 5 $\beta-$ strands in a barrel shape with an $\alpha-$ helix in between, a common topology of all class IIb members. The C-terminal domain composed of 6 antiparallel $\beta-$ strands, includes the protein's active site and the interface between the two monomers. The 3D structure of the AsnRS:AsnAMP complex is shown in Fig. 8.15.

Figure 8.14: Sequence alignment of the asparaginyl-tRNA from *Thermus thermophilus* and the aspartyl-tRNA from *Escherichia coli*. Small, hydrophobic and aromatic (except Tyr) residues are colored red, negative blue, positive magenta, polar (including hydroxyl) green and others gray. $*$ notes identical sidechains, : highly conserved and . semi conserved. The five active positions of AsnRS employed in the design calculation of Chapter 9 are indicated with inclined numbers.

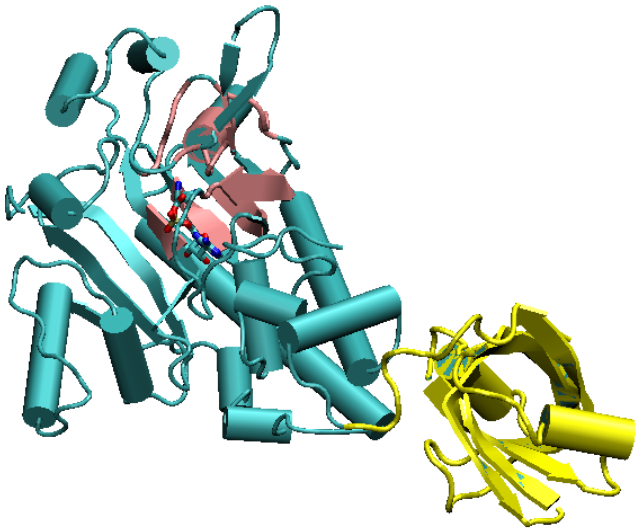


Figure 8.15: Cartoon representation of the (Tt) asparaginyl-tRNA synthetase, in complex with its cognate ligand AsnAMP. The N-terminal domain is colored yellow and the part of the active site in the vicinity of the ligand is shown in pink.

The AsnRS Active Site

Figure 8.16 presents the active site of AsnRS:AsnAMP complex, including the magnesium Mg^{2+} ion. The class II synthetases bind three Mg^{2+} ions [140] but the two of them capture the released pyrophosphate during the first step of aminoacylation.

The top panel describes the interactions of the common part of the ligand. The base is held fixed by two hydrogen bonds between the two adjacent amide atoms (N1 and N6) of the adenine and the backbone carbonyl- and amino groups of Leu218; amide atom N6 forms another hydrogen bond with Glu210. The base ring forms also $\pi-$ stacking interactions with Phe221. The adjacent ribose sugar interacts with Gly409 and Glu361.

The magnesium ion interacts with Glu361 and Asp352 (conserved in all class II synthetases, as described above), and keeps the phosphate group fixed with the assistance of a hydrogen-bond from the Arg208 side chain in the opposite site.

The bottom panel describes the most important interactions between the amino acid and the active site. The side chain carboxylate of the AsnAMP ligand interacts with Arg368 and Glu225 and its ammonium group interacts with Ser185, Gln187 and Glu164. Among the active-site residues, the side-chain of Gln187 interacts with the ligand ammonium and Glu191; Glu225 makes a hydrogen bond with Ser185 and contacts Met233, in addition to its interaction with the ligand side chain; The Glu227 side-chain forms an electrostatic interaction (salt bridge) with the Arg368 side chain.

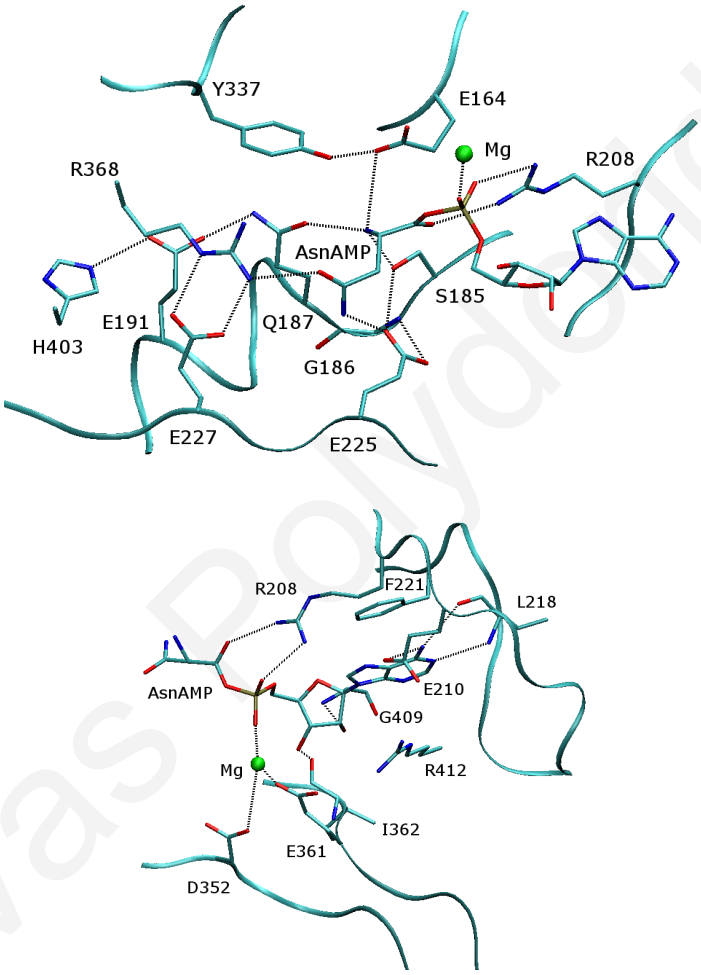


Figure 8.16: The active site of the AsnRS:AsnAMP complex. The top panel indicates the most important interactions of the Asn sidechain with surrounding residues (dashed lines). The bottom panel shows the most important interactions of the AMP moiety.

8.3.2 The Aspartyl-tRNA Synthetase (AspRS)

Aspartyl-tRNA synthetase of *Escherichia coli* is a homodimeric protein [189]. Each monomer contains 590 residues and consists of four domains: (i) The N-terminal domain, that is responsible for the tRNA anticodon recognition. This domain folds into a $\beta-$ barrel / $\alpha-$ helix, characteristic of all class IIb synthetases; (ii) The hinge domain (composed of a few aminoacids), that is involved in tRNA recognition; (iii) the catalytic domain, that consists of the C-terminal, including the protein active site and the three signature motifs of class II; finally, (iv) the insertion domain, that is formed by an antiparallel $\beta-$ sheet with three $\alpha-$ helix on the sides Fig. 8.17.

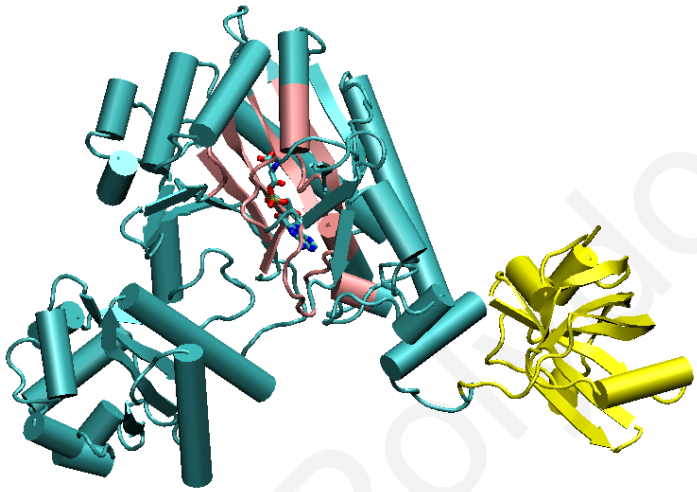


Figure 8.17: Cartoon representation of the aspartyl-tRNA synthetase from Ec in complex with its cognate ligand AspAMP. The N-terminal domain is colored yellow and the part of the active site in the vicinity of the ligand is shown in pink.

The AspRS Active Site

The active site structure of the complex formed by the AspRS synthetase and the aspartyl-adenylate (AspAMP) is shown in Fig. 8.18. The cognate AspAMP ligand forms a group of hydrogen bonds with residues of the binding pocket. The aspartic acid moiety is stabilized by several important residues conserved in several AspRSs. Arg489 (class IIb invariant) forms two hydrogen bonds with the Asp side chain; Lys198 and His448 contribute two more hydrogen bonds. Lys198 interacts also strongly with Asp233, and His448 is part of the so-called "histidine loop", formed by residues 436 - 449. The aspartyl backbone ammonium group makes three hydrogen bonds with Glu171, Ser193 and Gln195. Glu171 is part of the "flipping loop" (residues 167 - 173), which adopts a close conformation in the bound state and an open in the unbound. The above residues 190 to 199, compose an invariant oligopeptide chain LXQSPQXXKQ in all AspRS synthetases from different organisms.

The AMP moiety is stabilized by interactions with residues Arg217, Phe229 and Arg537. Arg217 makes two hydrogen bonds with the carbonyl of the Asp main chain and the $\alpha-$phosphate oxygen (O1P), while the adenine is placed in between Phe229 and Arg537. Gln231 is an important residue, responsible for recognizing the ligand AspAMP [190]. It forms two hydrogen bonds with the carbonyl group of the Asp backbone and the $\alpha-$phosphate oxygen (O5'). Residues Arg217, Phe229, Arg537, Gln192, Asp475 and Glu482 are conserved in all class II synthetases [140].

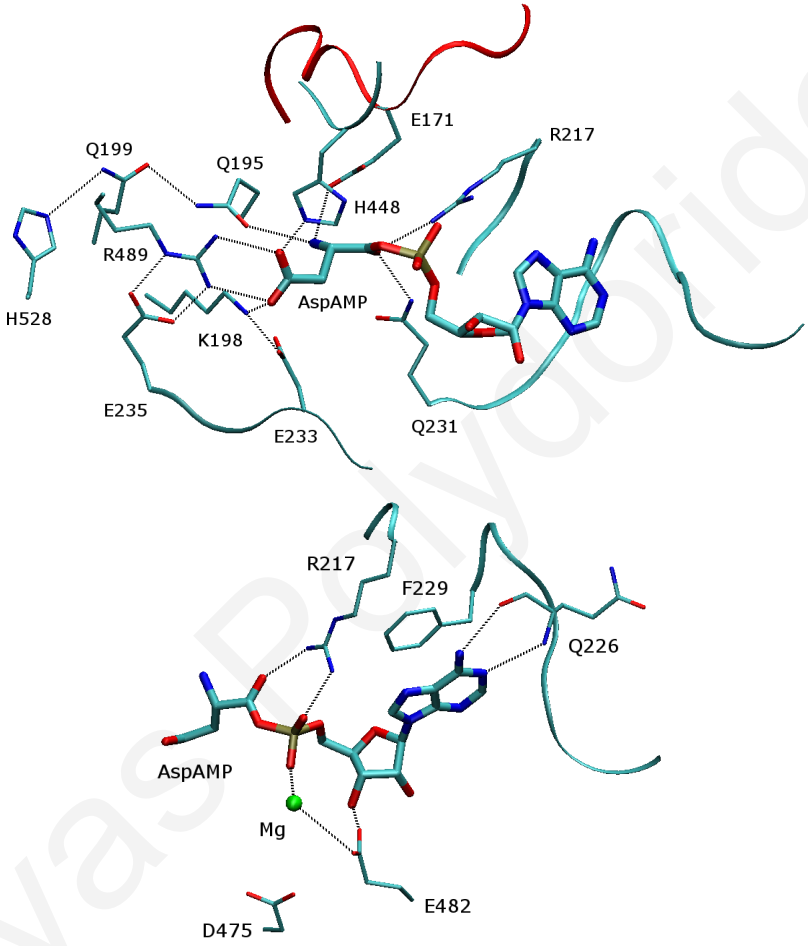


Figure 8.18: The active site of AspRS:AspAMP complex from *E. coli*. The top panel indicates (dash lines) the most important interactions of the Asp sidechain with surrounding residues. The bottom panel shows the most important interactions of the AMP moiety. The flipping loop is represented by a red ribbon.

8.3.3 Binding Specificity of AspAMP and AsnAMP

Although the two proteins belong to the same category of synthetases, with similar structures, they have a few important differences which enhance the binding specificity of each native complex. The Arg489/Glu235 pair of AspRS, which interacts with the Asp ligand carboxylate) is conserved in an identical conformation in AsnRS (pair Arg368/ Glu227). Glu235 (Glu227 in AsnRS) fixes the Arg368 (Arg489) side chain by

two direct hydrogen bonds. In the AspRS:AspAMP complex, the ligand carboxylate is positioned face-to-face with Arg368, forming two simultaneous hydrogen bonds. This is unfavorable conformation for the asparagine ligand. Therefore, in the AsnRS:AsnAMP complex the ligand retains one hydrogen bond with Arg225, but also rotates and interacts with Glu225. Thus, the Glu225 favors the binding of asparagine and disfavors the aspartic acid, improving the specificity of the AsnRS. Glu225 is replaced by Asp233 in AspRS, an amino acid with the same charge but a smaller side chain, that is easily placed near the Lys198 and away from the negatively-charged ligand.

Three residues (Glu164, Ser185 and Gln187) orient the ammonium group of asparagine, which changes orientation when the ligand rotates (see above), preserving only the two interactions of AspRS (with Glu171 and Ser193 in AspRS).

The active site of AsnRS:AsnAMP favors the replacement of the long Lys198 sidechain (in AspRS) by a small hydrophobic side chain (Ala190); The Glu191 side chain of AsnRS holds in place the adjacent Gln187, to form a hydrogen bond with the asparagine ligand ($\alpha-$ amino); in the AspRS:AspAMP complex it changes to its polar analogue Gln199, which forms a direct hydrogen bond with the carboxyl oxygen of aspartic ligand, due to the ligand rotation. These differences in the active sites of AsnRS and AspRS are systematically conserved in corresponding synthetases from other organisms.

Table 8.2 lists the equivalent active site residues in AsnRS (Tt) and AspRS (Ec). The differences between the two structures explain how the synthetases can differentiate between the two ligands. Their ligand-specific hydrogen-bonding network interactions disallow the binding of the non-cognate ligand. For example, placing the AsnAMP ligand in the active site of AspRS restores the neutral state of His448, disrupts the Asp ligand hydrogen bond with the Lys198. In the native AspRS:AspAMP complex, the negatively charged Glu171 stabilizes the doubly protonated state of the crucial His448 side chain, enhancing the aspartic acid preference over its neutral analogue asparagine through a salt bridge with the carboxylate of the ligand. The main magnesium ion present in the AspRS:AspAMP complex increases the protein's specificity for the native ligand. The protein preference for AspAMP versus AsnAMP is diminished in the absence of Mg^{++} [140].

These considerations show that the design of new amino acid specificites in the synthetases is an especially challenging problem. In Chapter 9, we undertake the design of aspartic-acid specificity into the protein AsnRS.

Chapter 9

CPD with a GB Solvent Model: Application to Asparaginyl - tRNA Synthetase

The protein Asparaginyl-tRNA synthetase (AsnRS) belongs to the family of aminoacyl-tRNA synthetases (aaRSs), which were analyzed in detail in Chapter 8. These proteins catalyze the first step in the translation of the genetic code by attaching a specific amino acid to a cognate tRNA molecule [191; 192]. The specificity of aaRSs for their amino acid and tRNA ligands is crucial for the correct translation of the genetic code [192; 193]. Several groups have investigated the contributions of various residues to aaRS binding and catalysis, and have engineered aaRSs with modified amino acid specificity [93; 133; 194; 195]. *In silico* site-directed mutagenesis and free energy simulations studied the amino acid specificity of aspartyl-tRNA synthetase (AspRS) [139–141; 196]. Such studies contribute to our understanding of aaRS function and can lead to engineered organisms with a modified genetic code [93].

Chapter 6 presented recent CPD calculations with the polar-hydrogen/CASA model focused on the AsnRS complex with the non-native ligand Aspartyl adenylate (AspAMP) [77]. This work explored five active-site positions (residues 187, 190, 225, 227, 366 in *Thermus thermophilus* AsnRS [197]), seeking to identify sequences/structures with low folding free energies (high stabilities). Molecular dynamics (MD) simulations of selected designed sequences and Poisson-Boltzmann Free Energy (PBFE) calculations showed that the AsnRS specificity was reversed (AspAMP binding was favored by 11–37 kcal/mol over the natural substrate asparaginyl adenylate, or AsnAMP). However, the computed AspAMP affinities were substantially worse than the native AsnAMP affinity. Furthermore, in the simulations the active site structures became distorted with respect to the native AsnAMP complex, with a bent ligand geometry.

Because of these shortcomings, in the present thesis we reconsider the AsnRS aminoacid specificity problem. Here, we use the improved residue-GB solvent treatment of Chapter 5, together with various design criteria of Chapter 7; these criteria take into account not only stability, but also affinity or specificity. We describe protein

and peptide interactions by an all-atom energy model [92] and treat solvent effects implicitly by the GB/HCT formulation [90]. Two recent advances make this model attractive for CPD calculations. First, we have shown that a careful parameterization of the GB/HCT [90] approximation can yield accurate protein solvation free energies and free-energy changes due to mutations in fully or partly buried positions [76]. The accuracy of this model was also tested further by binding-affinity calculations for several point mutants of the Aspartyl-tRNA synthetase and Tyrosyl-tRNA synthetase (Suplementary Material of Ref. [198] in Appendix A). Second, we recently introduced an accurate residue-pairwise variant of the GB model [91], which is suitable for CPD. We used this model successfully to study acid/base equilibria in proteins (Appendix B, [181]). A major goal of the present thesis is to check the performance of this GB model in a challenging CPD calculation.

A second objective and novelty of the present work with respect to the previous design of the AsnRS:AspAMP complex [77] is that we compare the performance of three design criteria: a maximum stability criterion, which minimizes the folding free energy of the complex; an (absolute) affinity criterion, which minimizes the AsnRS binding free energy for the AspAMP ligand; finally, a relative affinity criterion, which optimizes binding of AspAMP relative to AsnAMP. These criteria were analyzed in detail in Chapter 7.

The sequences suggested by the present CPD calculations are more consistent with the properties of the AspRS active site, compared to the results of the earlier, polar-hydrogen/CASA design [77]. The combined stability/affinity criteria often predict the insertion of a charged (Lys) or polar (His) sidechain at position 187, which form new ligand interactions. The Lys187-containing sequences have an AspRS-like ligand-recognition mode, in which the ligand carboxylate interacts with the key residue Arg388 (Arg489 in *E. coli* AspRS) and with Lys187 (Lys198 in *E. coli* AspRS). The two combined criteria predict, respectively, eleven and twelve sequences that bind AspAMP more strongly than AsnAMP.

To assess the quality of the designed sequences, we study selected complexes with AspAMP or AsnAMP by explicit-solvent molecular dynamics simulations and Poisson-Boltzmann binding free energy calculations (PBFE). The active-site conformations of the designed complexes are well maintained and the ligand-recognition mode of Lys187-containing AsnRS sequences is AspRS-like [189]. Furthermore, the sequences are predicted to have an inverted specificity, favoring Asp.

The AspRS-like protein-AspAMP interactions in the AsnRS active site, the conformational stabilities of the designed sequences in the MD simulations and their increased relative (Asp - Asn) affinities suggest that the GB-HCT implicit-solvent treatment and the combined stability/affinity criteria constitute improvements over our earlier stability/CASA design [77]. This is the main result of the present work. To test further the success of our design, we performed activity measurements with the seven most promis-

ing sequences. Unfortunately, the tested sequences could not catalyze the adenylation reaction of Asp or Asn with ATP.

9.1 Methodology

9.1.1 Effective Energy Function

Protein and ligand interactions were modeled by the AMBER all-atom energy function (Section 3.1.1 and Ref. [92]). Solvent effects were taken into account by the residue GB approximation (Section 4), with the GB/HCT variant (Section 4.1.2 and Ref. [90]) and parameters recently optimized for protein-design calculations [76]. The approximate residue-GB treatment of Section 5.3 was employed. A protein/ligand dielectric constant $\epsilon_p = 8$ and a solvent dielectric constant $\epsilon_w = 80$ were used in the electrostatic interactions.

9.1.2 System

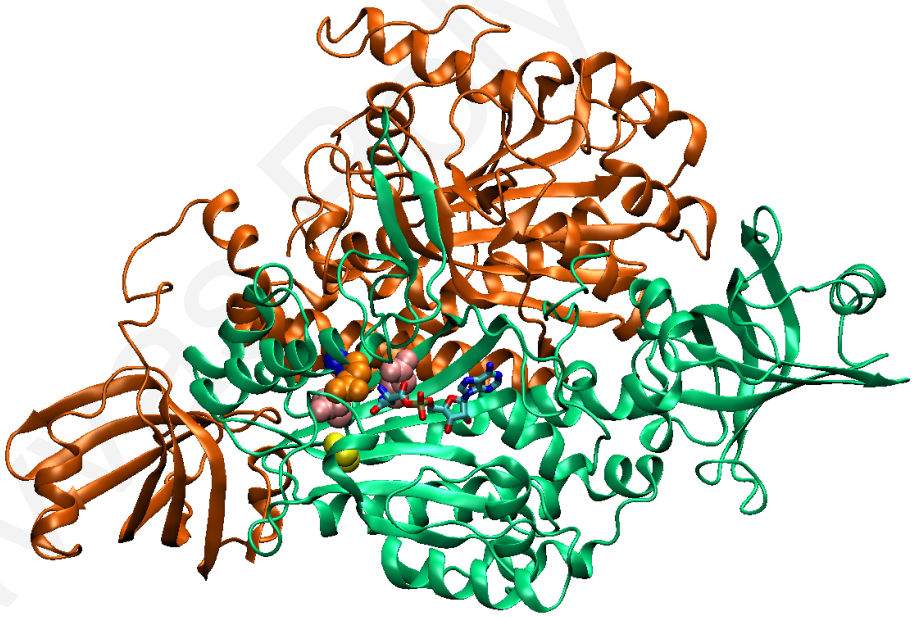


Figure 9.1: The Asparaginyl-tRNA synthetase homodimer, in complex with the AsnAMP ligand (shown in thin licorice). The five mutatable (active) residues are indicated by vdW representation. The system corresponds to a 20Å sphere, centered around the AspAMP ligand including the five active residues Gln187, Ala190, Glu225, Glu227 and Ser366.

The calculations employed the protein AsnRS of *Thermus thermophilus* (Tt). This protein is a homodimer with 438 amino acids per monomer. The protein is too large to be entirely included in the calculation of the residue-pair interaction-energy matrices. To render the calculations efficient, we employed a subset of the protein, corresponding

to a 20 Å-radius sphere centered on the AsnAMP ligand. The main chain atoms of the protein (including the C_β atoms and the sidechains of Cys, Pro and Gly residues) constituted the "fixed" part of the design; that is, their chemical type was invariant, and their atomic coordinates were fixed to their crystallographic position. Five positions near the ligand were selected as "active": 187 (Gln in native AsnRS), 190 (Ala), 225 (Glu), 227 (Glu) and 366 (Ser). These positions were allowed to change chemical states (side-chain type and orientation); the 216 chemical states employed are listed in Table 7.1. All other sidechains were "inactive"; i.e. they maintained the chemical type of the native sequence, but were allowed to explore different conformations. The AsnAMP ligand also was allowed to explore 161 distinct rotamer orientations (see Fig. 9.2).

Non-bonded interactions were truncated at a distance of 15Å. Solvent effects were taken into account by the residue GB / HCT approximation [91], with parameters recently optimized for protein design calculations [76]. The performance of these parameters for chemical mutations and rotamer changes was tested in [199] (see Section 5). A protein / ligand dielectric constant $\epsilon_p = 8$ and a solvent dielectric constant $\epsilon_w = 80$ were used in conjunction with the Coulomb/GB interactions.

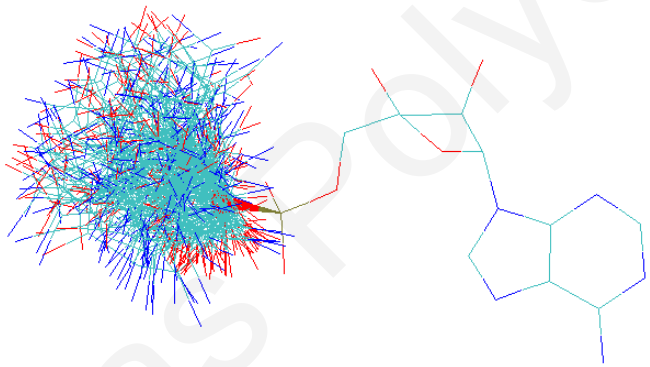


Figure 9.2: The ligand is represented by 161 distinct rotamer orientations.

9.2 Computational Design

9.2.1 Calculation of the Interaction Energy Matrix

The interaction-energy calculation was done by the program XPLOR, using in-house input files.

Our computational design (CPD) procedure consisted of two stages. In the first stage, we partitioned the ligand and protein into segments. The AspAMP and AsnAMP ligands were divided into five segments, corresponding to the adenine base, the ribose sugar, the phosphate backbone, the amino acid backbone and the amino acid sidechain. Gly, Ala, Cys and Pro amino acids were considered as a single segment; all other amino

acids were divided into a backbone segment (including the C_β atom) and a sidechain segment. For each segment, we computed the corresponding GB self-energy, with the approximation that the rest of the molecule had the native sequence and conformation. The resulting self-energies and the corresponding residue coefficients [B_R; see Eq. (5.9)] were stored. In the second stage, we computed the interaction energies between pairs of sidechain and backbone segments, or between pairs of sidechain segments, taking into account all posible sidechain chemical types and orientations. In this calculation, the interaction GB energy terms [Eq. (5.11)] employed the residue coefficients derived in the first stage.

The chart of Fig. 9.3 illustrates the computation procedure. In the first stage, we computed residue self-energies as described in Section 5.1. In the residue-GB approximation, all atoms within a particular residue R are assigned a common residue-GB solvation radius B_R. For this approach to be useful, individual residues should not be very large (so that the environment of the residue atoms does not vary significantly, justifying the common radius). To satisfy this requirement, we partitioned the protein and ligand into segments. Gly, Ala, Cys and Pro amino acids were considered as single segments; all other amino acids were divided into a backbone segment (including the C_β atom) and a sidechain segment. The AspAMP and AsnAMP ligands were divided into five segments, corresponding to the adenine base, the ribose sugar, the phosphate backbone, the amino acid backbone and the amino acid sidechain.

The GB self-energy of each backbone and side chain residue was calculated by Eq. (5.7) of Section 5.1, and was inverted [Eq. (5.9)] to compute the corresponding residue solvation radius B_R. Subsequently, the stored solvation radii were used to compute GB residue-pair interaction energies between all possible pairs of sidechain–backbone or sidechain–sidechain segments, by Eq. (5.11). This calculation was performed by XPLOR scripts, which also computed all other residue-pair interaction energy terms of the AMBER energy function. The scripts contained nested loops over active positions, and chemical states (amino acid chemical types and rotamer conformations) of individual amino acids or amino acid pairs. For each chemical state, the side chain conformations were subjected to 15–30 steps of conjugate-gradient minimization while the backbone was kept fixed.

For each active (inactive) position i, all possible chemical types/rotamers (rotamers) were minimized by fifteen steps with the Powell conjugate-gradient algorithm in the presence of the fixed protein backbone. During the minimization, solvent effects were included by scaling the Coulomb energy with a constant dielectric factor $\epsilon_p = 8$. At the end of the minimization, the interaction energy between the sidechain and the backbone was computed in the residue-GB approximation and stored into a file. Subsequently, the interactions between the sidechain pairs (i, j) were considered.

To increase computational efficiency, interactions between residues in a pair were neglected if the distance between the corresponding C_β atoms was larger than a cutoff

distance of 15 Å. This was based on the assumption that the inter-residue interaction decays fast and becomes small beyond the cutoff distance. The residue pair interaction was also neglected if the minimum distance between any two sidechain atoms of the given residue pair was greater than 12 Å; otherwise, the pair was subjected to thirty minimization steps. During the minimization, solvent effects were included by scaling the Coulomb energy with a constant dielectric factor $\epsilon_p = 8$. At the end of minimization, the interaction energy between the pair was computed in the residue-GB approximation and stored in a file. All calculations were performed with the XPLOR program [183], using in-house scripts. The ligand-protein interactions were also tabulated.

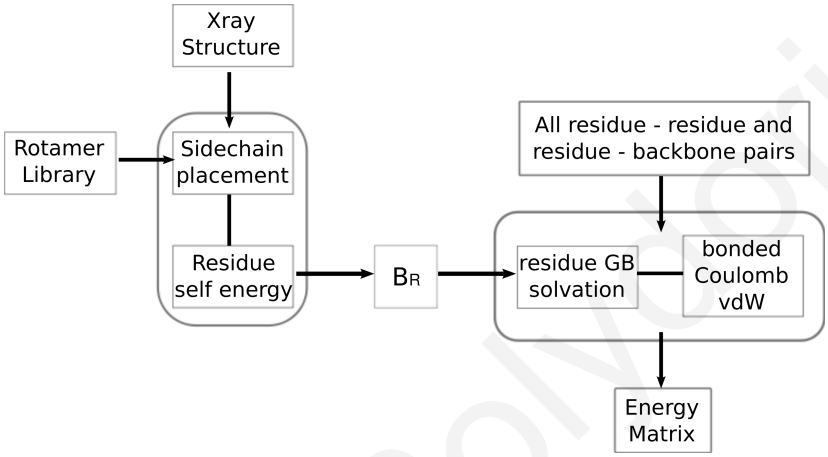


Figure 9.3: Flowchart of the interaction energy matrix computation.

9.2.2 Calculation of the Unfolded State Reference Free Energies

We employed the independent-amino acid approximation of the unfolded state (Section 7.2.3). To compute the reference free energies $\{G_X^{\text{ref}}\}$ for the $X = 18$ chemical types (Table 7.1), we assumed that the side chain X was part ot a tripeptide with sequence Ala-X-Ala (Fig. 7.7) [73]. For each type X, we extracted a large number of tripeptide backbone conformations requiring only the middle segment to be X, from six proteins (lysozyme (PDB ID code 2LZM), bovine pancreatic trypsin inhibitor (4PTI), staphylococcal nuclease (1STN), a-toxin (1PTX), ribonuclease A (2RN2) and cyclophilin (2CPL). On each of these backbone conformation we fitted all rotamers compatible with type X (Table 7.1); for each backbone-side chain combination we performed 140 steps of conjugate-gradient minimization, keeping the backbone fixed. The interaction energy of a side chain with itself, its backbone and the solvent was computed for each member of the collection. The minimum-energy backbone / side chain rotamer combination was selected as the optimum tripeptide structure of the unfolded state for each amino acid type X. The optimum structures of all amino acid types were

Table 9.1: Amino acid energies corresponding to the unfolded state (in kcal/mol)

Amino Acid	Energy
Ala	1.60
Asp	-5.67
Asn	-4.16
Arg	-17.80
Glu	-0.98
Gln	-0.26
His	20.14/24.13[a]
Ile	9.27
Leu	7.93
Lys	8.44
Met	3.76
Phe	3.91
Ser	1.84
Tyr	4.22
Thr	-1.24
Trp	5.95
Val	7.62

[α] Neutral/Protonated Histidine.

used to calculate the reference energies G_X^{ref}, taking into account only the interactions of a particular side chain X with itself, its own backbone and the solvent; as in the folded-state calculations, solvent effects were described by the GB/HCT variant [90], with a protein dielectric constant $\epsilon_p = 8$ and a solvent dielectric constant $\epsilon_w = 80$. The resulting reference energies for all amino acid chemical types are listed in Table 9.1.

9.2.3 AsnRS and AspRS Active Sites

Fig. 9.4 (A) shows the active site of the native AsnRS:AsnAMP complex. The strong affinity of AsnRS for the Asn aminoacid ligand is achieved by a network of hydrogen-bonding interactions, involving the AsnAMP side chain carboxylate and proximal residues Arg368 and Glu225. Furthermore, the ammonium group of the ligand interacts with the side chains of residues Ser185, Gln187 and Glu164.

In the set of five "active" residues of the present study, the side chain of Gln187 interacts with the ligand ammonium and Glu191; Ala190 makes a non-polar contact with the C_β atom of Glu227; Glu225 makes a hydrogen bond with Ser185 and contacts Met223, in addition to its hydrogen-bond with the ligand side chain; The Glu227 side chain forms a salt bridge with Arg368, and the Ser366 side chain makes one interaction with the main chain carbonyl of Gln367.

Fig. 9.4 (B) shows the active site of the protein aspartyl-tRNA synthetase (AspRS) from *E. coli*, in complex with its cognate ligand, AspAMP. The ligand carboxylate

forms direct interactions with the key-recognition residues Arg489 and Lys198 and with His448. Residues Arg489 and Lys198 form salt bridges, respectively with Glu235 and Asp233. The ligand ammonium group interacts with Ser193, Gln195 and Glu171.

Sequence alignment of the proteins AsnRS and AspRS shows that the active sites are highly homologous. In particular, Arg368 (tt AsnRS) is homologous to Arg489 (*E. coli* AspRS), Glu225 to Glu235 and Gln187 to Gln195. The main differences between the two proteins are located in three residues: Ala190 (AsnRS) is replaced by Lys198 in (AspRS), Glu225 by Asp233 and Glu191 by Gln199.

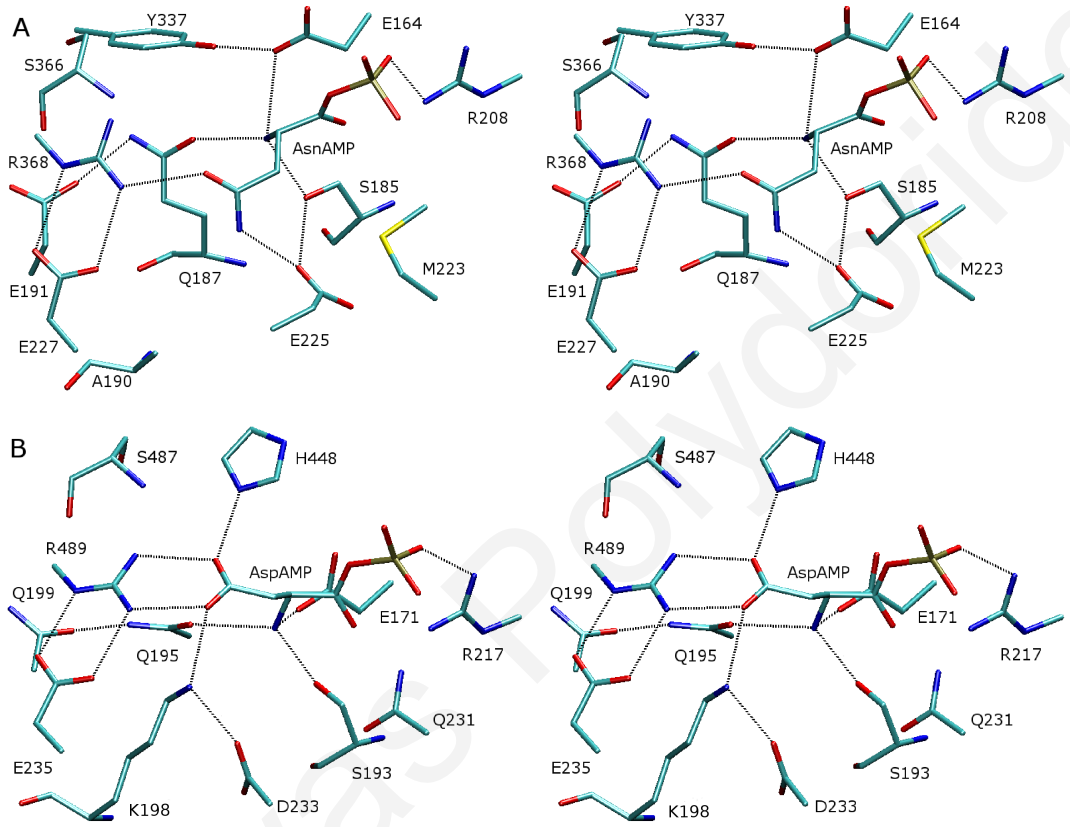


Figure 9.4: Stereo representations of: (A) the active site of the complex *Thermus thermophilus* AsnRS:AsnAMP [197]; (B) the active site of the complex *Escherichia coli* AspRS:AspAMP [190].

9.2.4 Structure Optimization of the AsnRS:AspAMP Complex

In rotamer optimization, the algorithm searches for low-energy conformations while retaining the chemical composition of the target molecule. A stability criterion is employed, which minimizes the free-energy of the folded conformation (the unfolded state has fixed chemical composition and a constant free energy, in the independent-aminoacid approximation of Section 7.2.3).

To optimize the sequence and conformation of the AsnRS:AspAMP complex, we

placed the AspAMP ligand in the AsnRS active site, with its non amino acid moiety at the same position as the corresponding moiety of the AsnAMP ligand and the side chain carboxylate in a very similar orientation as in the active site of the AspRS:AspAMP complex. In this orientation, the carboxylate formed two direct interactions with Arg368. The ligand ammonium group retained its interactions with Ser185 and Gln187, as in the AsnRS:AsnAMP complex. The two ligands AsnAMP and AspAMP are shown in Fig. 7.2

The rotamer optimization calculations of the AsnRS:AspAMP complex employed 20,000 heuristic cycles of the maximum stability protocol, and produced 9,504 different rotameric combinations of the complex, with folding free energies between -795 kcal/-mol and -682 kcal/mol and a mean of -763±16 kcal/mol. The distribution of obtained free energies is plotted in Fig. 9.5; the rotamer conformations of active-site residues, for selected, low-energy conformers, are listed in Table 9.2. Low energy (high stability) conformations have a well-estabilished pattern of rotameric combinations: Residues in the proximity of the ligand (Glu164, Gln187, Arg368) explore a modest number (2-6) of rotamers, while the other active site residues retained their native orientation. Fig. 9.6 shows the conformation of lowest free energy (-794.4 kcal/mol) together with the x-ray structure of the native complex AsnRS:AsnAMP (in thin lines). Arg368 does not maintain its initial position (opposite the ligand carboxylate and near Glu227). Instead, it rotates toward Glu191; it retains one interaction with the ligand side chain carboxylate and forms new interactions with Glu191 and Ser366. Gln187 loses its interactions with the ligand ammonium and Glu191 and forms a new interaction with Glu227. The ligand ammonium compensates its lost Gln187 interaction by improved interactions with Glu164 and Tyr337. Fig. 9.7 shows a second low-energy conformation (-786.3 kcal/mol). Arg368 interacts with Glu191 and Gln187 and packs against the sidechain ring of Tyr337. The sidechain of Gln187 approaches the sidechain of Glu227.

An analogous rotamer optimization was performed in the case of the free AsnRS protein. This calculation produced 7,156 different conformations, with folding free energies between -682 kcal/mol and -604 kcal/mol and a mean of -657±13 kcal/mol. As in the AsnRS : AspAMP complex, residues Glu164, Gln187, Arg208 and Arg368 explore a modest number of rotamer conformations. In the absence of the ligand ammonium, Glu164 forms an electrostatic interaction (salt-bridge) with Arg208. The favored side chain rotamers are similar to the ones in the AsnRS:AspAMP complex, with the exception of Glu164. The most stable conformation is shown in Fig. 9.8.

For completeness we also performed a rotamer optimization of the native AsnRS:AsnAMP complex. The mean folding free energy was -770±14 kcal/mol. The rotamer optimization calculations suggest that some of the residues in the vicinity of the ligand can explore alternative conformations without destabilizing the complex. In particular the key ligand recognition residue Arg368 can rotate away from Glu227 and

Table 9.2: Rotamer optimization of the complex AsnRS:AspAMP.

Energy	E164	S185	Q187	E191	E225	E227	Y337	S366	R368	Ligand
-794.4	6	3	9	4	4	5	6	1	17	140
-791.4	2	3	9	4	4	5	6	1	17	140
-786.3	6	3	4	4	4	5	6	1	8	140
-783.4	2	3	4	4	4	5	6	1	8	140
-783.4	6	3	4	4	4	5	6	1	7	140
-783.4	6	3	4	4	4	5	6	1	14	140
-775.2	6	3	5	4	4	5	6	1	31	102
-773.9	6	3	4	1	4	5	6	1	16	140
-681.7	6	3	4	4	4	5	2	1	39	52

The first row represents the rotamer combination of lowest free energy, identified by Proteus; the last row includes the rotamer combination of minimum stability (intermediate combinations have been omitted). Rotamers are numbered as in the Tuffery library [135].

form a new interaction with Glu191, while maintaining one interaction with the ligand sidechain carboxylate. This suggests that Glu227 can be substituted by non negative residues, improving the protein affinity for AspAMP without a negative impact on stability. The reorientation of Arg368 near Glu191 leaves room for Gln187 to move toward Glu227 and Glu225 and / or accept chemical substitutions, as will be shown below.

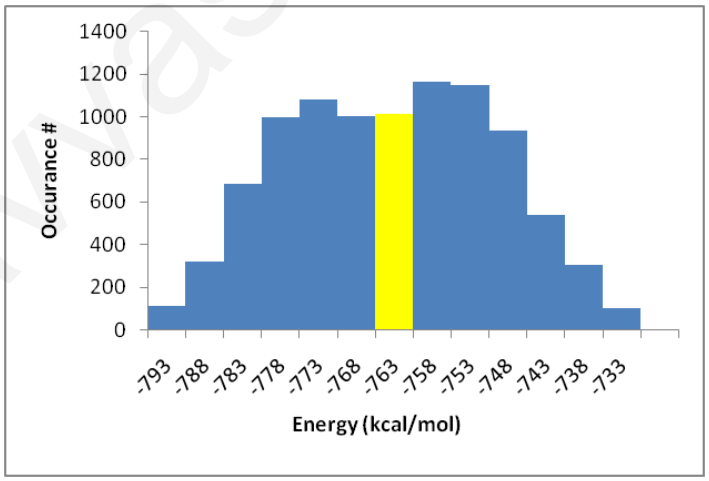


Figure 9.5: Energy profile of the rotameric structures for the wild-type AsnRS:AspAMP produced by Proteus with maximum stability criterion. The mean energy -763 kcal/mol is indicated by the yellow bar.

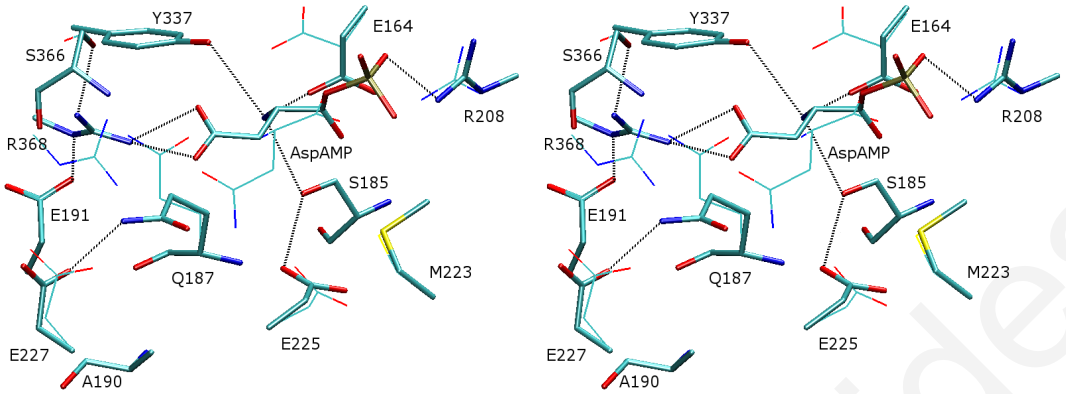


Figure 9.6: Stereo representation of the conformation of maximum stability (folding free energy =-794.4 kcal/mol) for the complex between native AsnRS and the AspAMP ligand. In thin lines is shown the crystallographic conformation of the AsnRS:AsnAMP complex.

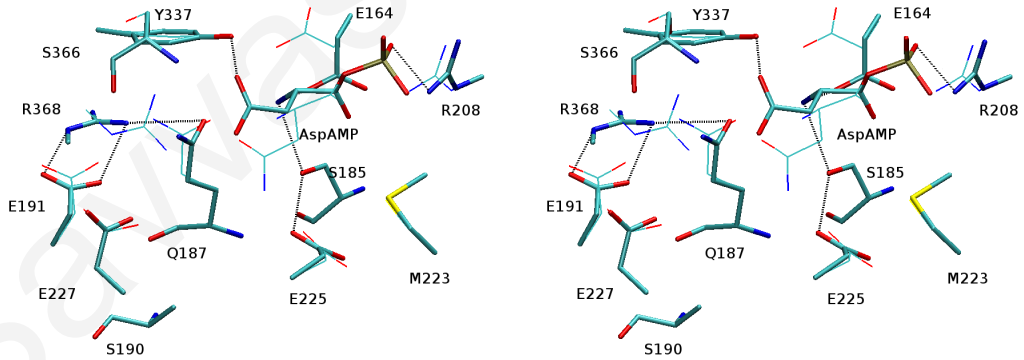


Figure 9.7: Stereo representation of the third highest-stability conformation (folding free energy =-786.3 kcal/mol) of the complex between native AsnRS and the AspAMP ligand. In thin lines is shown the crystallographic conformation of the AsnRS:AsnAMP complex.

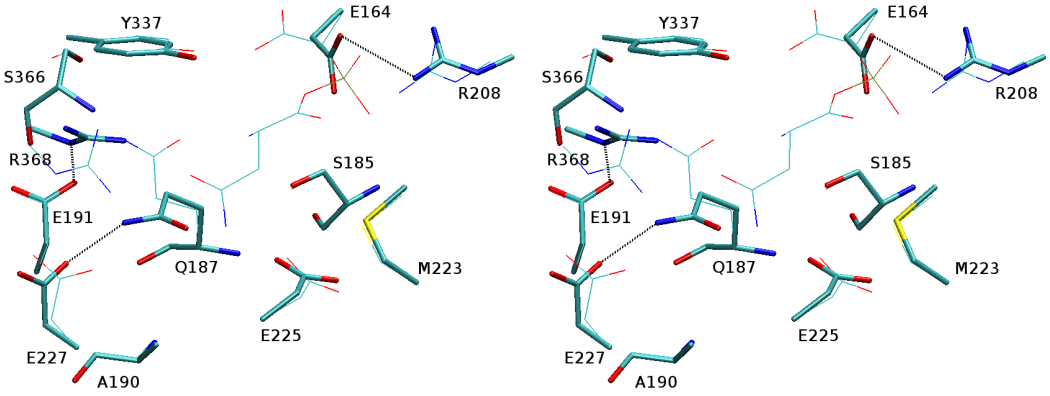


Figure 9.8: Stereo representation of the conformation of maximum stability (folding free energy =-657.0 kcal/mol) of the *free* native AsnRS. In thin lines is shown the crystallographic conformation of the AsnRS:AsnAMP complex.

9.2.5 Sequence / Structure Optimization with the Criterion of Maximum Stability

Our first application involved the design of sequences / structures with low folding free energies of the free AsnRS protein, or the target complex AsnRS:AspAMP. In the case of the free protein, we conducted 100,000 cycles of heuristic design, which identified 44,376 distinct combinations of sequence/conformations involving 31 sequences with foding energies between -675.9 and -782.7 kcal/mol. Table 9.3 lists sequences with lower folding free energies than the native sequence (-657.0 kcal/mol), sorted in terms of decreasing stability. Each value is averaged over the rotamer conformations of the corresponding sequence. According to the bar chart of Fig. 9.9 the most frequent solution corresponds to sequence is R187 / S190 / E225 / M227 / S366 (14319 rotamer conformations), with a mean energy of -669.0 kcal/mol. Other highly populated sequences are RSESS, VSEMS, YSEMS, QSESS. BLOSUM62 [200] scores (last column of the table) assess the similarity of the designed sequences, with respect to the native sequence (QAEES). The sequence Q187 / S190 / E225 / S227 / S366 has the highest similarity with the native molecule (a score of 15); it contains differences in two positions (A190S and E227S), which lower the average folding energy by -13.8 kcal/mol. The sequence V187 / S190 / R225 / E227 / S366 has the lowest average folding energy (-675.9 kcal/mol) but differs in three positions from the native sequence, and is less frequently found (171 rotameric combinations).

Additional design calculations optimized the folding free energy of AsnRS in the presence of the target ligand aspartyl-adenylate (AspAMP). We conducted 100,000 design cycles, which identified 26028 rotameric conformations with 31 distinct sequences. The resulting sequences are listed in Table 9.4, sorted in terms of folding free energy. A histogram of the frequencies of observed sequences is plotted in Fig. 9.10.

Table 9.3: Sequences of the free AsnRS protein, designed with the maximum stability criterion.

Sequence						
187	190	225	227	366	Energy	Score
		Native:				
Q	A	E	E	S	-657.0 ± 13.2	23
		Designed:				
V	S	R	E	S	-675.9 (11.7)	8
R	S	E	S	S	-673.9 (13.0)	11
Y	S	R	E	S	-672.7 (13.0)	9
Y	S	E	E	S	-672.6 (13.0)	14
Q	S	E	S	S	-670.8 (13.2)	15
Y	S	E	M	S	-670.6 (12.8)	7
Q	S	R	S	S	-670.5 (13.1)	10
V	S	E	M	S	-670.0 (13.1)	6
V	S	R	M	S	-669.6 (3.9)	1
R	S	I	S	S	-669.3 (12.9)	3
R	S	E	M	S	-669.0 (13.1)	9
Y	S	I	E	S	-668.3 (12.7)	6
Q	S	I	S	S	-667.7 (13.1)	7
V	S	I	M	S	-667.7 (12.8)	-2
R	S	I	E	S	-665.5 (12.4)	8
R	S	I	M	S	-657.5 (12.8)	1

For each sequence, the reported values are averaged over all rotamer conformations of the Proteus optimization. Standard deviations (over rotamer conformations) are included in parentheses. Sequences with higher average folding free energies than the native are omitted.

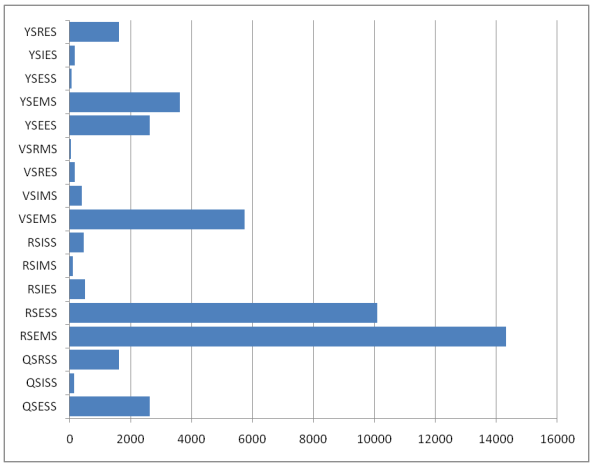


Figure 9.9: Probability histogram of the low-free energy sequences of AsnRS, produced by 100,000 heuristic cycles with a maximum stability criterion.

In general, the obtained sequences are similar with those of the free AsnRS design (Table 9.3). All optimized sequences have a serine residue at positions 190 (replacing alanine) and 366 (as in the native sequence). As will be shown below, this is a persistent feature encountered in optimizations with all criteria. In the native sequence, the Ala190 sidechain makes a non-polar contact with the C_β atom of Glu227. The hydroxyl group of the inserted serine at position 190 makes an additional hydrogen bond with the main chain carbonyl group of the residue at position 225.

The most stable sequence (YSEES) differs from the native sequence in two positions (Q187Y, A190S) and has an average folding energy that is lower by 19.4 kcal/mol. Other low free energy sequences have the pattern S190 / E225 / M227 / S366 in the last four positions, combined with a tyrosine (-781.9 kcal/mol) or a valine (-781.5 kcal/mol) at position 187. The most frequently produced sequence is KSESS with mean energy -775.9 kcal/mol. Other highly populated sequences are QSESS, YSEMS, VSVMS and KSEES.

Position 187 has the maximum variability in the design: it can stay invariant (serine-Q), become aromatic (tyrosine-Y), positively charged (arginine-R), or non-polar (valine-V). Glu225 stays unchanged, but can also be replaced by serine (S) or methionine (M). The Arg368 sidechain occupies two possible orientations, which depend on the chemical type of residue 227. If the native glutamic acid is maintained at this position (sequences YSEES, YSIES, YSRES), the Tyr187 ring is inserted between the sidechains of the ligand and the Arg368, and the Arg368-ligand interaction is disrupted. This is shown in Fig. 9.11 where the conformation of sequence YSEES is displayed against the native conformation of highest stability of Fig. 9.6 (in thin lines). The sidechain of Arg368 occupies a position that is different from its position in the crystal structure or in the structure of lowest stability; it moves toward Glu191 and retains two interactions with Glu227. Glu191 moves also slightly backwards to accomodate the Tyr187 ring and the new Arg368 position. Other residues are oriented as in the structure of highest stability. Obviously, the insertion of the tyrosine ring at this position may contribute to the total stability, but eliminates the Arg368-ligand interaction, preventing thereby the specific recognition of the aspartic acid carboxylate sidechain by the protein.

If a serine (S) or methionine (M) is present at position 227 (with the exception of sequence RSESS), Arg368 maintains an orientation similar to what is observed in the structure of lowest stability. An example is shown for sequence QSRSS, Fig. 9.11. The QSRSS conformation is very similar to the highest stability conformation of native AsnRS. The Arg225 sidechain makes a hydrogen bonding interaction with the main chain carbonyl oxygen of the ligand. Sequences QSESS and QSISS have also similar conformations. In the designed sequences, a methionine residue at position 227 is combined with an aromatic (Y) or non-polar (V) residue at position 187. The two residues form non-polar contacts that contribute to stability. Mutations that create more extensive non-polar contacts in the active site are also observed. For example,

sequence VSVMS, shown in Fig. 9.12, contains a chain of non-polar contacts involving residues Val187, Met223, Val225 and Met227.

One of the designed sequences (RSESS) contains an arginine residue at position 187. The conformation of this sequence is shown in Fig. 9.12. The Arg187 guanidinium group occupies a similar position as the same group of Arg368 in the native conformation of lowest stability; in this position it forms hydrogen bonding interactions with the sidechains of the ligand, Glu191 and Ser366, and packs against the Tyr337 ring. Tyr337 also forms a hydrogen bond with the ligand carboxylate; Arg368 rotates toward the sidechain of His403 and hydrogen-bonds with the serine sidechain at position 227. In the case of the Q187K mutant (KSEES, KSESS) the inserted lysine interacts with Glu227; Arg368 stays in place. Glu225 remains invariant or changes to non-polar icoleucine. Glu227 remains invariant or changes to serine.

Table 9.4: Sequences of the complex AsnRS:AspAMP, designed with the maximum stability criterion.

Sequence						
187	190	225	227	366	Energy	Score
Native:						
Q	A	E	E	S	-763.3 (15.5)	23
Designed:						
Y	S	E	E	S	-782.7 (13.2)	14
Y	S	E	M	S	-781.9 (13.2)	7
V	S	E	M	S	-781.5 (12.9)	6
Y	S	V	E	S	-780.9 (13.5)	7
Q	S	E	S	S	-780.6 (13.1)	15
V	S	I	M	S	-779.2 (13.2)	-2
V	S	V	M	S	-778.9 (12.9)	-1
Q	S	V	S	S	-778.8 (12.9)	8
Y	S	I	E	S	-778.1 (13.0)	6
Q	S	I	S	S	-777.3 (12.9)	7
Y	S	R	E	S	-777.0 (12.5)	9
K	S	E	E	S	-776.9 (13.7)	16
K	S	E	S	S	-775.9 (14.4)	11
R	S	E	S	S	-772.0 (13.1)	11
Q	S	R	S	S	-771.9 (11.3)	10
K	S	I	E	S	-770.5 (12.3)	8
K	S	I	S	S	-769.9 (14.7)	3
Y	S	E	S	S	-768.7 (13.1)	9
Y	S	R	S	S	-768.0 (13.5)	4
V	S	R	E	S	-767.0 (15.0)	8
R	S	E	E	S	-766.2 (12.9)	16

For each sequence, the reported values are averaged over all rotamer conformations of the Proteus optimization. Standard deviations (over rotamer conformations) are included in parentheses. Sequences with higher average folding free energies than the native are omitted.

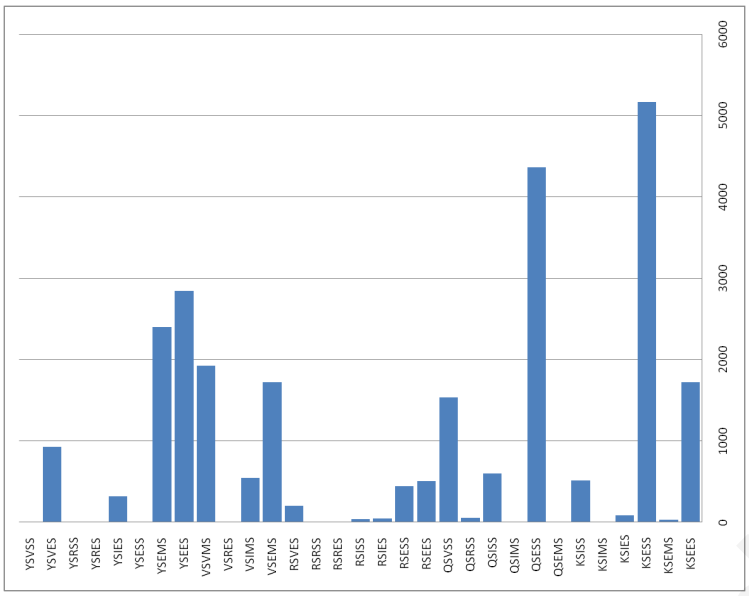


Figure 9.10: Probability histogram of the low-free energy sequences of the complex AsnRS:AspAMP, produced by 100,000 heuristic cycles with a maximum stability criterion. A total of 26,028 non redundant sequence/rotamer combinations were produced (31 distinct sequences).

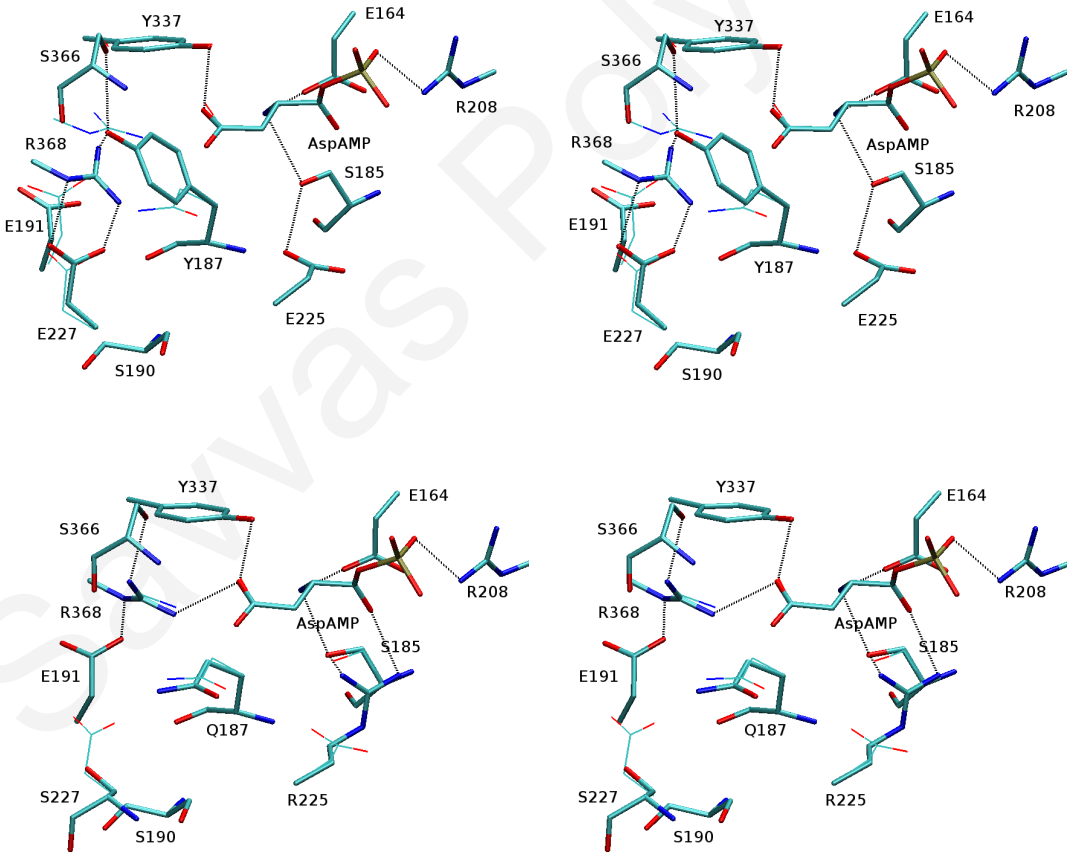


Figure 9.11: Stereo representation of the YSEES (top) and QSRSS (bottom) conformations produced in sequence optimization for the complex AsnRS:AspAMP. In thin lines is shown the conformation of maximum stability for the native AsnRS:AspAMP. The most important interactions are shown in dashed lines.

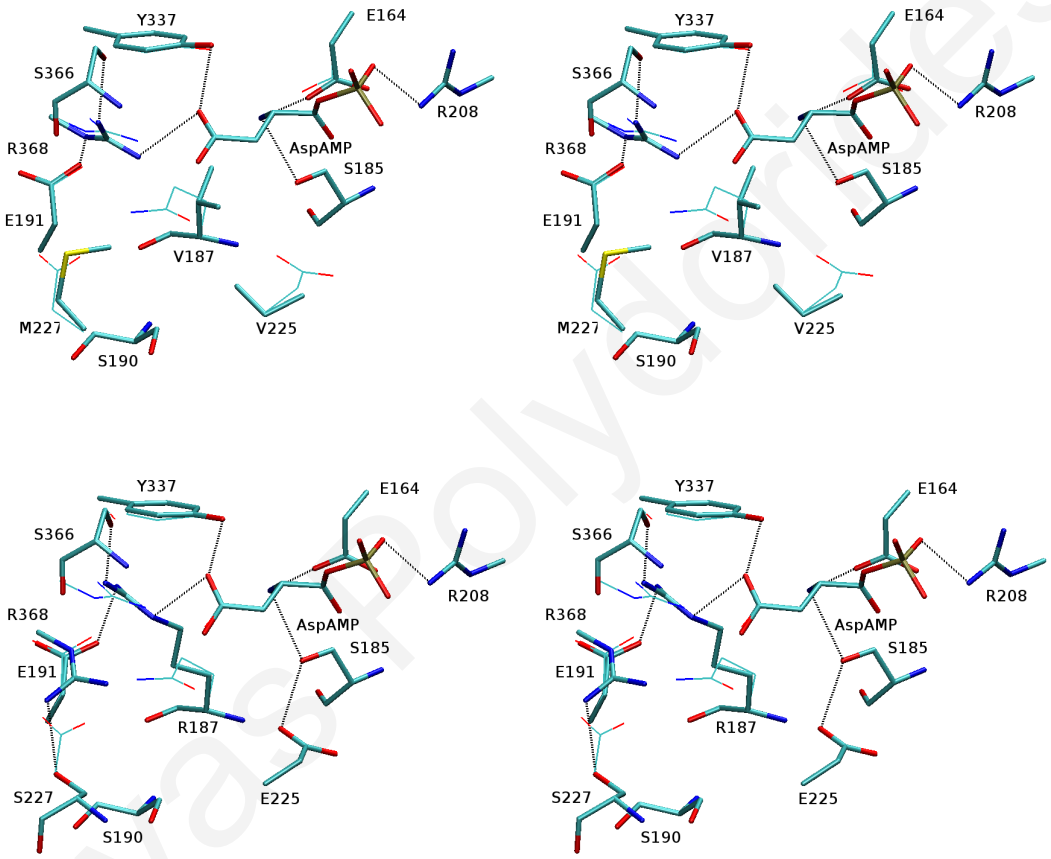


Figure 9.12: Stereo representation of the VSVMS (top) and RSESS (bottom) conformations produced in sequence optimization for the complex AsnRS:AspAMP. In thin lines is shown the conformation of maximum stability for the native AsnRS:AspAMP. The most important interactions are shown in dashed lines.

The average folding free energies of the designed sequences can be used to estimate affinity free energies of AsnRS for the AspAMP ligand, via the relation ($\Delta G = G_{\text{AsnRS:AspAMP}} - G_{\text{AsnRS}}$). In this relation, the constant contribution (from the free energy of the invariant ligand AspAMP) is omitted. The resulting values are listed in the last column of Table 9.5, sorted with respect to affinity. Eight sequences have improved affinity, compared to the native protein (-106.3 kcal/mol). All these sequences (with the exception of QSISS) have a tyrosine or valine at position 187. The best four sequences have a methionine or serine at position 227. The replacement of Glu227 by one of these residues prompts Arg368 to interact with Glu191 and Ser366, while maintaining one interaction with the ligand, and eliminates one negative charge in the vicinity of AspAMP. These factors contribute both to a high stability of the AsnRS:AspAMP complex and a high ligand affinity. The substitution of Glu227 by serine or methionine does not destabilize the free protein (column 3 of Table 9.5), because Arg368 rotates and interacts with Glu191 and Ser366. Introduction of a positive charge at position 225 is not strictly correlated with higher affinity; for example, sequences YSRES, QSRSS, VSRES have low affinities, because they are associated with higher stabilities of the free protein. These results suggest that the protein can gain stability and affinity by introducing an aromatic or non-polar residue at position 187, together with a non-negative residue at position 227. The elimination of Glu227 is tolerated, because Arg368 replaces its salt-bridge with Glu227 by an interaction with Glu191 and maintains one interaction with the ligand carboxylate.

We performed an additional folding free-energy optimization of AsnRS, in the presence of the native asparaginyl-adenylate (AsnAMP) ligand. We run 100,000 heuristic cycles producing 43 distinct sequences (Fig. 9.13). The results are listed in Table 9.6. In general, the presence of the native (uncharged) ligand facilitates the emergence of a somewhat larger number of sequences, presumably due to the weakening of Coulombic interactions. Overall, the obtained sequences are similar with respect to the AsnRS:AspAMP complex, but with different conformations. For example, when Gln187 is mutated to arginine (sequence R187 / S190 / E225 / E227 / S366), it forms two hydrogen bonds with the ligand and another one with Ser366. Tyr337 forms a hydrogen bond with the aspartate (AspAMP). When a lysine is inserted at position 187 (sequence K187 / S190 / E225 / E227 / S366), the interactions are almost the same as in the native structure, except for an interaction between Lys187 and Glu191. Position 187 has higher variability with respect to the complex with AspAMP, with the possibility for insertion of Glu and Hie.

Table 9.5: Sequences of the complex AsnRS:AspAMP and the free protein AsnRS, optimized with a maximum-stability criterion. All values are in kcal/mol.

Sequence							
187	190	225	227	366	AsnRS:AspAMP	AsnRS	ΔG
Native:							
Q	A	E	E	S	-763.3	-657.0	-106.3
Designed:							
Y	S	E	S	S	-768.7	-656.2	-112.5
V	S	E	M	S	-781.5	-670.0	-111.5
V	S	I	M	S	-779.2	-667.7	-111.5
Y	S	E	M	S	-781.9	-670.6	-111.3
Y	S	E	E	S	-782.7	-672.6	-110.1
Q	S	E	S	S	-780.6	-670.8	-109.8
Y	S	I	E	S	-778.1	-668.3	-109.8
Q	S	I	S	S	-777.3	-667.7	-109.6
Y	S	R	E	S	-777.0	-672.7	-104.3
Q	S	R	S	S	-771.9	-670.5	-101.4
R	S	E	S	S	-772.0	-673.9	-98.1
V	S	R	E	S	-767.0	-675.9	-91.1
R	S	I	S	S	-763.4	-669.3	-94.1

Columns "AsnRS:AspAMP" and "AsnRS" list, respectively, the folding free energies of the complex AsnRS:AspAMP and the free protein AsnRS; ΔG is the corresponding binding affinity for AspAMP, obtained by the difference of the preceding two columns. For each sequence, the reported values are averaged over different rotamer conformations obtained in the optimization.

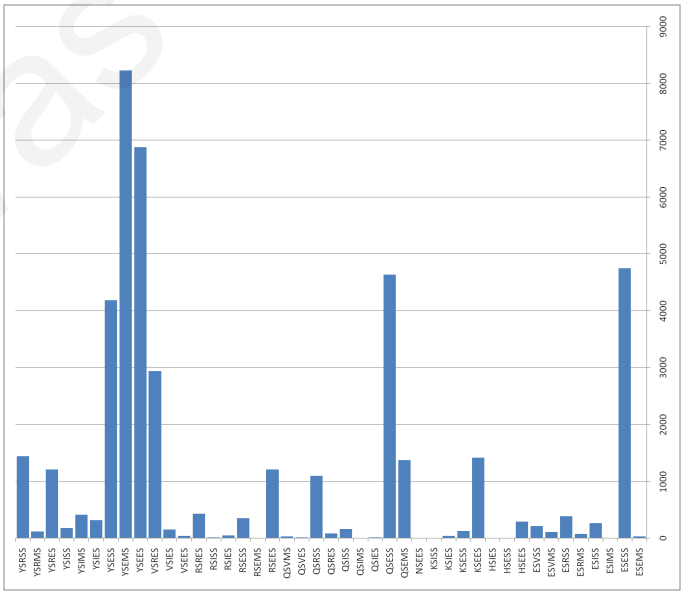


Figure 9.13: Probability histogram of the low-free energy sequences of the complex AsnRS:AspAMP, produced by 100,000 heuristic cycles with a maximum stability criterion. A total of 43,267 non redundant sequence/rotamer combinations were produced (43 distinct sequences).

Table 9.6: Sequences of the native complex AsnRS:AsnAMP, designed with the maximum stability criterion.

Sequence						
187	190	225	227	366	Energy	Score
Native:						
Q	A	E	E	S	-770.6(13.5)	23
Designed:						
H	S	I	E	S	-791.6(0.0)	7
Y	S	R	E	S	-789.1(13.3)	9
Y	S	E	E	S	-786.6(13.1)	14
V	S	R	E	S	-784.8(13.2)	8
Q	S	E	S	S	-784.7(13.2)	15
Y	S	R	S	S	-783.4(13.1)	4
Y	S	E	M	S	-783.3(13.2)	7
V	S	E	E	S	-782.6(12.3)	13
E	S	E	S	S	-782.5(13.1)	12
Y	S	I	E	S	-781.5(12.4)	6
V	S	I	E	S	-781.3(13.0)	5
Y	S	E	S	S	-781.2(13.1)	9
K	S	E	S	S	-780.5(12.2)	11
K	S	E	E	S	-780.3(13.2)	16
R	S	E	E	S	-779.4(13.1)	16
Q	S	R	S	S	-779.3(13.3)	10

For each sequence, the reported values are averaged over all rotamer conformations of the Proteus optimization. Standard deviations (over rotamer conformations) are included in parentheses. A zero standard deviation implies that a single rotamer conformation was found by the design. Sequences with higher average folding free energies than the native are omitted.

9.2.6 Sequence / Structure Optimization with the Criterion of Absolute Affinity

In the next stage of design, we searched for sequences which minimized the binding free energy of AsnRS to the non-native ligand AspAMP. We used the same "active" positions as in the stability calculations (187, 190, 225, 227, 366). The searching protocol was described in Section 7.2.2. As explained in that section, the use of the affinity criterion without consideration of stability may direct the optimization toward sequences with high folding free energies. This was indeed observed in preliminary calculations shown in Table 9.7. All sequences listed in that table have lower stabilities by at least 100 kcal/mol (column 7), even though they also have very improved affinities (column 6).

To obtain sequences with high affinities and stabilities, we optimized the weighted sum $w_\alpha \, \Delta G + w_s \, G^*$, where ΔG was the binding free energy between the complex and the protein, ligand free states, and $G^* \equiv (G_P + G_{PL})/2$ was the arithmetic mean of the free-protein and complex folding free energies.

We started with affinity / stability weights $w_\alpha = 0.25$, $w_s = 0.75$, and progressively increased the affinity weight w_α to a value of 0.9, decreasing respectively the stability weight $w_s = 1 - w_\alpha$. With small values of the affinity weight w_α, we obtained very similar sequences as in the stability optimization of the previous section. On the other hand, the introduction of small stability weights ($w_s = 0.1 - 0.25$) was sufficient to yield sequences with high affinity for AspAMP and high stability (with $w_\alpha = 1$ we got the sequences listed in Table 9.7). Since we are trying to design sequences according to the absolute affinity criterion we focus on the last two cases of increased affinity weight, 0.75 and 0.90. We performed 100,000 heuristic cycles for each set of weights (w_α, w_s) =(0.75,0.25) and (0.9,0.1) resulting in 92 and 188 different sequences succeeding the first filtering step. For each sequence, we computed the average binding free energy $\langle \Delta G \rangle \equiv \langle G_{PL} \rangle - \langle G_P \rangle$ and the average of the arithmetic mean of the folding free energies $\langle G^* \rangle = \langle (G_{PL} + G_P)/2 \rangle$ (the averages $\langle \cdots \rangle$ are over all rotamer conformations of each sequence, found in the optimization). A sequence was selected, if its average free energies $\langle \Delta G \rangle$ and $\langle G^* \rangle$ were lower than the corresponding values for native AsnRS (-106.3 kcal/mol and -710.2 kcal/mol, respectively).

A large number of preliminary sequences were associated with one rotameric conformation. To identify sequences that maintained high affinities when the protein was allowed to sample a larger number of conformations, the results were post-processed by a filtering analysis in multiple steps. First, for each of the selected sequences we performed a rotamer optimization of the complex between the mutated AsnRS protein and the AspAMP ligand. We produced 10,000 rotamer conformations, reconstructed the 100 lowest energy conformations for each sequence, and minimized them by 90 steps of a Powell conjucated-gradient algorithm, keeping the backbone fixed. Subsequently,

Table 9.7: Sequences of the complex AsnRS:AspAMP, designed with a simple affinity criterion.

Sequence							
187	190	225	227	366	$\langle \Delta G \rangle$	$\langle G^* \rangle$	Score
Native:							
Q	A	E	E	S	-106.3	-710.2	23
Designed:							
W	M	Y	Q	W	-227.6 ± 2.4	-546.8 ± 9.6	-6
H	H	V	T	K	-193.8 ± 3.4	-563.2 ± 15.5	-5
K	H	V	T	K	-191.4 ± 1.9	-562.4 ± 10.9	-4
Y	R	Y	W	R	-177.1 ± 28.1	-323.6 ± 12.6	-8
T	H	V	H	M	-170.9 ± 33.6	-544.8 ± 6.2	-6
I	H	V	Y	R	-149.8 ± 30.3	-371.7 ± 22.5	-10
H	Y	Y	W	K	-148.0 ± 3.7	-367.5 ± 14.7	-7
H	Y	Y	W	H	-147.0 ± 4.9	-360.4 ± 11.5	-8
Y	W	I	R	K	-134.2 ± 2.5	-441.7 ± 18.9	-7
K	W	I	R	H	-133.9 ± 2.4	-447.4 ± 15.3	-6
H	H	V	Y	H	-132.9 ± 2.5	-452.0 ± 11.2	-7
K	W	I	R	K	-131.5 ± 2.5	-452.7 ± 13.7	-5
Q	I	Y	R	K	-130.9 ± 0.0	-621.8 ± 0.0	2
H	H	V	Y	K	-130.4 ± 4.0	-448.7 ± 12.0	-6
H	E	A	V	H	-129.6 ± 0.0	-371.1 ± 0.0	-5
F	H	V	H	K	-128.8 ± 0.0	-534.0 ± 0.0	-7
H	R	Y	D	H	-127.4 ± 0.0	-580.5 ± 0.0	-2
Q	H	V	F	E	-123.0 ± 0.0	-446.1 ± 0.0	-2
H	E	Y	V	K	-122.2 ± 2.3	-607.9 ± 7.6	-5
H	E	Y	V	H	-121.6 ± 1.3	-595.0 ± 20.2	-6

For each sequence, the reported values are averaged over all rotamer conformations of the Proteus optimization. Standard deviations (over rotamer conformations) are included in parentheses. A zero standard deviation implies that a single rotamer conformation was identified by the design.

Table 9.8: Sequences of the complex AsnRS:AspAMP, selected after reconstruction, rotamer optimization and minimization of the sequences designed with affinity filter $w_a = 0.75 - 0.90$.

Sequence							
187	190	225	227	366	$\langle G^* \rangle$	$\langle \Delta G \rangle$	$G^{\min}_{\text{bind}}$
Native:							
Q	A	E	E	S	-736.5	-111.2	-61.4 (9.5)
Designed:							
K	S	T	E	S	-737.8	-112.1	-66.3 (5.1)
K	S	E	E	S	-746.6	-112.0	-64.6 (3.5)
K	S	I	E	S	-741.2	-112.0	-63.3 (4.2)
H	S	T	E	S	-738.6	-112.3	-63.1 (5.0)
H	S	E	K	S	-740.0	-112.0	-63.1 (5.5)
H	S	T	M	S	-736.9	-111.8	-62.6 (6.3)
H	S	V	M	S	-744.5	-111.3	-62.5 (7.0)
Y	S	Q	K	S	-749.0	-111.7	-62.5 (6.5)
Y	S	E	M	S	-739.1	-111.7	-62.3 (5.3)
Y	S	E	K	S	-739.4	-111.7	-61.7 (6.0)
K	S	E	D	S	-737.0	-113.1	-61.6 (4.4)

we continued the minimization for 10 steps in the presence or absence of the ligand. We employed the final minimized energies of the complex and free protein to compute the binding free energy, and averaged the result over the 100 rotamer conformations. The binding free energy of the native sequence is $\Delta G = -61.4 \pm 9.5$ kcal/mol, averaged over the 100 best rotamers and minimized under the same protocol, was employed as a final selection filter, to narrow down the designed sequences. This value set the cut off for the final selection stage.

Table 9.8 lists sequences produced by both affinity / stability weighted optimization functions, passing through all selection filters. Compared to the native protein, all sequences have slightly improved affinities after rotamer optimization; these affinities are further improved after minimization. The average binding energy $\langle \Delta G \rangle$ and the average mean energy of the complex and free protein $\langle G^* \rangle$ are compared to the values obtained for the native sequence, after rotamer optimization.

Positions 190 and 366 are occupied by serines, as before. Position 187 is occupied by a positive (lysine-K) or aromatic residue (histidine-H,tyrosine-Y). The sequences of highest affinity combine the introduction of a lysine or histidine at position 187 with the conservation of Glu227. Alternatively, an aromatic residue at 187 can be combined with a lysine or methionine at position 227. Position 225 is conserved (glutamic acid-E), but can accept polar (threonine-T,glutamine-Q) or non-polar residues (isoleucine-I,valine-V). Fig. 9.14 shows the conformations of sequences KSTES and KSEES, against the lowest energy conformation of the native sequence. The ligand ammonium interacts with Ser185, Tyr337 and Glu164. The ligand sidechain remains in its initial position and interacts with Arg368, Lys187 and Tyr337. Arg368 rotates to a similar position,

as in the lowest energy conformations of the folding optimization calculations, Fig. 9.6. It loses its salt bridge with Glu227, maintains one interaction with the ligand and forms interactions with Glu191 and Ser366. Lys187 forms one direct interaction with the ligand sidechain carboxylate. Interestingly, the position and orientation of a Lys198 (in *E. coli* numbering) in the active site of the protein Aspartyl-tRNA synthetase (see Fig. 9.4). Lys198 is responsible, together with Arg489 (the residue analogous to Arg368 in AsnRS), for the specific recognition of aspartic acid by AspRS, by forming a direct interaction with the ligand sidechain carboxylate. This is shown in Fig. 9.15, where the AspRS active site is superposed against the active site conformation of sequence KSEES. Unlike the active site of AspRS, where Lys198 interacts with Asp233, here Lys187 is distanced from the corresponding aminoacid (Glu225). Among other residues, Glu227 keeps its initial position and interacts with Lys187. Glu191 interacts with His403, Arg368 and Tyr392. The Glu225 sidechain interacts with the Ser185 sidechain and the Lys187 mainchain carbonyl. In the sequence KSTES, Thr225 forms a hydrogen bond with the mainchain carbonyl of Met223. The last groups of sequences contains a histidine or tyrosine residue at position 187, combined with a methionine (as in the stability calculations) or lysine at position 187. Arg368 approaches Glu191 and retains one interaction with the ligand. The aromatic rings of Tyr187 (or His187), Arg368 and Tyr337 are placed parallel to each other, forming pi-stacking interactions. Lys227 interacts with the Tyr187 and His403 sidechain, and makes a more distant interaction with Glu191.

9.2.7 Sequence / Structure Optimization of AsnRS with the Criterion of Relative Affinity

In this case, we searched for sequences which minimized the affinity free energy of AsnRS for the non-native ligand AspAMP, *relative* to the native ligand AsnAMP. We used the same mutation positions as in the stability and absolute affinity calculations (187, 190, 225, 227, 366). Preliminary calculations yielded sequences of very low stability. We thus used the combined relative affinity / stability criterion, described earlier. We performed 50,000 heuristic cycles of the Proteus optimization algorithm producing 86 different sequences. The designed sequences were post-processed by an analogous filtering analysis, as in the absolute affinity calculations. For each optimized sequence, we computed the average relative binding free energy $\langle \Delta G^* \rangle \equiv \langle G_{PL_2} - G_{PL_1} \rangle$ and the average folding free energy of the free protein $\langle G^* \rangle \equiv \langle G_P \rangle$ (averages $\langle \cdots \rangle$ are over the rotameric conformations of each sequence, found during the optimization). A total of 65 sequences were selected, with $\langle \Delta G^* \rangle$ and $\langle G^* \rangle$ values lower than the corresponding values for the native AsnRS sequence, 7.3 kcal/mol, and -657.0 kcal/mol, respectively.

Subsequently, for each of the selected sequences we produced 10,000 optimized rotameric conformations of the complexes AsnRS:AspAMP, AsnRS:AsnAMP, and the

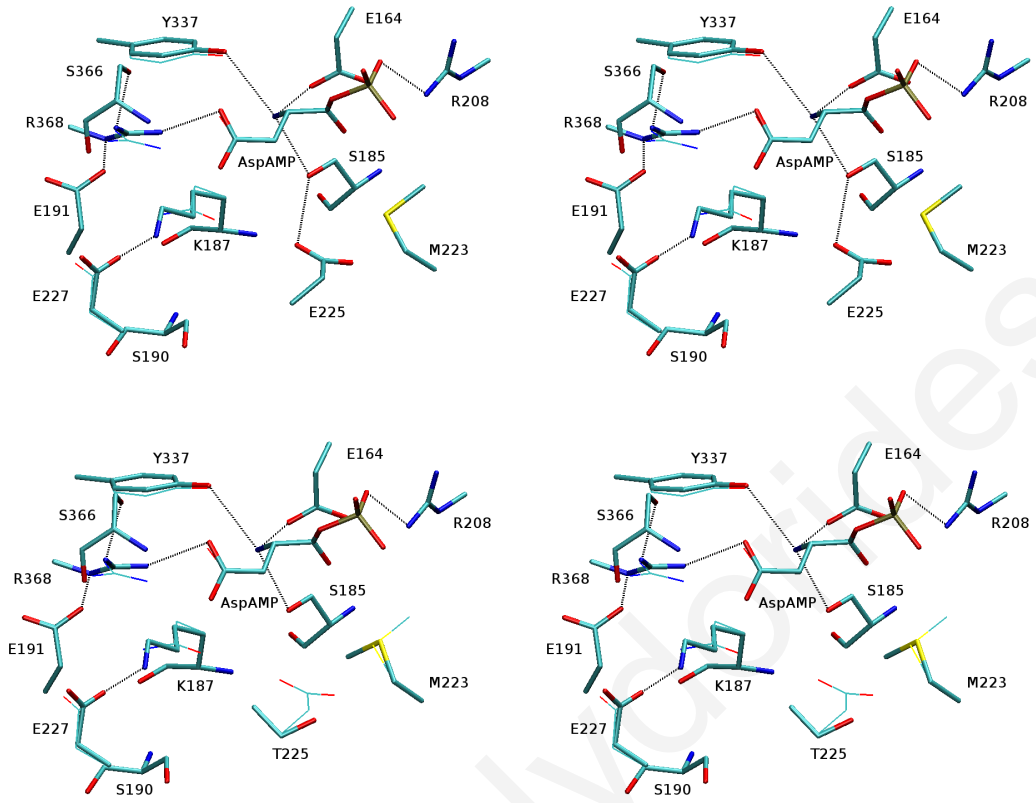


Figure 9.14: Stereo representation of the KSEES (top) and KSTES (bottom) conformations produced in sequence optimization for the complex AsnRS:AspAMP, using the weighted stability / affinity criterion. In thin lines is shown the conformation of maximum stability for the native AsnRS:AspAMP. The most important interactions are shown in dashed lines.

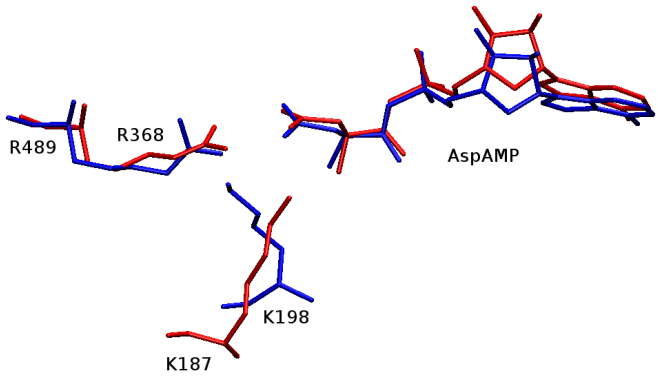


Figure 9.15: Structural alignment of the AspRS:AspAMP complex from (Ecoli- colored blue), with the reconstructed mutant sequence KSEES of AsnRS:AspAMP (from T.Thermophilus - colored red).

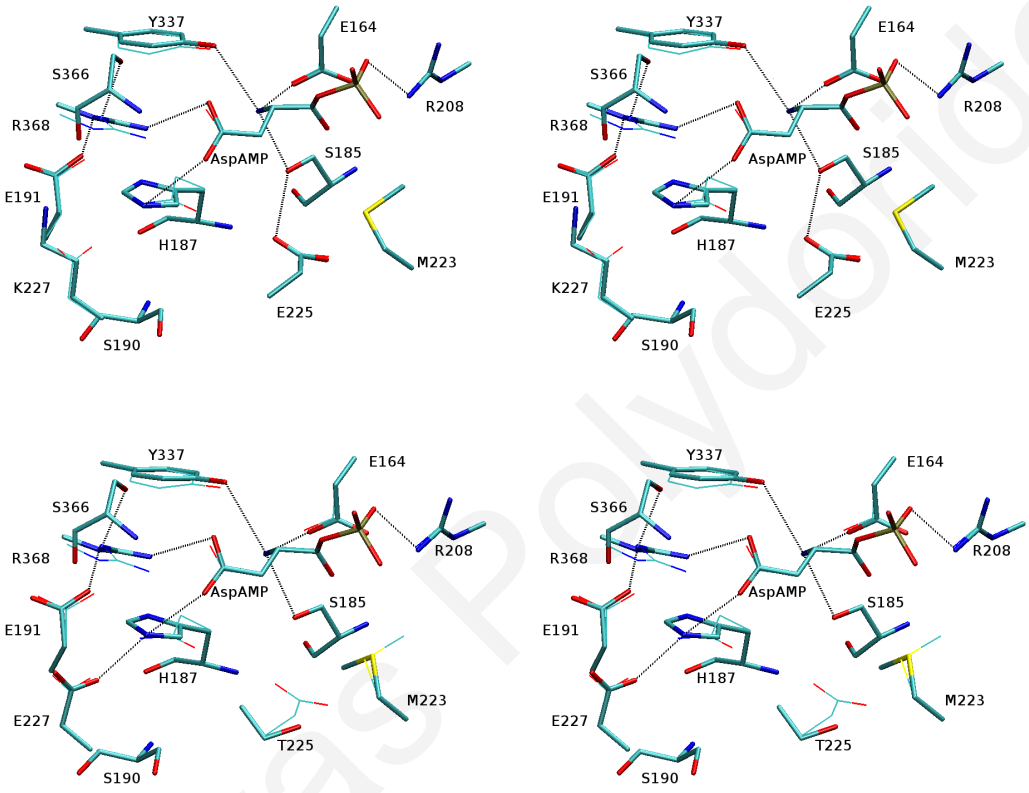


Figure 9.16: Stereo representation of the HSEKS (top) and HSTES (bottom) conformations produced in sequence optimization for the complex AsnRS:AspAMP, using the weighted stability / affinity criterion. In thin lines is shown the conformation of maximum stability for the native AsnRS:AspAMP. The most important interactions are shown in dashed lines.

free protein, and computed the average (over rotamers) stabilities of the two complexes and the free protein. We chose sequences for which the relative affinity and the stability of the free protein were improved compared to the corresponding values for the native AsnRS sequence, 5.2 kcal/mol and -680.9 kcal/mol (values after rotamer optimization). We reconstructed the 100 lowest energy rotamer conformations of the complexes AsnRS:AspAMP and AsnRS:AsnAMP for these 37 sequences (passed through the second filter), and minimized them by 90 steps of a Powell conjucated gradient algorithm, keeping the backbone fixed. An additional 10 steps of minimization were performed, in the presence or absence of the ligand. The final minimized energies of the complexes and free protein were employed to compute the absolute binding free energy of AspAMP, relative to AsnAMP; for each sequence, the result was averaged over the 100 rotameric conformations. The results are shown in Table 9.9.

The relative binding free energy of the native sequence for AspAMP is 3.9 kcal/mol, compared to AsnAMP, signifying that AsnRS is optimized for asparagine recognition. The designed sequences have negative relative affinities, ranging from -3.8 to -13.8 kcal/mol. Furthermore, all sequences have a lower folding free energy than the native protein. The absolute binding affinities of these sequences for AspAMP at the end of minimization are shown in the second column of the table. With the exception of MSKMS and QSKSS, these affinities are somewhat better than the corresponding affinity of the native sequence. Thus, most of the obtained sequences are at least as stable as the native protein; at the same time they have at least as strong absolute affinity for AspAMP and a weaker binding for AsnAMP. The mutations S190 / K225 / S366 are conserved for all sequences (except for S190 / V,I225 / S366 and S190 / E225 / S366), while Glu227 is replaced by histidine, methionine or serine. Gln187 is either maintained or replaced by lysine, histidine, methionine, valine or icoleucine. Some of the sequences / conformations were obtained earlier, with absolute affinity and stability optimization criteria.

The sequences with optimum relative and absolute affinities KSVES, KSVSS have the same conformations with the optimum structure KSEES of the absolute affinity calculation (Fig. 9.14) with the exception of residue 225. Sequence HSVMS is shown in Fig. 9.17. The sidechains of His227, Arg368 and Tyr337 form a ladder of pi-stacking interactions. Met227, Val225 and Met223 make non-polar contacts. Sequence VSIMS was found in the stability optimization of the AsnRS:AsnAMP complex. Val187 makes a non-polar contact with Met227 and Ile225; Ile225 also contacts Met223. In sequence MSVSS, Met187 makes a non-polar cluster with Val225 and Met223. In all sequences, Arg368 interacts with the ligand sidechain. When position 187 is occupied by a non-polar residue (sequences VSKMS, VSIMS, MSVSS) Arg368 is slightly rotated toward the 187 sidechain. In the rest of the sequences (KSVSS, KSVES, HSVMS, and HSKMS) Arg368 is rotated slightly upwards and maintains a parallel orientation to the Tyr337 ring. The total charge of the active positions in the designed sequences ranges between

Table 9.9: Sequences of the complex AsnRS:AspAMP, selected from the relative-affinity design, following reconstruction, rotamer optimization and minimization (see text).

Sequence						
187	190	225	227	366	$\langle \Delta G \rangle$	$\langle \Delta\Delta G \rangle$
Native:						
Q	A	E	E	S	-61.4	3.9
Designed:						
K	S	V	E	S	-66.3	-13.8
K	S	V	H	S	-62.1	-10.9
H	S	V	M	S	-61.6	-9.7
K	S	V	S	S	-62.9	-9.6
K	S	I	E	S	-63.3	-9.5
M	S	V	S	S	-62.0	-9.0
H	S	K	M	S	-61.5	-8.2
I	S	V	M	S	-62.9	-8.0
V	S	I	M	S	-63.5	-7.9
Q	S	V	S	S	-61.6	-7.6
V	S	V	M	S	-63.1	-5.7
M	S	K	M	S	-59.0	-5.7
V	S	K	M	S	-62.4	-4.9
K	S	K	S	S	-60.3	-3.8
V	S	K	S	S	-61.5	-2.4
K	S	E	S	S	-61.9	-1.4

The designed sequences were filtered by a post-processing analysis, involving rotamer - optimization and energy minimization. ΔG is the binding free energy for AspAMP after minimization; $\Delta\Delta G$ is the relative affinity for AspAMP (compared to AsnAMP) after minimization.

0 and +2, compared to the negative value (-2) in the native sequence. The inversion of the charge sign is related to the relative affinity criterion.

9.2.8 Exploring an Alternate Active Position

The sequences identified by CPD obviously depend on the choice of mutatable ("active") positions. The five positions considered above (187, 190, 225, 227, and 366) were chosen for their proximity to the ligand sidechain and their homology to important recognition residues in AspRS (195, 198, 233, 235, and 487 in *E. coli* AspRS). The design calculations then predicted several mutations that increase the similarity between the AsnRS and AspRS active sites, especially the Q187K mutation. Three key interactions seen in AspRS, involving the AspAMP carboxylate, AspRS-Arg489, and AspRS-Lys198, are also present in the Q187K-AsnRS complexes. However, the AspRS active site contains a fourth key interaction, between the ligand carboxylate and His448 (see Fig. 9.4B). This interaction does not exist in the AsnRS:Asn com-

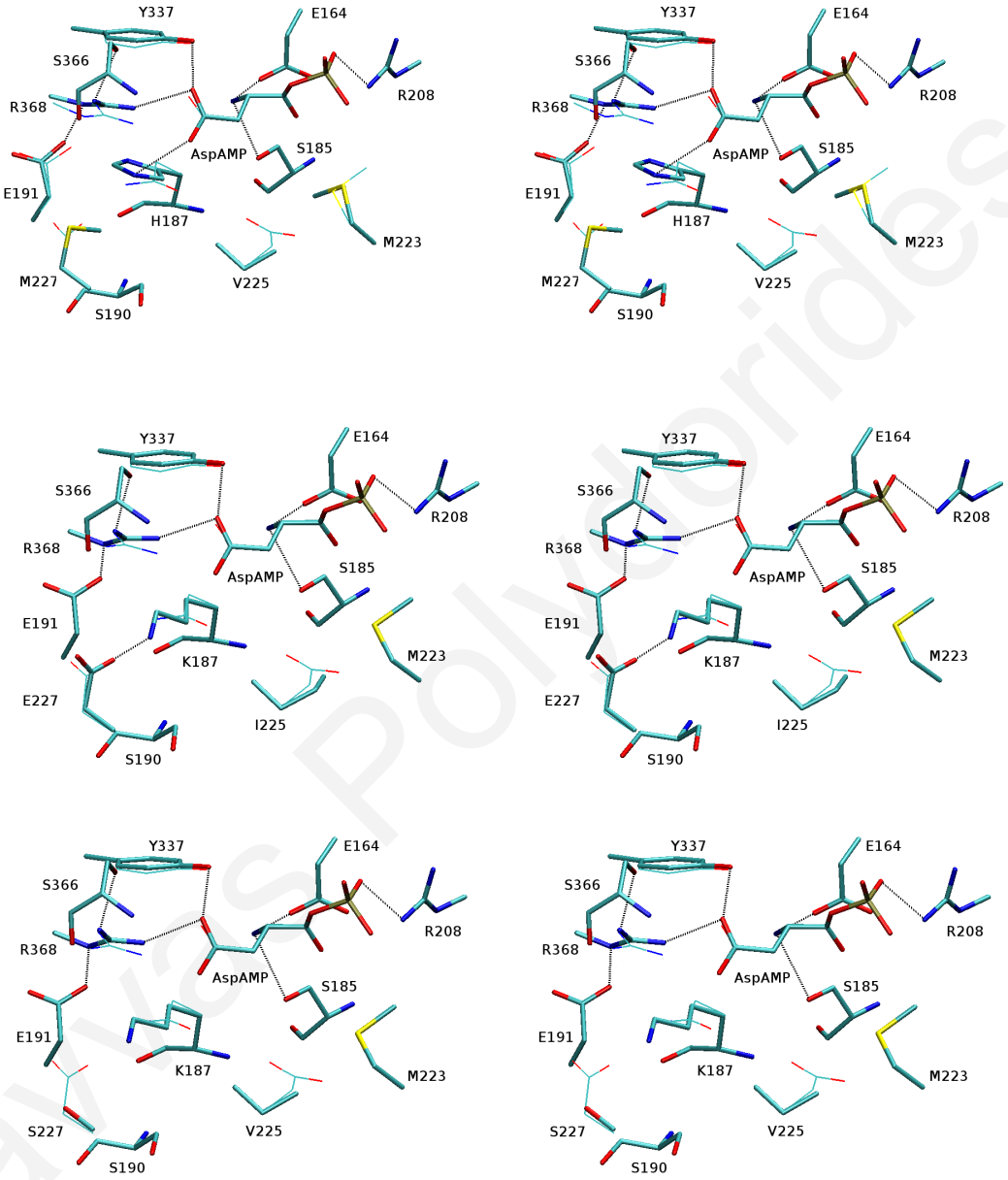


Figure 9.17: Stereo representation of the HSVMS (top) KSIES (middle) and KSVSS (bottom) conformations produced in sequence optimization for the complex AsnRS:AspAMP, using the combined relative affinity / stability criterion. In thin lines is shown the conformation of maximum stability for the native AsnRS:AspAMP. The most important interactions are shown in dashed lines.

plex, and is not inserted by our CPD calculations, since there is no active homologue of AspRS-His448. We therefore examined whether a sixth position could be identified, homologous to the AspRS-His448, which might allow additional mutations stabilizing AspAMP in AsnRS. Sequence alignment and visual inspection of the AspRS and AsnRS active sites show that the His448/His449 pair in *E. coli* AspRS does not have a close homologue in AsnRS. His448 is replaced by a nonpolar residue in AsnRS, and His449 is replaced by Lys334, but the Lys334 sidechain points directly away from the ligand. In fact, the sidechain that appears most likely to accept a His and form a strong interaction with the ligand is Tyr337. This residue is the nearest (after Arg368) to the ligand sidechain within the Proteus-optimized native complex; its C_β atom is about 8Å from the AspAMP carboxylate. Other positions are either too far away, poorly oriented to form an interaction with AspAMP, or both. A Tyr337His mutation was therefore introduced with a "minimum" protocol (see Methods). The His337 sidechain did indeed interact closely with the ligand carboxylate, just 2.6Å away, but it interfered sterically and electrostatically with the ligand ammonium group (which hydrogen-bonds to native Tyr337). As a result, the protein-ligand interaction energy actually increased (disfavoring binding) by 7 kcal/mol (charged His337) or 3 kcal/mol (neutral His337). Thus, there is no obvious, additional, active position, that is consistent with the AsnRS backbone fold and appears likely to allow a strong interaction, homologous to the His448-ligand interaction in AspRS. Allowing for structural rearrangements of the AsnRS backbone during the design might facilitate the insertion of such an interaction. This is beyond the scope of the present fixed-backbone design study, but is worth exploring more systematically in the future.

9.2.9 Molecular Dynamics Simulations

Selected mutant sequences were studied further by MD simulations and PB free-energy calculations. We chose the native *Thermus thermophilus* AsnRS protein (QAEES in Table 9.10), seven mutants determined by the absolute-affinity criterion (KSEES, KSTES, KSIES, HSEKS, HSTES, HSTMS, YSQKS), and three mutants determined by the relative-affinity criterion (KSVSS, KSVES, HSVMS). All eleven sequences were simulated in complex with the target ligand AspAMP. The native sequence and four mutants were also simulated in complex with the native ligand, AsnAMP.

The initial coordinates of main chain atoms were taken from the crystallographic structures of *T. thermofilus* AsnRS in complex with the asparaginyl adenylate [197]; the ligand was fitted into the position seen in the active site of *E. coli* Asprs, in complex with the aspartyl adenylate [190]. The protein side chains were initially placed in positions predicted by the design calculations, by superposing on the dimer the 20 Å sphere from the PROTEUS calculations. The structure was then truncated to a sphere of radius 24 Å centered on the ligand. To model the complex of a synthetase

Table 9.10: Binding affinities for AspAMP and AsnAMP of selected designed AsnRS sequences. The affinities were computed by PBFE calculations on equilibrium conformations, obtained by explicit-solvent MD runs. All values in kcal/mol.

Sequence					AspAMP binding		AsnAMP binding		Asp-Asn binding	
187	190	225	227	366	GB/HCT[a]	PBFE	GB/HCT[a]	PBFE	GB/HCT[a]	PBFE
Q	A	E	E	S	-111.2	-23.7	-116.4	-25.7	5.2	+2.0
K	S	T	E	S	-112.1	-22.7				
H	S	V	M	S	-111.0	-22.4	-109.5	-11.9	-1.5	-10.5
K	S	V	S	S	-109.0	-22.3	-108.8	-17.1	-0.2	-5.2
K	S	E	E	S	-111.7	-21.4				
K	S	V	E	S	-111.7	-19.9	-111.2	-16.6	-0.5	-3.3
H	S	E	K	S	-112.0	-19.3				
K	S	I	E	S	-112.0	-17.9				
Y	S	Q	K	S	-111.7	-15.9				
H	S	T	E	S	-112.3	-14.0				
H	S	T	M	S	-111.8	-13.8	-103.3		-9.2	-4.6

[a] Values at the end of Proteus optimization (before minimization), from Tables 9.8 - 9.9.

with the non-cognate ligand, we fit the asparaginyl or aspartyl molecule using their adenylate moiety. We removed any crystallographic waters far from the ligand (> 16 Å) and overlayed a large water box of edge 64 Å. Any overllaping waters with the protein or the ligand were also removed, by applying a minimum acceptable distance between water molecules and heavy atoms of 2.8 Å. The complete model contained the protein:ligand complex, about 8500 water molecules and a single Mg^{+2} ion bound to the ligand $\alpha-$phosphate. A few positive (sodium) or negative (chloride) ions were also included, in order to neutrilize the total charge of the system (their number and kind depended on the mutant sequence). Protein atoms of the outer shells of the sphere ($20 < r < 24$ Å) were harmonically restrained to their experimental positions.

The CHARMM22 forcefield was employed for the protein and the AMP moiety of the ligands [201]. The linkage between the AMP and aminoacid moieties of the ligands was parameterized as in Ref. [202]. The water was represented by a TIP3P model [203]. Long-range electrostatic interactions were computed without cutoff by the particle-mesh Ewald method [204], with a parameter k=0.34 for the charge screening and 6th-order splines for the mesh interpolations. The Lennard-Jones interactions between atom pairs were switched to zero at a cutoff distance of 12 Å. The temperature was kept at 300 K by a Nosé-Hoover thermostat [205; 206] using a mass of 1000 kcal/mol ps^2 for the thermostat. The pressure was maintained at 1 Atm with a Langevin piston [207], using a 500 amu mass and a 5 ps^{-1} collision frequency for the piston. The classical equations of motion were integrated by the Leap-Frog integrator, using a time step of 1 fs. Bond lengths to hydrogen atoms and the internal water geometry were constrained

to standard values with the SHAKE algorithm [208]. All simulations were performed with the CHARMM program, version c35b3 [209].

Initially, the system was subjected to 100 steps of steepest descent minimization followed by another 100 steps of the adopted basis Newton-Raphson method, to correct the structure from any bad contacts between the protein and the solvent. We then equilibrated the system by 60 ps of dynamics, during which all non-hydrogen protein atoms were harmonically restrained around the initial positions. The harmonic restraints were progressively varied from 5 to 1 kcal/mol/Å^2 (mainchain) and 0.5 kcal/mol/Å^2 (sidechain atoms). Subsequently, all protein atoms in the inner 17 Å-sphere were left free; the remaining protein atoms (excluding hydrogens) were harmonically restrained to their experimental positions; the employed force constants reproduced the corresponding crystallographic B factors. This production phase lasted 4 ns in most simulations.

The simulations were conducted in explicit water and lasted 4 ns. 1000 snapshots, extracted at 4-ps intervals, were used to compute the AspAMP or AsnAMP binding affinities by Poisson calculations. We did not include nonpolar contributions to the affinities; in our earlier work, we showed that these contributions were small, not very sensitive to the mutations, and within the PBFE uncertainty [77].

The conformations of the various complexes are quite stable in the MD simulations. The rmsd from the starting conformation is 0.5-0.7 Å for mainchain and 1.2 Å for sidechain atoms in the inner (unrestrained) sphere of the complexes. In the AspAMP complexes of the Lys187-containing sequences, the active sites maintain the Asp ligand recognition mode of AspRS, shown in Fig. 9.4B. This is illustrated in Figs. 9.18C-D, which display the final MD conformations of the two most promising sequences from the absolute-affinity (KSEES) and relative-affinity design (KSVSS). The ligand AspAMP carboxylate retains two hydrogen-bonding interactions with Arg388 and one hydrogen bond with Lys187. Furthermore, in KSEES, the orientation of Arg388 is further stabilized by two interactions with Glu227. In the His187 complexes, the ligand carboxylate retains two interactions with Arg388 during the simulations, but loses its interactions with the His187 sidechain (not shown).

9.2.10 Poisson-Boltzmann Free Energy Calculations

The electrostatic contribution to the ligand binding free energy was computed by subtracting the electrostatic free energy in the complex and the isolated ligand and protein. The electrostatic potential was computed by numerical solution of the Poisson-Boltzmann equation, using a finite-difference algorithm implemented in CHARMM [210]. The solvent-protein dielectric boundary was defined by a probe sphere with radius of 2.0 Å. The solvent dielectric constant was set to 80 and the protein/ligand dielectric constants were set to 4, as in [77]. The ionic strength was set to 100 mM.

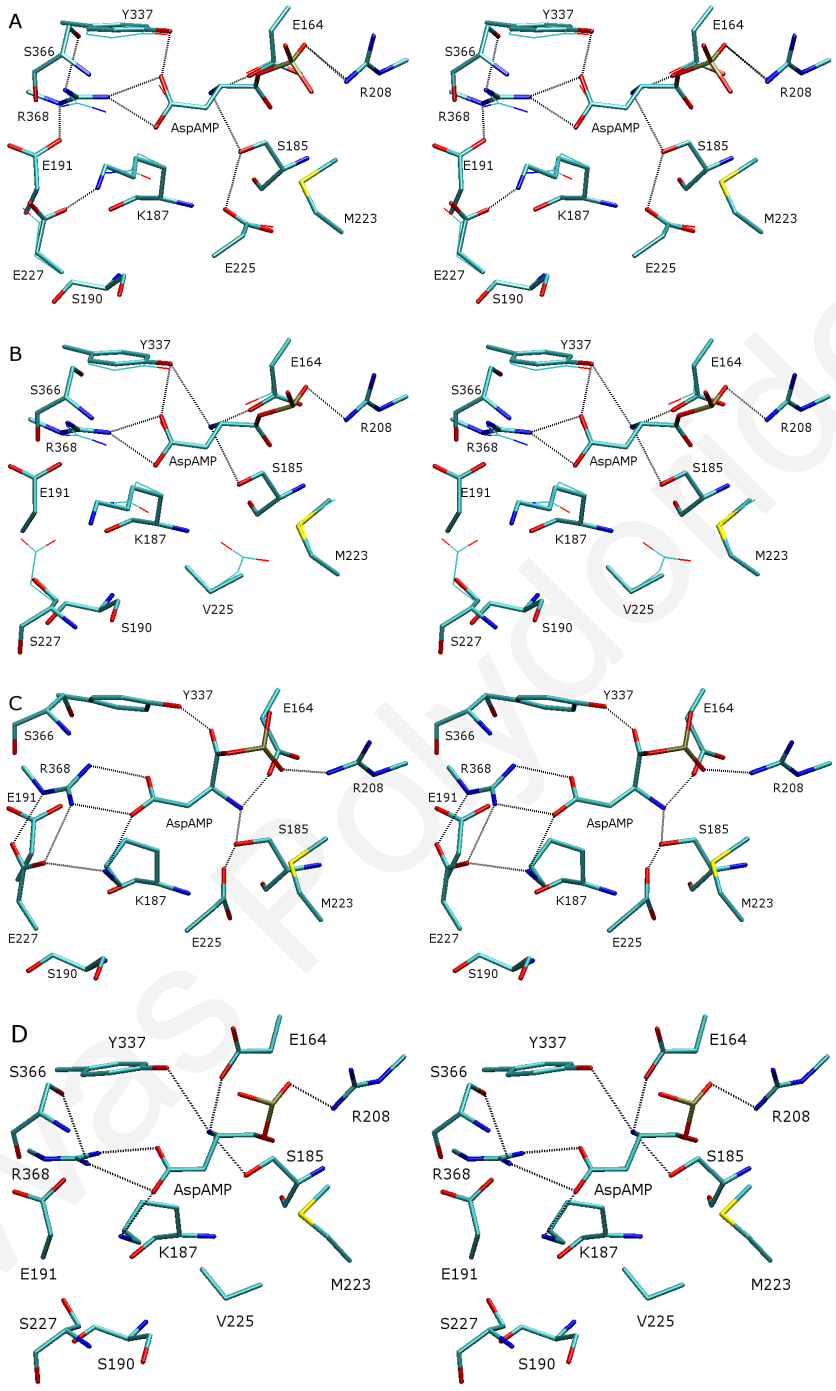


Figure 9.18: Active-site conformations of the most promising AsnRS:AspAMP complexes from the stability/absolute affinity design, KSEES (shown in **A**) and stability/relative affinity design, KSVSS (in **B**). In thin lines is shown the native-AsnRS:AspAMP conformation of maximum stability (Fig. 9.4). Figs **(C)** and **(D)** show, respectively, the conformations of complexes KSEES and KSVSS at the end of 4-ns MD simulations in explicit water.

PB calculations were performed for 1000 structures, sampled every 4 ps of the MD simulation.

The PBFE affinity results are summarized in Table 9.10. The calculations predict that native AsnRS has a PBFE binding affinity of -25.7 kcal/mol for the native ligand (AsnAMP) and an AspAMP affinity that is weaker by 2 kcal/mol. With a zero ionic strength, the corresponding numbers are -26.0 and -29.8 kcal/mol, resulting in a relative affinity of 3.8 kcal/mol in favor of AsnAMP. The four AsnRS mutants simulated with bound AsnAMP have inverted binding affinities, favoring AspAMP by -3.3 kcal/mol to -10.5 kcal/mol. Three of these sequences (KSVSS, KSVES, HSVMS) were indeed determined by the stability/relative affinity criterion. Thus, our design was successful in decreasing Asn binding compared to Asp binding. At the same time, according to PBFE, none of the sequences binds Asp more strongly than the native sequence. This is consistent with the marginally better affinity of the stability/absolute-affinity design sequences (Table 9.8). Furthermore, the AspAMP affinities are smaller, by several kcal/mol, than the native affinity for AsnAMP.

Even though the PBFE analysis employed here is less accurate with respect to alchemical free energy calculations [76; 100; 141; 143], in the past we have employed it succesfully to compute protein-ligand affinities in various systems, including the proteins AspRS and AsnRS [139; 141; 142]. The PBFE affinity estimates obtained here are consistent with the experimental activity measurements on selected designed sequences, as explained below.

9.3 Discussion of Results

Compared to a previous CPD study of the same system [77], our present work has two methodological innovations. In the previous study [77], protein interactions were computed by a polar-hydrogen energy function [62] and solvent effects by a Coulomb / Accessible Surface Area (CASA) approximation [76; 199]. Here, we use an all-atom energy function [138] for the protein and ligand interactions, and a continuum electrostatics generalized Born model based on the GB-HCT formalism [90] for solvent effects. We use a parameterized version of the GB-HCT model, shown to yield accurate protein solvation free energies and free-energy changes due to mutations in fully or partly buried positions [76]. We deal with the many-body nature of the GB model in two steps. (i) We use a "residue-GB" approximation [91; 181], in which all atoms of a residue are assigned a common, "residue Born" radius, defined as a harmonic average over the Born radii of the individual atoms within the residue; (ii) We compute and tabulate prior to the design calculations the residue Born radii, assuming that the environment of each residue corresponds to the chemical structure and geometry of the native state. An analogous approximation has been used in conjunction with continuum electrostatics treatments in protein design [70; 71; 117; 177].

Second, we compared the performance of three optimization criteria, which target, respectively, the folding free energy (stability), the affinity for the non-cognate ligand (AspAMP), or the affinity relative to the cognate ligand, AsnAMP. In contrast, our earlier study only used the stability criterion.

The all-atom / GB (this work) and polar-atom / CASA [77] free-energy functions yield different optimized AsnRS sequences. We first compare the predictions of the stability criterion, which is common to both studies. Sequence alignment of AspRS and AsnRS active sites [96] shows that the five AspRS active-site residues corresponding to the designed AsnRS positions Gln187, Ala190, Glu225, Glu227, Ser366 are, respectively, Gln195, Lys198, Asp233, Glu235 and Ser487 (in *E. coli* numbering). Thus, the chemical identity is preserved at positions 187 (Gln), 227 (Glu) and 366 (Ser); the negative charge is retained at position 225, and a positive charge is inserted at position 190.

The majority of sequences designed with the polar-hydrogen/CASA free energy function have negatively-charged residues (Asp, Glu) at positions 187 and 366 [77]; position 190 either retains its native identity (Ala), or was set manually to be a Lys (in analogy with Lys198 in AspRS), and positions 225, 227 mostly become hydrophobic (Ala, Met) or polar (Ser, Asn). The sequences designed with the GB-HCT/all-atom energy function and the stability criterion (Table 9.5) are more consistent with the properties of the AspRS active site. Position 366 remains invariant (Ser) and positions 225, 227 often retain a negatively-charged residue (Glu). Position 187 can remain invariant (Gln), but mostly accepts an aromatic (Tyr) or non-polar (Val) residue which disrupts the ligand carboxylate-Arg388 interactions.

With a combined stability / absolute-affinity criterion (Table 9.8), position 187 is changed to a charged (Lys) or aromatic (His, Tyr) sidechain. The Lys insertion at this position constitutes a major improvement, compared to the stability / CASA prediction (mostly Asp or Glu): The Lys187 ammonium group forms an interaction with the ligand sidechain carboxylate, analogous to the Lys198-ligand interaction in the AspRS:Asp active site. The resulting conformations are very stable during the MD simulations, retaining the AspRS-like geometry (Fig. 9.4B). In contrast, the CASA-derived sequences experienced significant distortions with respect to the native AsnAMP complex in the MD runs of Ref. [77], due to repulsions between the ligand sidechain carboxylate and the negatively charged residue inserted at position 187 (and sometimes 366). The CASA design did not insert a lysine residue at position 190 spontaneously; rather, this Lys was inserted by hand [77].

The good conformational stability and AspRS-like protein-ligand interactions of the simulated complexes indicate that the mutations introduced by the present model were physically reasonable. Stability optimization of the native complex AsnRS:Asn with the same five active positions produced various sequences. With the exception of position 190 (100% serine), the native aminoacids are observed with the highest frequency in all

other positions. Sequence alignment shows that the four active positions Q187, E225, E227 and S366 are conserved across AsnRS from various species; position 190 is more variable, containing methionine(45%), alanine(15%), glycine(12%), leucine(11%) and valine(6%). Thus, Ala190 is replaced by much bulkier residues (Leu, Val); presumably such an insertion is not favored by our design due to the fixed backbone. Our model predicts a serine at position 190 with 100% probability; at the same time, it predicts no serine at positions 187 and 225, suggesting that there is no consistent bias for this residue.

The Proteus design results with combined stability/affinity criteria (Tables 9.8-9.9) and the PBFE analysis of the trajectories (Table 9.10) suggested that the obtained sequences had inverted specificities, but their Asp binding was not as strong as Asn binding by native AsnRS. This is consistent with the inability of the experimentally tested sequences to adenylate Asp or (Asn) with ATP. A pure absolute-affinity or relative-affinity criterion yielded sequences with strong Asp affinity; these sequences had decreased stability by 100 kcal/mol or more with respect to native AsnRS and were not considered further.

9.4 Experimental Results

To further test the success of our design, our collaborators (C. Aubard and P. Plateau, Ecole Polytechnique) measured the activity against L-asparagine and L-aspartate of wild-type AsnRS and several mutant sequences. Specifically, they chose the sequences KSTES, KSEES, KSIES, HSEKS and HSTES (stability/absolute-affinity criterion) and KSVES, KSVSS (stability/relative-affinity criterion). Furthermore, they tested the most promising sequence, DKMMD, from our earlier CPD [77]. Sequence DKMMD was predicted earlier (with CASA) to have a native-like binding mode and an inverted affinity, favoring Asp by 11.8 kcal/mol.

The experiments failed to show a detectable Asp or Asn activity for any of the designed sequences. The initial rate of L-asparagine-dependent isotopic ATP-PP_i exchange was 0.83 sec^{-1} for wild type AsnRS, and less than 1×10^{-2} sec^{-1} for all the mutants. The corresponding initial rate of L-aspartate-dependent isotopic ATP-PP_i exchange was 6×10^{-3} sec^{-1} for wildtype AsnRS. This last value is likely to reflect contamination of the experimental aspartate sample by asparagine. The initial rates of exchange were less than 5×10^{-3} sec^{-1} for the mutants.

Several factors may have contributed to the failure to obtain sequences able to adenylate Asp. The computationally simpler approximation to the residue-GB/HCT model is less accurate than the original atomic- or residue-GB/HCT model; the heuristic algorithm searches a small subset of the conformational space: side-chains are placed into a small set of rotamer states [135] and the protein backbone is retained into the conformation of native AsnRS. Using a much more extended rotamer library [211],

and/or introducing backbone flexibility [8] could yield additional sequences, missed by the present design. Finally, the inactivity of the mutant proteins could be due to a disruption of transition state stabilization. It could also be that structural rearrangements due to the design interfere with ATP binding, necessary for the initial adenylation reaction to occur. Further testing is needed to determine the importance of these factors.

9.5 Conclusion

In conclusion, we have implemented and tested a CPD methodology in which solvent effects are modelled with the generalized-Born approximation and sequences are selected according to both stability and affinity. Using this methodology, we have engineered Asp specificity into AsnRS. The engineered AsnRS active sites have some of the structural features seen in the Asp-specific protein Aspartyl-tRNA synthetase; their conformations and interactions are maintained in MD simulations. Compared with the earlier CPD study [77], which treated solvent effects by a Coulomb / Accessible Surface Area approxination, the present, GB treatment appears promising.

Bibliography

[1] F. Boas and P. Harbury., "Potential energy functions for protein design," *Curr. Opin. Struct. Biol.*, vol. 17, pp. 199–204, 2007.

[2] C. Floudas, H. Fung, S. Mcallister, M. Monnigmann, and R. Rajgaria., "Advances in protein structure prediction and de novo protein design: A review," *Chem. Engin. Sci.*, vol. 61, pp. 966–988, 2006.

[3] S. Lippow and B. Tidor., "Progress in computational protein design," *Curr. Opin. Biotechn.*, vol. 18, pp. 305–311, 2007.

[4] R. Das and D. Baker., "Macromolecular modelling with rosetta," *Ann. Rev. Biochem.*, vol. 77, pp. 363–382, 2008.

[5] M. S. am Busch, D. Mignon, and T. Simonson., "Computational protein design as a tool for fold recognition," *Proteins*, vol. 77, pp. 139–158, 2010.

[6] J. Karanicolas and B. Kuhlman., "Computational design of affinity and specificity at protein-protein interfaces," *Curr. Opin. Struct. Biol.*, vol. 19, pp. 458–463, 2009.

[7] J. Dambrovsky and J. Brezovsky., "Computational tools for designing and engineering biocatalysts," *Curr. Opin. Chem. Biol.*, vol. 13, pp. 26–34, 2009.

[8] D. Mandell and T. Kortemme., "Backbone flexibility in computational protein design," *Curr. Opin. Biotechn.*, vol. 20, pp. 420–428, 2009.

[9] M. Suarez and A. Jaramillo., "Challenges in the computational design of proteins," *J. R. Soc. Interface*, vol. 6, pp. 5477–5491, 2009.

[10] E. Mashiach, R. Nussinov, and H. Wolfson, "Flexible induced-fit backbone refinement in molecular docking," *Proteins*, vol. 78, pp. 1503–1519, 2009.

[11] M. Schmidt am Busch, A. Sedano, and T. Simonson, "Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition," *PloS One*, vol. 5(5), p. e10410, 2010.

[12] J. Saven., "Computational protein design: Advances in the resign and redesign of biomolecular nanostructures," *Curr. Opin. Colloid Interf. Sci.*, vol. 15, pp. 13–17, 2010.

[13] M. Bellows and C. Floudas., "Computational methods for de novo protein design and its applications to the human immunodeficiency virus 1, purine nucleoside phosphorylase, ubiquitin specific protease 7 and histone demethylases," *Current drug targets*, vol. 11, pp. 264–278, 2010.

[14] B. Kuhlman, G. Dantas, G. Ireton, G. Varani, B. Stoddard, and D. Baker., "Design of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, pp. 1364–1368, 2003.

[15] A. Korkegian, M. Black, D. Baker, and B. Stoddard., "Computational thermostabilization of an enzyme," *Science*, vol. 308, pp. 857–860, 2005.

[16] X. Ambroggio and B. Kuhlman., "Computational design of a single aminoacid sequence that can switch between two distinct protein folds," *J. Am. Chem. Soc.*, vol. 128, pp. 1154–1161, 2006.

[17] P. Shah, G. Hom, S. Ross, J. Lassila, K. Crowhurst, and S. Mayo., "Full-sequence computational design and solution structure of a thermostable protein variant," *J. Mol. Biol.*, vol. 372, pp. 1–6, 2007.

[18] E. Bae, R. Bannen, and G. Phillips, "Bioinformatic method for protein thermal stabilization by structural entropy optimization," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 9594–9597, 2008.

[19] D. Bolon and S. Mayo, "Enzyme-like proteins by computational design," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 14274–14279, 2001.

[20] J. Calhoun, H. Kono, S. Lahr, W. Wang, W. deGrado, and J. Saven., "Computational design and characterization of a monomeric helical dinuclear metalloprotein," *J. Mol. Biol.*, vol. 334, pp. 1101–1115, 2003.

[21] L. Looger, M. Dwyer, J. Smith, and H. Hellinga., "Computational design of receptor and sensor proteins with novel functions," *Nature*, vol. 423, pp. 185–189, 2003.

[22] J. Kaplan and W. F. DeGrado, "De novo design of catalytic proteins," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 11566 – 11570, 2004.

[23] F. Cohran, S. Wu, W. Wang, V. Nanda, J. Saven, M. Therien, and W. DeGrado, "Computational de novo design and characterization of a four helix bundle protein that selectively binds a non-biological cofactor," *J. Am. Chem. Soc.*, vol. 127, pp. 1346–1347, 2005.

[24] J. Calhoun, W. Liu, K. Spiegel, M. D. Peraro, M. Klein, K. Valentine, J. Wand, and W. deGrado., "Solution nmr structure of a designed metalloprotein and complementary molecular dynamics refinement," *Structure*, vol. 16, pp. 210–215, 2008.

[25] M. Faiella, C. Andreozzi, R. deRosales, V. Pavone, O. Maglio, F. Nastri, W. deGrado, and A. Lombardi., "An artificial di-iron oxo-protein with phenol oxidase activity," *Nature Chem. Biol.*, vol. 5, pp. 882–884, 2009.

[26] D. Bolon, C. Voigt, and S. Mayo, "De novo design of biocatalysts," *Curr. Opin. Chem. Biol.*, vol. 6, pp. 125 – 129, 2002.

[27] U. Bornscheuer and R. Kazlauskas, "Catalytic promiscuity in biocatalysts: Using old enzymes to form new bonds and follow new pathways," *Angew. Chem. Int. Ed. Engl.*, vol. 43, pp. 6032–6040, 2004.

[28] B. Tynan-Conolly and J. Nielsen., "Redesigning protein pk_a values.," *Prot. Sci.*, vol. 16, pp. 239–249, 2007.

[29] D. Röthlisberger and O. Khersonsky and A.M. Wollacott and L. Ziang and J. Dechncie and J. Betker and J.L. Gallaher and E.A. Althoff and A. Zanghellini and O. Dym and S. Albeck and K.N. Houk and D.S. Tawfik and D. Baker., "Kemp elimination catalysts by computational enzyme design," *Nature*, vol. 453, pp. 190–195, 2008.

[30] L. J. et. al., "De novo computational design of retro-aldol enzymes," *Science*, vol. 319, pp. 1387 –, 2008.

[31] M. Dwyer, L. Looger, and H. Hellinga, "Computational design of a biologically active enzyme," *Science*, vol. 304, pp. 1967–1971, 2004.

[32] J. Lassila, J. Keeffe, P. Oelschlaeger, and S. Mayo, "Computationaly designed variants of escherichia coli chorismate mutase show altered catalytic activity," *Prot. Eng. Design and Selection*, vol. 18, pp. 161–163, 2005.

[33] G. Lazar, "Engineered antibody fc variants with enhanced effector function," *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 16710 – 16715, 2006.

[34] J. Shifman, M. Choi, S. Mihalas, S. Mayo, and M. Kennedy, "Ca^{2+}/ calmodulin-dependent protein kinase ii (camkii) is activated by calmodulin with two bound calciums.," *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 13968 – 13973, 2006.

[35] M. Altman, E. Nalivaika, M. Prabu-Jeyabalan, C. Schiffer, and B. Tidor., "Computational design and experimental study of tighter binding peptides to an inactivated mutant of hiv-1 protease," *Proteins*, vol. 70, pp. 678–694, 2008.

[36] K. Reynolds, M. Hanes, J. Thomson, A. Antczak, J. Berger, R. Bonomo, J. Kirsch, and T. Handel., "Computational redesign of the shv-1 beta-lactamase/beta-lactamase inhibitor protein interface," *J. Mol. Biol.*, vol. 382, pp. 1265–1275, 2008.

[37] J. Haidar, B. Pierce, Y. Yu, W. Tong, M. Li, and Z. Weng., "Structure-based design of a t-cell receptor leads to nearly 100-fold improvement in binding affinity for pepmhc," *Proteins*, vol. 74, pp. 948–960, 2009.

[38] M. Bellows, H. Fung, M. Taylor, C. Floudas, A. L. de Victoria, and D. Morikis., "New compstatin variants through two de novo protein design frameworks," *Biophys. J.*, vol. 98, pp. 2337–2346, 2010.

[39] J. Shifman and S. Mayo., "Exploring the origins of binding specificity through the computational redesign of calmodulin," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 13274 – 13279, 2003.

[40] D. Bolon, R. Grant, T. Baker, and R. Sauer., "Specificity versus stability in computational protein design," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 12724 – 12729, 2005.

[41] J. Ashworth, J. Havranek, C. Duarte, D. Sussman, R. M. Jr., B. Stoddard, and D. Baker., "Computational redesign of endonuclease dna binding and cleavage specificity," *Nature*, vol. 441, pp. 655–659, 2006.

[42] D. Green, A. Dennis, P. Fam, B. Tidor, and A. Jasanoff., "Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide," *Biochemistry*, vol. 45, pp. 12547–12559, 2006.

[43] E. Yosef, R. Politi, M. Choi, and J. Shifman, "Computational design of calmodulin mutants with up to 900-fold increase in binding specificity," *J. Mol. Biol.*, vol. 385, pp. 1470–1480, 2009.

[44] R. Chaloupkova, J. Sykorova, Z. Prokop, A. Jesenska, M. Monincova, M. Pavlova, M. Tsuda, Y. Nagata, and J. Dambrovsky., "Modification of activity and specificity of haloalkane dehalogenase from *sphingomonas paucimobilis* ut26 by engineering of its entrance channel," *J. Biol. Chem.*, vol. 278, pp. 52622–52628, 2003.

[45] M. Petrek, M. Otyepka, P. Banas, P. Kosinova, J. Koca, and J. Damborsky, "Caver: a new tool to explore routes from protein clefts, pockets and cavities," *BMC Bioinf.*, vol. 7, pp. 316 –, 2006.

[46] D. Guieysse, J. Cortes, S. Puench-Guenot, S. Barbe, V. Lafaquiere, P. Monsan, T. Simeon, I. Andre, and M. Remaud-Simeon, "A structure-controlled investigation of lipase enantioselectivity by a path-planning approach," *Chem. Bio. Chem.*, vol. 9, pp. 1308–1317, 2008.

[47] A. Slovic, H. Kono, J. Lear, J. Saven, and W. deGrado., "Computational design of water-soluble analogues of the potassium channel kcsa," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 1828–1833, 2004.

[48] L. Cristian, V. Nanda, J. Lear, and W. DeGrado, "Synergistic interactions between aqueous and membrane domains of a designed protein determine its fold and stability," *J. Mol. Biol.*, vol. 348, pp. 1225–1233, 2005.

[49] H. Yin, J. Slusky, B. Berger, R. Walters, G. Vilaire, R. Litvinov, J. Lear, G. Caputo, J. Bennett, and W. DeGrado, "Computational design of peptides that target transmembrane helices," *Science*, vol. 315, pp. 1817–1822, 2007.

[50] Y. Ofran and B. Rost., "Analyzing six types of protein-protein interfaces," *J. Mol. Biol.*, vol. 325, pp. 377–387, 2003.

[51] T. Kortemme and D. Baker., "Computational design of protein-protein interactions," *Curr. Opin. Chem. Biol.*, vol. 8, pp. 91–97, 2004.

[52] S. Lippow, K. Wittrup, and B. Tidor., "Computational design of antibody-affinity improvement beyond *in vivo* maturation," *Nat. Biotechnol.*, vol. 25, pp. 1171–1176, 2007.

[53] L. Joachimiak, T. Kortemme, B. Stoddard, and D. Baker., "Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface," *J. Mol. Biol.*, vol. 361, pp. 195–208, 2006.

[54] E. Humphris and T. Kortemme, "Design of multi-specificity in protein interfaces," *PLoS Comput. Biol.*, vol. 3, pp. 164 –, 2007.

[55] G. Grigoryan, A. Reinke, and A. Keating., "Design of protein interaction specificity gives selective bzip-binding peptides," *Nature*, vol. 458, pp. 859–864, 2009.

[56] B. Chevalier, T. Kortemme, M. Chadsey, D. Baker, R. Monnat, and B. Stoddard, "Design,activity and structure of a highly specific artificial endonuclease," *Mol. Cell*, vol. 10, pp. 895–905, 2002.

[57] J. Sander, P. Zaback, J. Joung, D. Voytas, and D. Dobbs, "Zinc finger targeter (zifit): And engineered zinc finger/target site design tool," *Nucleic Acids Res.*, vol. 35, pp. 599–605, 2007.

[58] M. Mena, T. Traynor, S. Mayo, and P. Daugherty, "Blue fluorescent proteins with enhanced brightness and photostability from a structurally targeted library," *Nat Biotechnol*, vol. 24, pp. 1569–1571, 2006.

[59] C. Otey, M. Landwehr, J. Endelman, K. Hiraga, J. Bloom, and F. Arnold, "Structure-guided recombination creates an artificial family of cytochromes p450," *PLos*, vol. 4, pp. 789–798, 2006.

[60] W. Aehle and D. Estell, "Systematic evaluation of sequence and activity relationships using site evaluation libraries for engineering multiple properties," *World Patent*, 2008.

[61] D. Bashford and D. Case., "Generalized born models of macromolecular solvation effects," *Ann. Rev. Phys. Chem.*, vol. 51, pp. 129–152, 2000.

[62] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus., "Charmm: A program for macromolecular energy, minimization and dynamics calculations," *J. Comput. Chem.*, vol. 4, pp. 187–217, 1983.

[63] P. Weiner and P. Kollman., "Amber-assisted model building with energy refinement-general program for modeling molecules and their interactions," *J. Comput. Chem.*, vol. 2, pp. 287–303, 1981.

[64] M. Schaefer, H. v. Vlijmen, and M. Karplus., "Electrostatic contributions to molecular free energies in solution," *Adv. Prot. Chem.*, vol. 51, pp. 1–57, 1998.

[65] A. Warshel and W. Parson., "Dynamics of biochemical and biophysical reactions: Insight from computer simulations," *Q. Rev. Biophys.*, vol. 34, pp. 563–679, 2001.

[66] T. Simonson., "Electrostatics and dynamics of proteins," *Rep. Prog. Phys.*, vol. 66, pp. 737–787, 2003.

[67] P. Koehl., "Electrostatic calculations: Latest methodological advances," *Curr. Opin. Struct. Biol.*, vol. 16, pp. 142–151, 2006.

[68] B. Roux and T. Simonson, "Implicit solvent models," *Biophys. Chem.*, vol. 78, pp. 1–20, 1999.

[69] T. Simonson., "Macromolecular electrostatics: Continuum models and their growing pains," *Curr. Opin. Struct. Biol.*, vol. 11, pp. 243–252, 2001.

[70] S. Marshall, C. Vizcarra, and S. Mayo., "One- and two-body decomposable poisson-boltzmann methods for protein design calculations," *Prot. Sci.*, vol. 14, pp. 1293–1304, 2005.

[71] N. Pokala and T. Handel., "Energy functions for protein design i: Efficient and accurate continuum electrostatics and solvation," *Prot. Sci.*, vol. 13, pp. 925–936, 2004.

[72] P. Koehl and M. Levitt., "De novo protein design i. in search of stability and specificity.," *J. Mol. Biol.*, vol. 293, pp. 1161–, 1999.

[73] L. Wernisch, S. Hery, and S. Wodak., "Automatic protein design with all-atom force-fields by exact and heuristic optimization," *J. Mol. Biol.*, vol. 301, pp. 713–736, 2001.

[74] N. Baker., "Poisson-boltzmann methods for biomolecular electrostatics," *Methods in Enzymology*, vol. 383, pp. 94–118, 2004.

[75] B. Lee and F. Richards., "The interpretation of protein structures: Estimation of static accessibility," *J. Mol. Bio.*, vol. 55, pp. 379–400, 1971.

[76] A. Lopes, A. Alexandrov, C. Bathelt, G. Archontis, and T. Simonson., "Computational sidechain placement and protein mutagenesis with implicit solvent models," *Proteins*, vol. 67, pp. 853–867, 2007.

[77] A. Lopes, M. S. am Busch, and T. Simonson., "Computational design of protein:ligand binding: Modifying the specificity of asparaginyl-trna synthetase," *J. Comp. Chem.*, vol. 31, pp. 1550–1560, 2010.

[78] B. Dahiyat and S. Mayo., "Protein design automation," *Prot. Sci.*, vol. 5, pp. 895–903, 1996.

[79] P. Ferrara, J. Apostolakis, and A. Caflisch., "Evaluation of a fast implicit solvent model for molecular dynamics simulations," *Proteins*, vol. 46, pp. 24–33, 2002.

[80] K. Ogata, A. Jaramillo, W. Cohen, J. B. F. Conana, and S. Wodak., "Automatic sequence design of mhc class-i binding peptides impairing cd8 + t-cell recognition.," *J. Biol. Chem.*, vol. 278, pp. 1281–1290, 2003.

[81] A. Onufriev, D. Bashford, and D. Case., "Modification of the generalized born model suitable for macromolecules," *J. Phys. Chem. B*, vol. 104, pp. 3712–3720, 2000.

[82] N. Calimet, M. Schaefer, and T. Simonson., "Protein molecular dynamics with the Generalized Born/ACE solvent model," *Proteins*, vol. 45, pp. 144–158, 2001.

[83] C. Simmerling, B. Strockbine, and A. Roitberg, "All-atom structure prediction and folding simulations of a stable protein," *J. Am. Chem. Soc.*, vol. 124, pp. 11258–11259, 2002.

[84] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, and A. Caflisch., "Exhaustive docking of molecular fragments with electrostatic solvation," *Proteins*, vol. 38, pp. 88–105, 1999.

[85] H. Liu and X. Zou., "Electrostatics of ligand binding: Parameterization of the generalized born model and comparison with the poisson-boltzmann approach," *J. Phys. Chem. B*, vol. 110, pp. 9304–9313, 2006.

[86] C. Simmerling, B. Strockbine, and A. Roitberg., "All-atom structure prediction and folding simulations of a stable protein," *J. Am. Chem. Soc.*, vol. 124, pp. 11258–11259, 2002.

[87] J. Chen, W. Im, and C. B. III., "Balancing solvation and intramolecular interactions: Towards a consistent generalized born force field," *J. Am. Chem. Soc.*, vol. 128, pp. 3728–3736, 2006.

[88] S. Jang, E. Kim, and Y. Pak., "Direct folding simulation of alpha-helices and beta-hairpins based on a single all-atom force field wth an implicit solvation model," *Proteins*, vol. 66, pp. 53–60, 2007.

[89] H. Lei, C. Wu, H. Liu, and Y. Duan., "Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations," *Proc. Natl. Acad. Sci. USA*, vol. 104, pp. 4925–4930, 2007.

[90] G. Hawkins, C. Cramer, and D. Truhlar., "Pairwise descreening of solute charges from a dielectric medium," *Chem. Phys. Lett.*, vol. 246, pp. 122–129, 1995.

[91] G. Archontis and T. Simonson., "A residue-pairwise generalized born scheme suitable for protein design calculations," *J. Phys. Chem. B*, vol. 109, pp. 22667–22673, 2005.

[92] D. Case, T. Darden, T. C. III, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M. Hsieh, G. Cui, D. Roe, D. Mathews, M. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. Kollman., *AMBER11*. 2010.

[93] J. Xie and P. Schultz., "Adding amino acids to the genetic repertoire," *Curr. Opin. Chem. Biol.*, vol. 9, pp. 548–554, 2005.

[94] D. Eisenberg and A. McClachlan., "Solvation energy in protein folding and binding," *Nature*, vol. 319, pp. 199–203, 1986.

[95] T. Simonson., "Free energy calculations," *In Computational Biochemistry and Biophysics.*, pp. 169–197, 2001. Edited by Becker OM, MacKerell AD, Roux B and Watanabe M, Marcel Decker Inc.

[96] G. Archontis, T. Simonson, D. Moras, and M. Karplus., "Specific Amino Acid Recognition by Aspartyl-tRNA Synthetase studied by Free Energy Simulations," *J. Mol. Biol.*, vol. 275, pp. 823–846, 1998.

[97] R. Rizzo, D. Wang, J. Tirado-Rives, and W. Jorgensen, "Validation of a model for the complex of hiv-1 reverse transcriptase with sustiva through computation of resistance profiles," *J. Am. Chem. Soc.*, vol. 122, pp. 12898 – 12900, 2000.

[98] C. Guimaraes, D. Boger, and W. Jorgensen, "Elucidation of fatty acid amide hydrolase inhibition by potent α-ketoheterocycle derivatives from monte carlo simulations," *J. Am. Chem. Soc.*, vol. 127, pp. 17377 – 17384, 2005.

[99] C. Chipota, X. Rozanskaa, and S. Dixit, "Can free energy calculations be fast and accurate at the same time? binding of low-affinity, non-peptide inhibitors to the sh2 domain of the src protein," *Journal of Computer-Aided Molecular Design*, vol. 19, pp. 765 – 770, 2005.

[100] M. Gilson and H. Zhou., "Calculation of protein-ligand binding affinities," *Ann. Rev. Biophys. Biomol. Struct.*, vol. 36, pp. 21–42, 2007.

[101] O. Guvench and A. D. M. Jr., "Computational evaluation of protein-small molecule binding," *Curr. Opin. Struct. Biol.*, vol. 19, pp. 56–61, 2009.

[102] G. Torrie and J. Valleau., "Nonphysical sampling distributions in monte carlo free-energy estimation:umbrella sampling," *Comput. Phys.*, vol. 23, pp. 187–199, 1977.

[103] S. Kumar, J. Rosenberg, D. Bouzida, R. Swendsen, and P. Kollman., "The weighted histogram analysis method for free-energy calculations on biomolecules i. the method.," *Comput. Chem.*, vol. 13, pp. 1011–1021, 1992.

[104] M. Lee and M. Olson, "Calculation of absolute protein-ligand binding affinity using path and endpoint approaches," *Biophys. J.*, vol. 90, pp. 864 – 877, 2006.

[105] H. Woo and B. Roux., "Calculation of absolute protein-ligand binding free energy from computer simulations," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 6825–6830, 2005.

[106] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus, "Absolute binding free energies: A quantitative approach for their calculation," *J. Phys. Chem. B*, vol. 107, pp. 9535 – 9551, 2003.

[107] Y. Lu, C. Yang, and S. Wang., "Binding free energy contributions of interfacial waters in hiv-1 protease/inhibitor complexes," *J. Am. Chem. Soc.*, vol. 128, pp. 11830–11839, 2006.

[108] J. Qvist, M. Davidovic, D. Hamelberg, and B. Halle, "A dry ligand-binding cavity in a solvated protein," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 6296–6301, 2008.

[109] M. Lee and M. Olson., "Calculation of absolute ligand binding affinity using path and endpoint approaches," *Biophys. J.*, vol. 90, pp. 864–877, 2006.

[110] M. Lee and M. Olson., "Assessment of detection and refinement strategies for de novo protein structures using force field and statistical potentials," *J. Chem. Theory Comput.*, vol. 3, pp. 312–324, 2007.

[111] M. Lee and M. Olson., "Calculation of absolute ligand binding free energy to a ribosome-targeting protein as a function of solvent model," *J. Phys. Chem. B*, vol. 112, pp. 13411–13417, 2008.

[112] S. Juffer, "Calculation of affinities of peptides for proteins," *J. Comp. Chem.*, vol. 25, pp. 393 – 412, 2004.

[113] I. Massova and P. A. Kollman, "Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding ," *Perspect. Drug Discov. Des.*, vol. 18, pp. 113–135, 2000.

[114] M. Lee, F. R. Salsbury, and C. L. Brooks III., "Constant pH molecular dynamics using continuous titration coordinates," *Proteins*, vol. 56, pp. 738–752, 2004.

[115] H. Gohlke, C. Kiel, and D. Case., "Insight into protein-protein binding by binding free energy calculation and free energy decomposition for the ras-raf and ras-ralgds complexes," *J. Mol. Biol.*, vol. 330, pp. 891–913, 2003.

[116] V. Zoete, M. Irving, and O. Michielin, "Mm-gbsa binding free energy decomposition and t cell receptor engineering," *J. Mol. Recognit.*, vol. 23, pp. 142–152, 2010.

[117] C. Vizcarra and S. Mayo., "Electrostatics in computational protein design," *Curr. Opin. Chem. Biol.*, vol. 9, pp. 622–626, 2005.

[118] J. Chen, C. B. III, and J. Khandogin., "Recent advances in implicit-solvent based methods for biomolecular simulations," *Curr. Opin. Struct. Biol.*, vol. 18, pp. 140–148, 2008.

[119] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson., "Use of mm-pbsa in reproducing the binding free energies to hiv-1 rt of tibo derivatives and predicting

the binding mode to hiv-1 rt of efavirenz by docking and mm-pbsa," *J. Am. Chem. Soc.*, vol. 123, pp. 5221–5230, 2001.

[120] P. Lyne, M. Lamp, and J. Saeh., "Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and mm-pbsa scoring," *J. Med. Chem.*, vol. 49, pp. 4805–4808, 2006.

[121] S. Polydoridis, D. Leonidas, N. Oikonomakos, and G. Archontis, "Recognition of ribonuclease a by 3'-5'-pyrophosphate-linked dinucleotide inhibitors: A molecular dynamics/continuum electrostatics analysis," *Biophys. J.*, vol. 92, pp. 1659 – 1672, 2007.

[122] L. Espinoza-Fonseca, D. Kast, and D. Thomas, "Thermodynamic and structural basis of phosphorylation-induced disorder-to-order transition in the regulatory light chain of smooth muscle myosin," *J. Am. Chem. Soc.*, vol. 130, pp. 12208–12209, 2008.

[123] F. Fogolari and S. Tosatto, "Application of mm/pbsa colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation," *Proteins Science*, vol. 14, pp. 889 – 901, 2005.

[124] S. Huo, I. Massova, and P. Kollman, "Computational alanine scanning of the 1:1 human growth hormonereceptor complex," *J. Comp. Chem.*, vol. 23, pp. 15–27, 2002.

[125] J. Carlsson, M. Ander, M. Nervall, and J. Aqvist., "Continuum models in the linear interaction energy method," *J. Phys. Chem. B*, vol. 110, pp. 12034–12041, 2006.

[126] T. Hansson, J. Marelius, and J. Aqvist, "Ligand binding affinity prediction by linear interaction energy methods," *J. Comp. Chem.*, vol. 12, pp. 27–35, 1998.

[127] R. Zhou, R. Friesner, A. Ghosh, R. Rizzo, W. Jorgensen, and R. Levy, "New linear interaction method for binding affinity calculations using a continuum solvent model," *J. Phys. Chem. B*, vol. 105, pp. 10388–10397, 2001.

[128] B. Brandsdal, F. Osterberg, M. Almlof, and I. Feierberg, "Free energy calculations and ligand binding," *Advances in Protein Chemistry*, vol. 66, pp. 123–158, 2003.

[129] S. Malakauskas and S. Mayo., "Design, structure and stability of a hyperthermophilic protein variant," *Nat. Struct. Biol.*, vol. 5, pp. 470–475, 1998.

[130] A. Filikov, R. Hayes, P. Luo, D. Stark, C. Chan, A. Kundu, and B. Dahiyat., "Computational stabilization of human growth hormone," *Protein Science*, vol. 11, pp. 1452–1461, 2002.

[131] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker., "A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins," *J. Mol. Biol.*, vol. 332, pp. 449–460, 2003.

[132] B. Mooers, D. Datta, W. Baase, E. Zollars, S. Mayo, and B. Matthews., "Repacking the core of t4 lysozyme by automated design," *J. Mol. Biol.*, vol. 332, pp. 741–756, 2003.

[133] T. Hendrickson, V. de Crecy-Lagard, and P. Schimmel., "Incorporation of nonnatural amino acids into proteins," *Ann. Rev. Biochem*, vol. 73, pp. 147–176, 2004.

[134] Y. Lu., "Design and engineering of metalloproteins containing unnatural amino acids or non-native metal-containing cofactors," *Curr. Opin. Chem. Biol.*, vol. 9, pp. 118–126, 2005.

[135] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery., "A new approach to the rapid determination of protein side chain conformations," *J. Biomol. Struct. Dyn.*, vol. 8, pp. 1267–, 1991.

[136] C. Pecore, J. Lecomte, and J. Desjarlais, "A de novo redesign of the ww domain," *Proteins*, vol. 12, pp. 2194–2205, 2003.

[137] C. Smith and T. Kortemme, "Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction," *J. Mol. Biol.*, vol. 380, pp. 752–756, 2008.

[138] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman., "A second generation force field for the simulation of proteins, nucleic acids and organic molecules," *J. Am. Chem. Soc.*, vol. 117, pp. 5179–5197, 1995.

[139] D. Thompson, P. Plateau, and T. Simonson., "Free energy simulations reveal long-range electrostatic interactions and substrate-assisted specificity in an aminoacyl-trna synthetase," *ChemBioChem*, vol. 7, pp. 337–344, 2006.

[140] D. Thompson and T. Simonson., "Molecular dynamics simulations show that bound mg^{2+} contributes to amino acid aminoacyl adenylate binding specificity in aspartyl-trna synthetase through long range electrostatic interactions.," *J. Biol. Chem.*, vol. 281, pp. 23792–23803, 2006.

[141] A. Aleksandrov, D. Thompson, and T. Simonson., "Alchemical free energy simulations for biological complexes: Powerful but temperamental," *J. Mol. Rec.*, vol. 23, pp. 117–127, 2010.

[142] G. Archontis, T. Simonson, and M. Karplus., "Binding free energies and free energy components from molecular dynamics and poisson-boltzmann calculations," *J. Mol. Biol.*, vol. 306, pp. 307–327, 2001.

[143] T. Simonson, G. Archontis, and M. Karplus, "Free-energy simulations come of age: Protein-ligand recognition," *Acc. Chem. Res.*, vol. 35, pp. 430–437, 2002.

[144] L. Jiang, B. Kuhlman, T. Kortemme, and D. Baker., "A 'solvated rotamer' approach to modeling water-mediated hydrogen bonds at protein-protein interfaces," *Proteins*, vol. 58, pp. 893–904, 2005.

[145] J. Schymkowitz, F. Rousseau, I. Martins, J. Borg, F. Stricherand, and L. Serran, "Prediction of water and metal binding sites and their affinities by using the fold-x force field," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 10147–10152, 2005.

[146] T. Ooi, M. Oobatake, G. Nemethy, and H. Scheraga., "Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides," *Proc. Natl. Acad. Sci. USA*, vol. 84, pp. 3086–3090, 1987.

[147] L. Wesson and D. Eisenberg., "Atomic solvation parameters applied to molecular dynamics of proteins in solution," *Prot. Sci.*, vol. 1, pp. 227–235, 1992.

[148] L. David, R. Luo, and M. Gilson., "Comparison of generalized born and poisson models: Energetics and dynamics of hiv protease," *J. Comp. Chem.*, vol. 21, pp. 295–309, 2000.

[149] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson., "Semianalytical treatment of solvation for molecular mechanics and dynamics," *J. Am. Chem. Soc.*, vol. 112, pp. 6127–6129, 1990.

[150] J. Jackson, *Classical Electrodynamics*. USA 3rd Ed: John Wiley and Sons Inc., 1999.

[151] D. Qiu, P. Shenkin, F. Hollinger, and W. Still., "A fast analytical method for the calculation of approximate born radii," *J. Phys. Chem. A*, vol. 101, pp. 3005–3014, 1997.

[152] A. Ghosh, C. Rapp, and R. Friesner., "Generalized born model based on a surface-area formulation," *J. Phys. Chem. B*, vol. 102, pp. 10983–10990, 1998.

[153] A. Romanov, S. Jabin, Y. Martynov, A. Sulimov, F. Grigoriev, and V. Sulimov., "Surface generalized born method: A simple, fast and precise implicit solvent model beyond the coulomb approximation," *J. Phys. Chem. A*, vol. 108, pp. 9323–9327, 2004.

[154] M. Schaefer and M. Karplus., "A comprehensive analytical treatment of continuum electrostatics," *J. Phys. Chem.*, vol. 100, pp. 1578–1599, 1996.

[155] I. Muegge, "Pmf scoring revisited," *J. Med. Chem.*, vol. 49, pp. 5895–5902, 2005.

[156] A. Ishchenko and E. Shakhnovich, "Small molecule growth 2001 (smog2001): An improved knowledge-based scoring function for protein-ligand interactions," *J. Med. Chem.*, vol. 45, pp. 2770–2780, 2002.

[157] H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions," *J. Mol. Biol.*, vol. 295, pp. 337 – 356, 2000.

[158] M. Vieth, A. Kolinsk, C. L. B. III, and J. Skolnick., "Prediction of quaternary structure of coiled coils application to mutants of the gcn4 leucine zipper.," *J. Mol. Biol.*, vol. 251, pp. 448–467, 1995.

[159] H. Bohm, "Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3d database search programs," *Journal of Computer-Aided Molecular Design*, vol. 12, pp. 309–323, 1998.

[160] A. G. Street and S. L. Mayo., "Pairwise calculation of protein solvent-accessible surface areas," *Folding and Design*, vol. 3, pp. 253–258, 1998.

[161] R. Goldstein., "Efficient rotamer elimination applied to protein side-chains and related spin glasses," *Biophys. J.*, vol. 66, pp. 1335–1340, 1994.

[162] P. Koehl and M. Delarue., "Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy," *J. Mol. Biol.*, vol. 239, pp. 249–275, 1994.

[163] D. Gordon, S. Marshall, and S. Mayo., "Energy functions for protein design," *Curr. Opin. Struct. Biol.*, vol. 9, pp. 509–513, 1999.

[164] L. L. Looger and H. W. Hellinga, "Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics," *J. Mol. Biol.*, vol. 307, pp. 429–445, 2001.

[165] D. T. Jones., "De novo protein design using pairwise potentials and a genetic algorithm," *Prot. Sci.*, vol. 3, pp. 567–574, 1994.

[166] J. Baker, "Reducing bias and inefficiency in the selection algorithm," pp. 14–21, 1987.

[167] D. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," pp. 69–93, 1991.

[168] G. Dantas, C. Corrent, S. Reichow, J. Havranek, Z. Eletr, N. Isern, B. Kuhlman, G. Varani, E. Merritt, and D. Baker., "High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design," *J. Mol. Biol.*, vol. 366, pp. 1209–1221, 2007.

[169] D. Sammond, Z. Eletr, C. Purbeck, R. Kimple, D. Siderovski, and B. Kuhlman., "Structure-based protocol for identifying mutations that enhance protein-protein binding affinities," *J. Mol. Biol.*, vol. 371, pp. 1392–1404, 2007.

[170] M. Roca, A. Vardi-Kilshtain, and A. Warshel., "Toward accurate screening in computer-aided enzyme design," *Biochemistry*, vol. 48, pp. 3046–3056, 2009.

[171] A. Warshel, "Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites," *J. Biol. Chem.*, vol. 273, pp. 27035–27038, 1998.

[172] R. Tan, T. Truong, J. McCammon, and J. Sussman, "Acetylcholinesterase: Electrostatic steering increases the rate of ligand binding," *Biochemistry*, vol. 32, pp. 401–403, 1993.

[173] R. Levy, L. Zhang, E. Gallicchio, and A. Felts, "On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy," *J. Am. Chem. Soc.*, vol. 125, pp. 9523–9530, 2003.

[174] M. S. am Buch, A. Lopes, N. Amara, C. Bathelt, and T. Simonson., "Testing the coulomb/accessible surface area solvent model for protein stability, ligand binding and protein design," *BMC Bioinformatics*, vol. 9, pp. 148–163, 2008.

[175] M. Wisz and H. Hellinga., "An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants," *Proteins*, vol. 51, pp. 360–377, 2003.

[176] T. Lazaridis and M. Karplus., "Effective energy functions for proteins in solution," *Proteins*, vol. 35, pp. 133–152, 1999.

[177] C. Vizcarra, N. Zhang, S. Marshall, N. Wingreen, C. Zeng, and S. Mayo., "An improved pairwise decomposable finite-difference poisson-boltzmann method for computational protein design," *J. Comp. Chem.*, vol. 29, pp. 1153–1162, 2008.

[178] J. Havranek and P. Harbury., "Tanford-kirkwood electrostatics for protein modeling," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 11145 – 11150, 1999.

[179] M. Gilson and B. Honig., "Total electrostatic energy of a protein," *Proteins*, vol. 4, 1988.

[180] M. Gerstein, J. Tsai, and M. Levitt, "The volume of atoms on the protein surface: Calculated from simulation, using voronoi polyhedra," *J. Mol. Bio.*, vol. 249, pp. 955–966, 1995.

[181] A. Aleksandrov, S. Polydorides, G. Archontis, and T. Simonson., "Predicting the acid/base behavior of proteins: A constant-ph monte carlo approach with generalized born solvent," *J. Phys. Chem. B*, vol. 114, pp. 10634–10648, 2010.

[182] J. Madura, J. Briggs, R. Wade, M. Davis, B. Luty, A. Ilin, J. Antosiewicz, M. Gilson, B. Baheri, L. Scott, and J. McCammon, "Electrostatics and diffusion of molecules in solution: Simulations with the university of houston brownian dynamics program," *Comp. Phys. Comm.*, vol. 91, pp. 57–95, 1995.

[183] A. Brunger, *X-PLOR Version 3.1, A system for X-ray crystallography and NMR.* New Haven: Yale University Press, 1992.

[184] M. S. am Busch, A. Lopes, D. Mignon, and T. Simonson., "Computational protein design: Software implementation, parameter optimization and performance of a simple model," *J. Comp. Chem.*, vol. 29, pp. 1092–1102, 2008.

[185] C. Voigt, D. Gordon, and S. Mayo., "Trading accuracy fo speed: A quantitative comparison of search algorithms in protein sequence design," *J. Mol. Biol.*, vol. 299, pp. 789–803, 2000.

[186] M. Delarue, "An asymmetric underlying rule in the assignment of codons: Possible clue to a quick early evolution of the genetic code via successive binary choices," *RNA*, vol. 13, pp. 161–169, 2007.

[187] G. Eriani, J. Cavarelli, F. Martin, L. Ador, B. Rees, J. Thierry, J. Gangloff, and D. Moras., "The class ii aminoacyl-trna synthetases and their active site: Evolutionary conservation of an atp binding site," *J. Mol. Evol.*, vol. 40, pp. 499–508, 1995.

[188] M. Delarue and D. Moras, "The aminoacyl-trna synthetase family: Modules at work," *BioEssays*, vol. 15, pp. 675–687, 1993.

[189] S. Eiler, A. Dock-Bregeon, L. Moulinier, J. Thierry, and D. Moras., "Synthesis of aspartyl-trna(asp) in *escherichia coli*: A snapshot of the second step," *EMBO J.*, vol. 18, pp. 6532–6541, 1999.

[190] L. Moulinier, S. Eiler, G. Eriani, J. Gangloff, J. Thierry, K. Gabriel, W. McClain, and D. Moras., "The structure of an asprs-trna ^{Asp} complex reveals a trna-dependent control mechanism.," *EMBO J.*, vol. 20, pp. 5290–5301, 2001.

[191] S. Cusack., "Eleven down and nine to go," *Nat. Struct. Biol.*, vol. 2, pp. 824–831, 1995.

[192] J. Arnez and D. Moras., "Structural and functional consideration of the aminoacylation reaction," *Trends Biochem. Sci.*, vol. 22, pp. 211–216, 1997.

[193] T. Meinnel, Y. Mechulam, and S. Blanquet, *In TRNA: Structure, Biosynthesis and Function.* Washington, D.C.: ASM Press, 1995.

[194] G. de Prat Gay, H. Duckworth, and A. Fersht., "Modification of the amino acid specificity of tyr-trna synthetase by protein engineering," *FEBS Lett.*, vol. 318, pp. 167–171, 1993.

[195] F. Agou, S. Quevillon, P. Kerjan, and M. Mirande., "Switching the amino acid specificity of an aminoacyl-trna synthetase," *Biochemistry*, vol. 37, pp. 11309–11314, 1998.

[196] D. Thompson, C. Lazennec, P. Plateau, and T. Simonson., "Probing electrostatic interactions and ligand binding in aspartyl-trna synthetase through site-directed mutagenesis and computer simulations," *Proteins*, vol. 71, pp. 1450–1460, 2008.

[197] C. Berthet-Colominas, L. Seignovert, M. Hartlein, M. Grotli, and S. Cusack., "The crystal structure of asparaginyl-trna synthetase from *thermus thermophilus* and its complexes with atp and asparaginyl-adenylate: The mechanism of discrimination between asparagine and aspartic acid," *EMBO J.*, vol. 17, pp. 2947–2960, 1998.

[198] S. Polydorides, N. Amara, C. Aubard, P. Plateau, T. Simonson, and G. Archontis, "Computational protein design with a generalized born solvent model: Application to asparaginyl-trna synthetase," *Proteins*, 2011.

[199] M. S. am Buch, A. Lopes, D. Mignon, and T. Simonson., "Computational protein design: Software implementation, parameter optimization and performance of a simple model," *J. Comp. Chem.*, vol. 29, pp. 1092–1102, 2008.

[200] S. Henikoff and J. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 10915–10919, 1992.

[201] A. D. Mackerell Jr., D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher III, B. Roux, , M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkewicz-Kuczera, D. Yin, and M. Karplus., "An all-atom empirical potential for molecular modelling and dynamics study of proteins," *J. Phys. Chem. B*, vol. 102, pp. 3586–3616, 1998.

[202] J. Arnez, K. Flanagan, D. Moras, and T. Simonson., "Engineering a mg^{2+} site to replace a structurally conserved arginine in the catalytic center of histidyl-trna synthetase by computer experiments.," *Proteins*, vol. 32, pp. 362–380, 1998.

[203] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein., "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, pp. 926 – 935, 1983.

[204] T. Darden, D. York, and L. Pedersen., "Particle Mesh Ewald: An n log(n) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, pp. 10089–10092, 1993.

[205] S. Nose., "A unified formulation of the constant temperature molecular dynamics method," *J. Chem. Phys.*, vol. 81, pp. 511–519, 1984.

[206] W. Hoover., "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, pp. 1695–1697, 1985.

[207] S. Feller, Y. Zhang, R. Pastor, and B. Brooks., "Constant-pressure molecular-dynamics simulation: The Langevin piston method," *J. Chem. Phys.*, vol. 103, pp. 4613–4621, 1995.

[208] J. Ryckaert, G. Ciccotti, and H. Berendsen., "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, pp. 327–341, 1977.

[209] B. Brooks, C. B. III, A. M. Jr., L. Nilsson, R. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. Pastor, C. Post, J. Pu, M. Schaefer, B. Tidor, R. Venable, H. Woodcock, X. Wu, W. Yang, D. York, and M. Karplus., "Charmm: The biomolecular simulation program," *J. Comp. Chem.*, vol. 30, pp. 1545–1614, 2009.

[210] W. Im, D. Beglov, and B. Roux., "Continuum solvation model: A computation of electrostatic forces from numerical solutions to the poisson-boltzmann equation," *Comp. Phys. Comm.*, vol. 111, pp. 59–75, 1998.

[211] R. Peterson, P. Dutton, and A. Wand, "Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library.," *Protein Sci*, vol. 13, pp. 735–751, 2004.

Appendix A

Computational Design

A.1 XPLOR Files for the Calculation of Interaction Matrix Elements

The residue coefficients B_R and the interaction-energy calculation was done by the program XPLOR, using in-house input files. The procedure is illustrated in Fig. 9.3 and detailed described in Section 9.2.1. In the first stage, we computed residue self-energies and the corresponding residue coefficients B_R. These residue solvation radii were employed in the second stage, for computing the interaction energies between all possible pairs of sidechain - sidechain and sidechain - backbone. Parts of the XPLOR input files used in the above calculations are shown below.

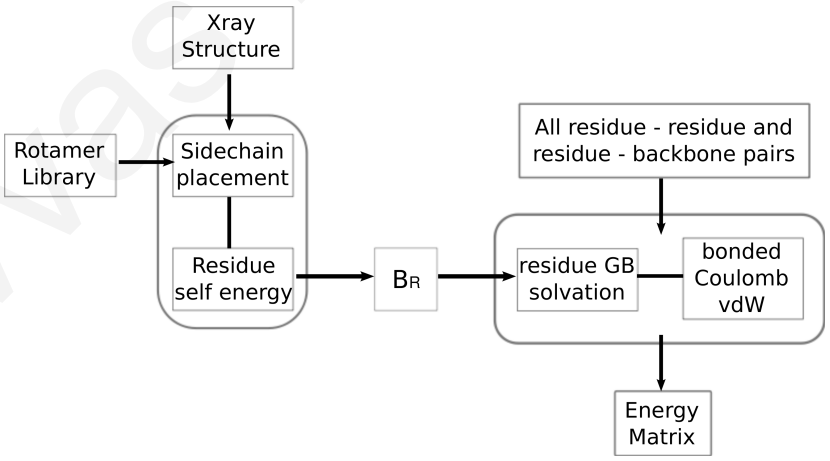


Figure A.1: Flowchart of the computational procedure for the preparation of interaction energy matrices prior to the design with Proteus.

A.1.1 Calculation of Residue Solvation Radii

```
!-------------------------------------------
! Read topology and parameter libraries
!-------------------------------------------
topology    @TOPPAR:Gianttoph-3_GB.pro end
topology    @TOPPAR:amber_s.inp end
topology    @TOPPAR:amberpatches_asn.pro end
topology    @TOPPAR:toph19.sol end
topology    @TOPPAR:topMG.inp end
topology    @TOPPAR:top_ASN_adenylate.inp end

parameters @TOPPAR:paramambernew.inp end
parameters @TOPPAR:param19.sol end
parameters @TOPPAR:par_ADEN.inp end

!-----------------------------
! Non-bonded energy options
!-----------------------------
parameters
BOND OC P 270. 1.60
ANGLe C OC P 20 127.5
ANGLe OS P OC 100 105.0
ANGLe O2 P OC 80 100.4
NONBONDED MG 0.0150    1.18500     0.0150    1.18500 ! Magnesium
end
parameter nbonds
tolerance=0.25 atom cdie shift eps=8.0 e14fac=0.83333
cutnb=21. ctofnb=20.5 ctonnb=20. vswitch
gbhct offset=0.0 lambda=1.0
?
end end

flags include bonds angl dihe impr vdw elec gbse gbin end


!-----------------
! Prepare protein
!-----------------

structure  @data/protein_C3_min.psf end
coordinates  @data/protein_C3_min.pdb

hbuild selec=(segid "SOLD" and name h* and not known) end

!-----------------
! Prepare ligand
!-----------------
```

```
structure  @data/bound_liga_min.psf end
coordinates  @data/bound_liga_min.pdb

hbuild selec=(segid "COM2" and name h* and not known) end


!Inactive residues
vector do (b=1)(all)
!Active residues
vector do (b=2)(resid 1 or resid 3 or resid 4 or resid 5
or resid 6 or resid 7 or resid 9 or resid 10 or resid 11
or resid 13)

! Delete everything else
!-------------------------------------
write coor selec = (not known)end
delete select=(not known) end

!-----------------------------
!   Create giant residue 999
!-----------------------------

segment
  name="GIAN"
  chain
  sequence ALL ALA ASP ASN ARG GLU GLN HIE HIP ILE LEU
           LYS MET PHE SER TYR THR TRP VAL end
  end
end

vector do (resid="999")(segid "GIAN")
patch gala  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ALA) end
patch gasp  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ASP) end
patch gasn  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ASN) end
patch garg  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ARG) end
patch gglu  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname GLU) end
patch ggln  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname GLN) end
patch ghie  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname HIE) end
patch ghip  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname HIP) end
patch gile  refe="1"=(resid 999 and resname ALL)
```

```
refe="2"=(resid 999 and resname ILE) end
patch gleu  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname LEU) end
patch glys  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname LYS) end
patch gmet  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname MET) end
patch gphe  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname PHE) end
patch gser  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname SER) end
patch gtyr  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname TYR) end
patch gthr  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname THR) end
patch gtrp  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname TRP) end
patch gval  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname VAL) end

vector do (segid= "") (segid GIAN)

!-------------------------
! Store residues of interest
!-------------------------

vector identify (store2) (attr b = 2 )

vector identify (store3) ((attr b < 2 ) and not
(resn PRO or resn GLY or resn CYS or resn ALA or resid 999))

vector identify (store1) ( store2 )


!-------------------
! Define the backbone
!-------------------

vector ident (store9)
(resn PRO or resn CYS or resn GLY or (resn ALA and not
(store2 or resid 999)) or (name ca or name ha or name n
or name hn or name h or name c or name o or
name ht* or name nt* or name hy* or name cay))

!-------------------------
! Loop over active residues
!-------------------------
for $i in id ( name CA and store1 ) loop m1
```

```
evaluate ($Act1 = 0)

vector show (resid) (id $i)
evaluate ($1 = decode($result))

coor copy end

!-------------------------------
! Adjust the volume parameters
!-------------------------------
@TOPPAR:volumes.amber
flags exclude * include gbse gbin end
parameter reduce selection=(all) overwrite=true mode=average
end end

flags include bonds angl dihe impr vdw elec gbse gbin end


!-------------------------------------------------
! Loop over amino acid types for current residue
!-------------------------------------------------
for $aa1 in (ALA ASP ASN ARG GLU GLN HIE HIP
             ILE LEU LYS MET PHE SER TYR THR TRP VAL ) loop maa1


evaluate ($rotfile = "DIREVO:macro/" + $aa1 + "_nrot.txt")
@@$rotfile

evaluate ($rot1 = 1)

while ($rot1 <= $nbrot) loop m2
evaluate ($Act1 = $Act1 + 1)

! ALL.pdb contains backbone coors for giant residue
coordinates  @DIREVO:rotamers/ALL.pdb

! Place 999 backbone coor on top of the current residue's backbone
@DIREVO:macro/Position_Giant_GB.inp

! Get coors of current rotamer
evaluate
($coor = "DIREVO:rotamers/" + $aa1 + "_" + encode($rot1) + ".pdb")
coordinates  @@$coor

! Move 999 on top of current residue by superimposing backbones,
! putting the 999 sidechain in place
coor fit sele ( resid 999 and
```

```
 (name ca or name n or name c or (name cb and resnam $aa1) ) ) end

! Restore the original protein backbone
coor swap sele (not (resid 999 and not (name ca or name ha or name
 cb or name hb1 or name hb2 or name n or name hn or name h or name
ht* or name c or name o or name nt* or name cay or name hy*))) end

!-------------------------------------
! Adjust bonds, angles, dihedrals
!-------------------------------------

! Fix backbone, PRO and CYS sidechains, chain terminals
constraints fix=(store9 or name CB) end

! Dihedral restraints for current sidechain
evaluate ($Dihe1 = "DIREVO:dihedrals/DIHE_"+ $aa1 + "_" +
 encode($rot1) + ".inp")
restraints dihedral reset nassign=300 scale=1.0 @@$Dihe1 end
flags include cdih end

! Current sidechain interacts with itself and backbone
constraints
inter (resid 999 and resn $aa1)
      (resid 999 and (resn ALL or resn $aa1))
inter (resid 999 and resn $aa1)
      ((store9 and not (resid 999 or resid $1)))
end

flags exclude gbse gbin end

! Minimize slightly to improve geometry
minimize powell drop=10 nstep=15 nprint=5 end
flags include gbse gbin end

! Energy call to compute correctly the atomic b

parameter nbonds
tolerance=0.25 atom cdie shift eps=8.0 e14fac=0.83333
cutnb=21. ctofnb=20.5 ctonnb=20. vswitch
gbhct bato  offset=0.0 lambda=1.0
?
end end

constraints
inter ((all and not (resid 999 or resid 998 or resid $1)) or
( resid 999 and ( resn $aa1 or resn ALL) ) )
((all and not ( resid 999 or resid 998 or resid $1)) or
( resid 999 and (resn $aa1 or resn ALL) )) end
```

```
energy end

parameter nbonds
tolerance=0.25 atom cdie shift eps=8.0 e14fac=0.83333
cutnb=21. ctofnb=20.5 ctonnb=20. vswitch
gbhct bcon offset=0.0 lambda=1.0
?
end end

flags include bonds angl dihe impr vdw elec gbse gbin cdih end

!=================================
! Calculate the Bres coefficients
!=================================

! Compute the self energy of sidechain i due to the rest

constraints
inter (residue 999 and resn $aa1)
((all and not (resid 999 or resid 998 or resid $1)) or
(resid 999 and (resn $aa1 or resn ALL))) end

energy end

vector do (store8 = charge * charge)
(resid 999 and resn $aa1)
vector show sum (store8) (resid 999 and resn $aa1)
evaluate ($sumq2 = $RESULT )

vector do (store8 = charge * charge / bsolv )
(resid 999 and resn $aa1)
vector show sum (store8) (resid 999 and resn $aa1)
evaluate ($sumq2b = $RESULT )

evaluate ($bres = $sumq2 / $sumq2b )

vector do (bsolv=$bres) (resid 999 and resn $aa1)
vector show ave (bsolv) (resid 999 and resn $aa1)
evaluate ($newb = $RESULT)

evaluate ($koto = "$bresi")
evaluate ($filename= "DIREVO:bsolv/bres.sc." +
 encode($1) + "." + $aa1 + "." + encode($rot1) + ".dat")
set display=$filename end
display  evaluate ( $koto = $newb )
close $filename end
```

```
! save current coordinates
coor copy end

evaluate ($rot1 = $rot1 + 1)

end loop m2

end if

end loop maa1

end loop m1

stop
```

A.1.2 Interaction Energy Matrix

```
eval ($rotactif =216)

!-----------------------------
! Create giant residue 999
!-----------------------------

segment
  name="GIAN"
  chain
  sequence ALL ALA ASP ASN ARG GLU GLN HIE HIP ILE
           LEU LYS MET PHE SER TYR THR TRP VAL end
  end
end

vector do (resid="999")(segid "GIAN")
patch gala  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ALA) end
patch gasp  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ASP) end
patch gasn  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ASN) end
patch garg  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ARG) end
patch gglu  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname GLU) end
patch ggln  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname GLN) end
patch ghie  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname HIE) end
patch ghip  refe="1"=(resid 999 and resname ALL)
```

```
refe="2"=(resid 999 and resname HIP) end
patch gile  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname ILE) end
patch gleu  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname LEU) end
patch glys  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname LYS) end
patch gmet  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname MET) end
patch gphe  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname PHE) end
patch gser  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname SER) end
patch gtyr  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname TYR) end
patch gthr  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname THR) end
patch gtrp  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname TRP) end
patch gval  refe="1"=(resid 999 and resname ALL)
refe="2"=(resid 999 and resname VAL) end

vector do (segid= "") (segid GIAN)

!-------------------------------
!  Create giant residue 998
!-------------------------------

segment
  name="GIAN"
  chain
  sequence ALL ALA ASP ASN ARG GLU GLN HIE HIP ILE
           LEU LYS MET PHE SER TYR THR TRP VAL end
  end
end

vector do (resid="998")(segid "GIAN")

patch gala refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname ALA) end
patch gasp refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname ASP) end
patch gasn refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname ASN) end
patch garg refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname ARG) end
patch gglu refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname GLU) end
```

```
patch ggln refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname GLN) end
patch ghie refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname HIE) end
patch ghip refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname HIP) end
patch gile refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname ILE) end
patch gleu refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname LEU) end
patch glys refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname LYS) end
patch gmet refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname MET) end
patch gphe refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname PHE) end
patch gser refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname SER) end
patch gtyr refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname TYR) end
patch gthr refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname THR) end
patch gtrp refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname TRP) end
patch gval refe="1"=(resid 998 and resname ALL)
refe="2"=(resid 998 and resname VAL) end

vector do (segid= "") (segid GIAN)


!-----------------------------
! Store residues of interest
!-----------------------------

evaluate ($7 = $pos999)


! ACTIVE residues are identified by their b value
vector identify (store2) (attr b = 2 )

! INACTIVE residues are put in store3
vector identify (store3) ((attr b < 2 ) and not
(resn PRO or resn GLY or resn CYS or resn ALA or resn AMO
or resid 999 or resid 998))

! Break-up the ligand in groups
!-----------------------------
! group A
```

```
vector ident (store4)(resn AMO and (name CB or name HB* or
name CG or name OD1 or name OD2))
! group B
vector ident (store5)(resn AMO and (name N or name HT* or
name CA or name HA or name C or name O or name OXT))
! the fixed part
! group C
vector ident (store6)(resn AMO and (name P or name O1P or
name O2P or name O5' or name C5' or name H5' or name H5''))
! group D
vector ident (store7)(resn AMO and (name C4' or name H4' or
name O4' or name C1' or name H1' or name C2'or name H2'' or
name O2' or name H2' or name O3' or name H3T))
! group E
vector ident (store8)(resn AMO and (name N9 or name C8 or
name H8 or name N7 or name C5 or name C4 or name N3 or
name C6 or name N1 or name C2 or name H2 or name N6 or
name H61 or name H62))


!--------------------
! Define backbone:
!--------------------

vector ident (store9) (segid LIGA or resn PRO or resn CYS or
resn GLY or (resn ALA and not (store2 or resid 999 or
resid 998)) or (name ca or name ha or name n or name hn or
name h or name c or name o or name ht* or name ot* ))

!-------------------------
! Loop over Active residues
!-------------------------
for $i in id ( name CA and store1 ) loop m1

! initialize the rotamer counter (will go from 1 to 216)
evaluate ($Act1 = 0)

! get current resid
vector show (resid) (id $i)
evaluate ($1 = decode($result))

@name2a.txt
set display=$matrix end

coor copy end

! adjust the volume parameters
!-----------------------------
```

```
@TOPPAR:volumes.amber
flags exclude * include gbse gbin end
parameter reduce selection=(all) overwrite=true mode=average
end end

flags include bonds angl dihe impr vdw elec gbse gbin end


!-------------------------------------------------------------
! Loop over amino acid types for current residue
! we broke the calculation in 4 files;
! here we compute ALA, ASP, ASN, ARG -> 56 out of 216
!-------------------------------------------------------------
for $aa1 in ( ALA ASP ASN ARG ) loop maa1

if ($aa1 = ALA) then

evaluate ($Act1 = $Act1 + 1)

! ALL.pdb contains backbone coors for giant residue
coordinates @DIREVO:rotamers/ALL.pdb

! place 999 backbone coors on top of the current residue's backbone
@DIREVO:macro/Position_Giant_GB.inp

! get coors of current rotamer
coordinates @DIREVO:rotamers/ALA.pdb

! Move 999 on top of current residue by superimposing backbones,
! putting the 999 sidechain in place
coor fit sele ( resid 999 and
(name ca or name n or name c or (name cb and resnam $aa1) ) ) end

! Restore the original protein backbone
coor swap sele (not (resid 999 and not (name ca or name ha or
name cb or name hb1 or name hb2 or name n or name hn or
name h or name ht* or name c or name o or name ot*))) end

!-------------------------------------------------------------
! Compute sidechain:backbone energy + sidechain with itself
!-------------------------------------------------------------

parameter nbonds
tolerance=0.25 atom cdie trunc eps=8.0 e14fac=0.83333
cutnb=15. ctofnb=14. ctonnb=13.
gbhct bcon offset=0.0 lambda=1.0
?
```

```
end end

flags include bonds angl dihe impr vdw elec gbse gbin end

!assign the residue B coefficient to the mutant sidechain
evaluate ($bresname = "DIREVO:bsolv/bres.sc." + encode($1) +
"." + $aa1 + ".dat")
@@$bresname
vector do (bsolv = $bresi) (resid 999 and resn $aa1)


! Assign B to the backbone segments
for $k in id (name CA and (store9 and not
(resid 999 or resid 998 or resid $1))) loop back

vector show (resid) (id $k)
evaluate ($bb = decode($result))

evaluate ($bresname = "DIREVO:bsolv/bres.bb." + encode($bb) +
".dat")
@@$bresname
vector do (bsolv = $bresi) (store9 and resid $bb)

end loop back


! assign B to the ligand groups
evaluate ($bresname = "DIREVO:bsolv/bres.lig.data")
@@$bresname
vector do (bsolv = $bresi) (store4)
evaluate ($bresname = "DIREVO:bsolv/bres.lig.datb")
@@$bresname
vector do (bsolv = $bresi) (store5)
evaluate ($bresname = "DIREVO:bsolv/bres.lig.datc")
@@$bresname
vector do (bsolv = $bresi) (store6)
evaluate ($bresname = "DIREVO:bsolv/bres.lig.datd")
@@$bresname
vector do (bsolv = $bresi) (store7)
evaluate ($bresname = "DIREVO:bsolv/bres.lig.date")
@@$bresname
vector do (bsolv = $bresi) (store8)


! assign B to the backbone part of the active sc
evaluate ($bresname = "DIREVO:bsolv/bres.bb."
+ encode($1) + ".dat")
@@$bresname
```

```
vector do (bsolv = $bresi) (resid 999 and resn ALL)


constraints
inter (residue 999 and resn $aa1)(residue 999 and resn $aa1)
inter (residue 999 and resn $aa1)(residue 999 and resn ALL)
inter (residue 999 and resn $aa1)(store9 and not
(resid 999 or resid 998 or resid $1))
end

energy end

evaluate ($Ubb = $ener - $elec - $gbse - $gbin)
evaluate ($Uelec = $elec)

display fENER $1 ACT $rotactif $Act1 $Ubb $Uelec $gbse $gbin
display 9999  $1 ACT $rotactif


! Save current coordinates
coor copy end

for $j in id ( name CA and store2 ) loop m3

evaluate ($flap = -9999)
evaluate ($Act2 = 0)

! get current resid
vector show (resid )(id $j)
evaluate ($2 = decode($result))

! get current resname
vector show (resname) (resid $2)
evaluate ($nomres2 = $RESULT)

! Initialize marker for a pair to be computed
evaluate ($mark = 1)

!-----------------------
! Apply filters
!-----------------------

if ($2 > $1) then evaluate ($mark = 0)
elseif ($2 = $1) then evaluate ($mark = 0)

else

! Prepare first distance filter
```

```
pick bond (resid $2 and name CB)(resid $1 and name CB) geom
evaluate ($cbcb = $result)

! Apply first distance filter
if ($cbcb > 15.0) then evaluate ($mark = 0)
end if

end if

if ($mark > 0.5) then

!-------------------------------------------------------
! Loop over amino acid types for current residue
!-------------------------------------------------------

for $aa2 in  ( ALA ASP ASN ARG GLU GLN HIE HIP ILE LEU
               LYS MET PHE SER TYR THR TRP VAL ) loop maa2


! Get number of rotamers from $fichier
evaluate ($fichier = "DIREVO:macro/" + $aa2 + "_nrot2.txt")
@@$fichier

evaluate ($rot2 = 1)
while ($rot2 <= $nbrot2) loop m4

evaluate ($Act2 = $Act2 + 1)

! Get 998 backbone coordinates
coordinates  @DIREVO:rotamers/ALL_998.pdb

! Place 998 backbone coors on top of the current residue's backbone
@DIREVO:macro/Position_Giant_998_GB.inp

! Read current rotamer
evaluate ($coor2 = "DIREVO:rotamers/" + $aa2 + "_"
+ encode($rot2) + "_998.pdb")
coordinates @@$coor2

coor fit sele ( resid 998 and
(name ca or name n or name c or
(name cb and resnam $aa2) ) ) end

! Restore the original protein backbone
coor swap sele  (not (resid 998 and not (name ca or name ha or
name cb or name hb1 or name hb2 or name n or name hn or
name h or name ht* or name c or name o or name ot*))) end
```

```
! Set up second distance filter
!-------------------------------
constraint inter (resid 999 and (resnam $aa1 or resnam ALL))
(resid 998 and (resnam $aa2 or resnam ALL) )end
vector do (rmsd = 0.00)(all)

parameter nbonds
cutnb=15.1
end end

distance from=(resid 999 and (resnam $aa1 or resnam ALL))
 to=(resid 998 and (resnam $aa2 or resnam ALL))
 cuton=1.0 cutoff=14.1 disp=rmsd end
vector show  (rmsd) (resid 999 and (resnam $aa1 or resnam ALL))
vector show mini (rmsd)(attr  rmsd > 0.00)
evaluate ($dmin= $result)

parameter nbonds
tolerance=0.25 atom cdie trunc eps=8.0 e14fac=0.83333
cutnb=15. ctofnb=14. ctonnb=13.
gbhct offset=0.0 lambda=1.0
end end

if ($dmin < 12.01) then

! Fix backbone
constraints fix=(store9 or resn AMO or name CB) end

! Apply dihedral restraints
evaluate ($Dihe2 = "DIREVO:dihedrals/DIHE_"+ $aa2 +
"_" + encode($rot2) + "_998.inp")
restraints dihedral nassign=300 reset @@$Dihe2 scale=1.0 end
flags include cdih end

! Current rotamer interacts with backbone
constraints
interaction (resid 998 and resnam $aa2)
            (resid 998 and (resnam ALL or resnam $aa2) )
interaction (resid 998 and resnam $aa2)
            (store9 and not (resid 999 or resid 998 or resid $2))
end

flags exclude gbse gbin end
! Minimize to improve pair geometry with respect to backbone
minimize powell drop=10 nstep=15 nprint=5 end
flags include gbse gbin end
```

```
!-------------------------------------------------
! Adjust 998 and 999 together, without backbone
!-------------------------------------------------

constraints fix=(store9 or resn AMO or name CB) end

if ($aa1 # ALA) then
! Apply dihedral restraints to 999 and 998
evaluate ($Dihe1 = "DIREVO:dihedrals/DIHE_"+ $aa1 + "_" +
encode($rot1) + ".inp")
restraints dihedral nassign=300 reset scale=1.0 @@$Dihe1
@@$Dihe2 end
end if

! Apply selected interactions
constraints
interaction (resid 998 and resnam $aa2)
            (resid 998 and (resnam ALL or resname $aa2))
weight * 0 bonds 1 angl 1 dihe 1 impr 1 end
interaction (resid 999 and resnam $aa1)
            (resid 999 and (resnam ALL or resnam $aa1))
weight * 0 bonds 1 angl 1 dihe 1 impr 1 end
interaction (resid 998 and resnam $aa2)
            (resid 999 and resnam $aa1)
weight * 0 vdw 1 elec 1 end
interaction (resid 999 and resname $aa1)
            (store9 and not (resid 999 or resid 998 or resid $1))
weight * 1 end
interaction (resid 998 and resnam $aa2)
            (store9 and not (resid 999 or resid 998 or resid $2))
weight * 1 end
end

flags exclude gbse gbin end

! Minimize 998 and 999 together
flags include cdih end
minimize powell drop=10 nstep=30 nprint=15 end
flags include gbse gbin end

evaluate ($bresname = "DIREVO:bsolv/bres.sc." + encode($2) +
"."  + $aa2 + "." + encode($rot2) + ".dat")
@@$bresname
vector do (bsolv = $bresi) (resid 998 and resn $aa2)


if ($aa1 # ALA) then
evaluate ($bresname = "DIREVO:bsolv/bres.sc." + encode($1) +
```

```
"."  + $aa1 + "." + encode($rot1) + ".dat")
@@$bresname
vector do (bsolv = $bresi) (resid 999 and resn $aa1)
end if

if ($aa1 = ALA) then
evaluate ($bresname = "DIREVO:bsolv/bres.sc." + encode($1) +
"."  + $aa1 + ".dat")
@@$bresname
vector do (bsolv = $bresi) (resid 999 and resn $aa1)
end if

!-------------------------------------
! Compute vdw and elec terms
!-------------------------------------
constraint
interaction (resid 999 and resnam $aa1)
(resid 998 and resnam $aa2)
end

energy end

evaluate ($PWvdw  = $vdw )
evaluate ($PWelec = $elec)


!---------------
! Write output
!---------------

if   ($flap =-9999) then
display $2
end if
evaluate ($flap = $Act2)
display  $Act2   $PWvdw    $PWelec   $gbse  $gbin

end if

!-----------------
! Terminate loops
!-----------------

flags include bonds angl dihe impr vdw elec gbse gbin end


! Swap main and comp to restore coors prior to minimization
coordinates swap end
```

```
coor copy end

evaluate ($rot2 = $rot2 + 1)

end loop m4

end if

end loop maa2

end if

end loop m3


stop
```

A.2 Computational Requirements of CPU Calculations

Table A.1 lists the time demands (in CPU hours on a single 3.06 GHz Xeon(TM) processor) needed for a complete design calculation.

Table A.1: Timetable of the computational design

Step	Process Description	Time
Before the design		
1	Residue solvation radii B_R	24
2	Interaction energy matrix (pairwise) 5 active and 233 inactive positions	720
3	Unfolded state reference energies	24
During the design		
4	100,000 heuristic cycles of stability	24
5	100,000 heuristic cycles of affinity	120
6	100,000 heuristic cycles of specificity	120
After the design		
7	Sequence post-treatment analysis filtering / rotamer optimization reconstruction / minimization	1
8	Explicit-solvent MD simulations (29750 atoms, 4ns)	720
9	PBFE calculations	0.1

All values in CPU hours on a single 3.06 GHz Xeon(TM) processor.
Steps 7, 8 and 9 correspond to a single sequence / structure.

Appendix B

CPD of AsnRS Amino Acid Specificity With a Residue-GB Solvent Model

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Computational protein design with a generalized Born solvent model: Application to Asparaginyl-tRNA synthetase

Savvas Polydorides,[1] Najette Amara,[2] Caroline Aubard,[2] Pierre Plateau,[2] Thomas Simonson,[2]* and Georgios Archontis[1]*

[1] Department of Physics, University of Cyprus, PO20537, CY1678, Nicosia, Cyprus
[2] Department of Biology, Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, 91128 Palaiseau, France

ABSTRACT

Computational Protein Design (CPD) is a promising method for high throughput protein and ligand mutagenesis. Recently, we developed a CPD method that used a polar-hydrogen energy function for protein interactions and a Coulomb/Accessible Surface Area (CASA) model for solvent effects. We applied this method to engineer aspartyl-adenylate (AspAMP) specificity into Asparaginyl-tRNA synthetase (AsnRS), whose substrate is asparaginyl-adenylate (AsnAMP). Here, we implement a more accurate function, with an all-atom energy for protein interactions and a residue-pairwise generalized Born model for solvent effects. As a first test, we compute aminoacid affinities for several point mutants of Aspartyl-tRNA synthetase (AspRS) and Tyrosyl-tRNA synthetase and stability changes for three helical peptides and compare with experiment. As a second test, we readdress the problem of AsnRS aminoacid engineering. We compare three design criteria, which optimize the folding free-energy, the absolute AspAMP affinity, and the relative (AspAMP-AsnAMP) affinity. The sequences and conformations are improved with respect to our previous, polar-hydrogen/CASA study: For several designed complexes, the AspAMP carboxylate forms three interactions with a conserved arginine and a designed lysine, as in the active site of the AspRS:AspAMP complex. The conformations and interactions are well maintained in molecular dynamics simulations and the sequences have an inverted specificity, favoring AspAMP over AsnAMP. The method is not fully successful, since experimental measurements with the seven most promising sequences show that they do not catalyze at a detectable level the adenylation of Asp (or Asn) with ATP. This may be due to weak AspAMP binding and/or disruption of transition-state stabilization.

Proteins 2011; 00:000–000.

Key words: computational protein design; implicit solvent models; generalized Born model; Poisson Boltzmann calculations; molecular dynamics simulations; Asparaginyl-tRNA synthetase; aminoacyl-tRNA synthetases; genetic code; protein-ligand interactions.

INTRODUCTION

Computational Protein Design (CPD) enables systematic, high throughput protein, and ligand mutagenesis and has been the focus of several laboratories in recent years, resulting in significant developments in methodology and applications.[1–13] CPD has been used to modify specificity,[14–17] to improve protein-ligand binding,[18–24] to increase stability,[25,26] to stabilize novel or alternative protein folds,[27,28] to perform fold recognition and homology searching,[5,11] to design new proteins,[29,30] and enzyme active sites,[31,32] to optimize ligand entrance and escape pathways,[33] to create water-soluble variants of membrane proteins,[34] to redesign protein-protein interfaces,[20,22,35,36] and to rewire biological networks.[37]

In this work, we will focus on the computational design of a protein-ligand complex, and especially on the effect of an improved treatment of aqueous solvent through a generalized Born model. This methodological question has broad relevance for protein modelling in general. Particularly, the CPD problem has a close relationship to the problem of computing protein acid/base constants, or pK_a's. Indeed, whereas CPD explores different aminoacid sidechain types and their preferred conformations, pK_a calculations explore different sidechain titration states and their preferred conformations. Thus, the methodology and software described below were also applied recently to the pK_a problem, as described in.[38]

Solvent-mediated effects contribute to protein stability and function,[39–42] including enzymatic reactions,[43] protein-protein association,[44] and ligand recognition.[45–50]

Additional Supporting Information may be found in the online version of this article.
The authors state no conflict of interest.
Grant sponsor: Cyprus Research Promotion Foundation; Grant number: PENEK-SUPPORT/0505/04; Grant sponsor: Cyprus-France ZENON; Grant number: CY-FR/0907/04
*Correspondence to: Georgios Archontis, Department of Physics, University of Cyprus, PO20537, CY1678, Nicosia, Cyprus E-mail: archonti@ucy.ac.cy (or) Thomas Simonson, Department of Biology, Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, 91128 Palaiseau, France, E-mail: thomas.simonson@polytechnique.fr.
Received 29 November 2010; Revised 25 February 2011; Accepted 3 March 2011
Published online 25 March 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23042

Nevertheless, they are rarely modeled explicitly in CPD studies, due to the combinatorial complexity introduced by solvent molecules.[51] Instead, they are usually incorporated implicitly through effective free energy terms.[52,53]

The simplest implicit-solvent (Coulomb/Accessible Surface Area or CASA) models include a surface-area-based solvation energy,[54–56] combined with a Coulombic energy that is screened by a constant or distance-dependent scaling factor. Such models[17,57] are routinely used in CPD calculations.[17,57–60]

A more accurate treatment is obtained by the Poisson-Boltzmann (PB) approximation.[41,53,61,62] This model has been succesfully used in many applications, including ligand binding,[49] protein-protein binding,[63] and protein dynamics.[64] However, the PB solvation energy is a many-body quantity, which depends on the shape of the entire solute and the complementary volume occupied by the high-dielectric solvent and cannot be ordinarily expressed as a sum of terms involving residue or atom pairs.[65] For this reason, it cannot be directly used in high throughput CPD calculations, which require the precomputation and storage of residue-residue interaction energies, evaluated by residue-pairwise expressions of the protein total energies.[27,65–68] To deal with this problem, Mayo and coworkers[68–70] have developed and employed in CPD calculations a pairwise approximation to the PB model.

A continuum-electrostatics approximation that is more efficient than PB is the generalized Born (GB) model.[71–74] It contains the same physics as the PB approximation (and employs also a many-body free-energy function) but allows for an analytical expression of the solvation free energy (and forces) in terms of the solute atomic coordinates.[71] The GB model has been used successfully in calculations of small molecule solvation,[75–78] peptide self-assembly,[79,80] protein solvation,[81] acid/base equilibria,[38,82–84] protein dynamics,[85,86] ligand binding,[87,88] protein folding,[89–92] structure refinement,[93–95] and protein design.[96]

Previously, we developed an automated CPD procedure, which used a polar-hydrogen molecular-mechanics energy function,[97] combined with a CASA implicit-solvent model.[5,57,98,99] We parameterized and tested this model for sidechain placement, protein solvation energies, protein stability changes, and ligand binding changes due to point mutations[57,99] and applied it to the complete redesign of 95 small proteins, obtaining predicted amino acid sequences with quality that was comparable with other CPD implementations.[5,11,99] We then employed this model in CPD calculations, aiming to engineer a modified aminoacid specificity into the protein Asparaginyl-tRNA synthetase (AsnRS).[17]

AsnRS belongs to the family of aminoacyl-tRNA synthetases (aaRSs), which catalyze the first step in the translation of the genetic code by attaching a specific amino acid to a cognate tRNA molecule.[100,101] The specificity of aaRSs for their amino acid and tRNA ligands is crucial for the correct translation of the genetic code.[101,102] Several groups have investigated the contributions of various residues to aaRS binding and catalysis, and have engineered aaRSs with modified amino acid specificity.[103–106] We have used *in silico* site-directed mutagenesis and free energy simulations to study the amino acid specificity of aspartyl-tRNA synthetase (AspRS).[45–47,107] Such studies contribute to our understanding of aaRS function and can lead to engineered organisms with a modified genetic code.[106]

Our recent CPD calculations with the polar-hydrogen/CASA model focused on the AsnRS complex with the non-native ligand Aspartyl adenylate (AspAMP).[17] We explored five active-site positions (residues 187, 190, 225, 227, 366 in *Thermus thermophilus* AsnRS[108]), seeking to identify sequences/structures with low folding free energies (high stabilities). Molecular dynamics (MD) simulations of selected designed sequences and Poisson-Boltzmann Free Energy (PBFE) calculations showed that the AsnRS specificity was reversed (AspAMP binding was favored by 11–37 kcal/mol over the natural substrate asparaginyl adenylate or AsnAMP). However, the computed AspAMP affinities were substantially worse than the native AsnAMP affinity. Furthermore, in the simulations the active site structures became distorted with respect to the native AsnAMP complex, with a bent ligand geometry.

Because of these shortcomings, in the present work, we reconsider the AsnRS aminoacid specificity problem. Here, we use an improved free energy function and a combination of design criteria, which take into account not only stability but also affinity or specificity. We describe protein and peptide interactions by an all-atom energy model[109] and treat solvent effects implicitly by the GB/HCT formulation.[110] Two recent advances make this model attractive for CPD calculations. First, we have shown that a careful parameterization of the GB/HCT [110] approximation can yield accurate protein solvation free energies and free-energy changes due to mutations in fully or partly buried positions.[57] Here, we test its accuracy further by binding-affinity calculations for several point mutants of the Aspartyl-tRNA synthetase and Tyrosyl-tRNA synthetase. Second, we recently introduced an accurate residue-pairwise variant of the GB model,[111] which is suitable for CPD. We used this model successfully to study acid/base equilibria in proteins.[38] A major goal of the present work is to check the performance of this GB model in a challenging CPD calculation.

A second objective and novelty of the present work with respect to our previous design of the AsnRS: AspAMP complex[17] is that we compare the performance of three design criteria: a maximum stability criterion, which minimizes the folding free energy of the complex; an (absolute) affinity criterion, which minimizes the AsnRS binding free energy for the AspAMP ligand; finally, a relative affinity criterion, which optimizes binding of AspAMP relative to AsnAMP.

The sequences suggested by the present CPD calculations are more consistent with the properties of the

AspRS active site, compared with the results of the earlier, polar-hydrogen/CASA design.[17] The combined stability/affinity criteria often predict the insertion of a charged (Lys) or polar (His) sidechain at position 187, which form new ligand interactions. The Lys187-containing sequences have an AspRS-like ligand-recognition mode, in which the ligand carboxylate interacts with the key residue Arg388 (Arg489 in *E. coli* AspRS) and with Lys187 (Lys198 in *E. coli* AspRS). The two combined criteria predict, respectively, 11 and 12 sequences that bind AspAMP more strongly than AsnAMP.

To assess the quality of the designed sequences, we study selected complexes with AspAMP or AsnAMP by explicit-solvent MD simulations and PBFE calculations. The active-site conformations of the designed complexes are well maintained and the ligand-recognition mode of Lys187-containing AsnRS sequences is AspRS-like.[112] Furthermore, the sequences are predicted to have an inverted specificity, favoring Asp.

The AspRS-like protein-AspAMP interactions in the AsnRS active site, the conformational stabilities of the designed sequences in the MD simulations and their increased relative (Asp - Asn) affinities suggest that the GB-HCT implicit-solvent treatment and the combined stability/affinity criteria constitute improvements over our earlier stability/CASA design.[17] This is the main result of the present work. To test further the success of our design, we performed activity measurements with the seven most promising sequences. Unfortunately, the tested sequences could not catalyze the adenylation reaction of Asp or Asn with ATP.

METHODOLOGY

Effective energy function

The design calculations employ the effective energy function

$$E = E_{\text{prot}} + E_{\text{solv}} \tag{1}$$

E_{prot} is the protein internal energy. In the present work, it is taken from the AMBER all-atom energy function[109]

$$E_{\text{prot}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihe}} + E_{\text{impr}} + E_{\text{vdw}} + E_{\text{Coul}} \tag{2}$$

The terms on the right-hand side (rhs) of Eq. (2) represent, respectively, energies of covalent bonds, bond angles, torsional angles, the chirality or planarity of certain atomic centers, van-der Waals and Coulomb electrostatic interactions.

E_{solv} is a solvation

$$E_{\text{solv}} = \left(\frac{1}{\varepsilon_{\text{p}}} - 1\right)\frac{q_i q_j}{r} + E_{\text{GB}} \tag{3}$$

The first term on the rhs is a screened Coulomb energy; it is different from zero only if the employed protein/ligand dielectric constant $\varepsilon_{\text{p}} \neq 1$. A constant $\varepsilon_p = 8$ was used in the present work. The second term is a generalized-Born (GB) model. Here, we use a residue-pairwise variant[111] of the GB/HCT approximation,[110] with parameters recently optimized for protein-design calculations.[57] Our residue-GB model is described in more detail below.

The residue GB approximation

In the standard, "atomic-GB" approximation, the solvation energy is computed by the Eq.[71]

$$E_{\text{GB}} = \sum_i E_i^{\text{self}} + \sum_{i<j} E_{ij}^{\text{inte}} = \tau \sum_i \frac{q_i^2}{2b_i} + \sum_{i<j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/(4b_i b_j)]}}, \tag{4}$$

where r_{ij} is the distance between atoms i and j, $\tau = 1/\varepsilon_w - 1/\varepsilon_p$, ε_w is the solvent dielectric constant (80 at room temperature), ε_p is the protein dielectric constant, and b_i, b_j are effective atomic "solvation radii" of the atoms i, j. The first term is a sum of "atomic self-energies," corresponding to the interaction of each atomic charge q_i with its own reaction field in the environment of the solvated biomolecule. The second term models the interaction of a charge q_i with the reaction field produced by a different charge q_j and accounts for the screening of electrostatic interactions by the high-dielectric solvent. The functional form of this interaction, adopted in Eq. (4), is the original and most frequently-used form of Still *et al.*[71] The employed atomic volumes and screening factors have been extensively optimized for protein design calculations in our earlier work.[57]

In the "residue-GB" approximation, the interaction between two residues R, R′ is given by the following ansatz formula[111]

$$E_{RR'}^{\text{inte}} \equiv \tau \sum_{i\in R,\, j\in R'} \frac{q_i q_j}{\sqrt{r_{ij}^2 + B_R B_{R'} \exp[-r_{ij}^2/(4B_R B_{R'})]}}, \tag{5}$$

where the summation is over all atoms of the two residues R, R'. Equation (5) has the same functional form as the atomic-GB expression [second term on the rhs of Eq. (4)], except that all atoms in a particular residue R are assigned a common solvation radius B_R. The two expressions have the same behavior in the united-atom and independent-atom limits.

To compute the residue-specific radii B_R in (4), we define the self-energy of residue R as

$$E_R^{\text{self}} \equiv \sum_{i\in R} E_i^{\text{self}} = \tau \sum_{i\in R} \frac{q_i^2}{2b_i}, \tag{6}$$

This "residue" self-energy is exactly equal to the sum over all residue atoms of the corresponding atomic self-energies [first term on the rhs of Eq. (5)]. We associate the residue self-energy with B_R via the relation

$$B_R = \tau \sum_{i \in R} \frac{q_i^2}{2E_R^{\text{self}}}. \quad (7)$$

Equations (6–7) imply that

$$\left(\sum_{i \in R} q_i^2 \right) \frac{1}{B_R} = \sum_{i \in R} \frac{q_i^2}{b_i}, \quad (8)$$

i.e., B_R is a harmonic average over the b_i, $i \in R$, weighted by the squared charges.

The self-energy term employed here corresponds to the GB/HCT functional form.[81,110] Because the GB/HCT self-energy is pairwise additive, it can be computed and stored in a residue-interaction matrix prior to the design calculation. Residue coefficients B_R can be computed from this matrix and Eq. (7) and used efficiently via a pairwise-decomposable approximation described in Ref. 111. Here, we employ a simpler approximation, as in Refs. 65,68–70. For each active and inactive sidechain, we compute a list of possible coefficients B_R, taking into account all possible sidechain chemical types and/or rotamer orientations; in this calculation, all other sidechains have their native chemical type and the conformation of the crystallographic structure.

System

Our calculations focused on AsnRS from *Thermus thermophilus*, shown in Figure 3.[108] AsnRS is a homodimer, with 438 amino acids per monomer. The crystallographic coordinates of the AsnRS complex with AsnAMP were supplied by the authors. In the calculations, we used a 20 Å-radius sphere centered on the AsnAMP ligand. The mainchain atoms of the protein (including the C_β atoms and the sidechains of Cys, Pro, and Gly residues) were fixed in their crystallographic positions. The sidechains of five residues near the ligand (Gln187, Ala190, Glu225, Glu227, and Ser366) were allowed to change sidechain chemical type and conformation; these residues are referred to as "Active". Mutations to the following 18 chemical types were allowed: Ala, Asp, Asn, Arg, Glu, Gln, His, protonated-His, Ile, Leu, Lys, Met, Phe, Ser, Tyr, Thr, Trp, and Val. All other sidechains maintained the chemical type of the native sequence but were allowed to explore different conformations from the backbone-independent Tuffery rotamer library[113]; these residues are referred to as "Inactive". The AspAMP and AsnAMP ligands were also treated as Inactive. Rotamers for these molecules were produced earlier through MD simulations, as described in Ref. 17. We used 161 rotamers for AspAMP and 163 for AsnAMP.

The unfolded state

In the unfolded state, we assumed that amino acid sidechains did not interact with each other but only with proximal backbone groups and solvent.[67,98] With this approximation, the unfolded state was represented by a collection of n tripeptide structures with the sequence Ala-X-Ala, with n the number of amino acids in the protein. The total free energy of the unfolded state was obtained by summing the contributions of the n individual amino acids in the protein. To compute the contribution of each amino acid chemical type, we used a large collection of backbone tripeptides from six protein structures; details are given in Ref. 17. We took into account solvent effects by the GB/HCT approximation[57,110] with a protein dielectric constant $\varepsilon_p = 8$ and a solvent dielectric constant $\varepsilon_w = 80$. The resulting contributions for all amino acid chemical types are listed in Table I.

Table I
Amino Acid Energies Corresponding to the Unfolded State (in kcal/mol)

Amino Acid	Energy
Ala	1.60
Asp	−5.67
Asn	−4.16
Arg	−17.8
Glu	−0.98
Gln	−0.26
His	20.14/24.13[a]
Ile	9.27
Leu	7.93
Lys	8.44
Met	3.76
Phe	3.91
Ser	1.84
Tyr	4.22
Thr	−1.24
Trp	5.95
Val	7.62

[a]Neutral/Protonated Histidine

Computational design

Our CPD procedure consisted of three stages. In the first stage, we partitioned the ligand and protein into segments. The AspAMP and AsnAMP ligands were divided into five segments, corresponding to the adenine base, the ribose sugar, the phosphate backbone, the amino acid backbone, and the amino acid sidechain. Gly, Ala, Cys, and Pro amino acids were considered as a single segment; all other amino acids were divided into a backbone segment (including the C_β atom) and a sidechain segment. For each segment, we computed the corresponding GB self-energy, with the approximation that the rest of the molecule had the native sequence and conformation. The resulting self-energies and the corresponding residue coefficients [B_R; see Eq. (7)] were stored. In the second stage, we computed the interaction energies between pairs of sidechain and backbone segments, or between pairs of sidechain segments, taking into account all posible sidechain chemical types and

orientations. In this calculation, the interaction GB energy terms [Eq. (5)] employed the residue coefficients derived in the first stage.

The interactions between all possible pairs of sidechain–backbone or sidechain–sidechain segments were computed as follows. For each active (inactive) position i, all possible chemical types/rotamers (rotamers) were minimized by 15 steps with the Powell conjugate-gradient algorithm in the presence of the fixed protein backbone. During the minimization, solvent effects were included by scaling the Coulomb energy with a constant dielectric factor $\varepsilon = 8$. At the end of the minimization, the interaction energy between the sidechain and the backbone was computed in the residue-GB approximation and stored into a file. Subsequently, the interactions between the sidechain pairs (i, j) were considered. Interactions were set to 0 if the $C_\beta(i) - C_\beta(j)$ distance was greater than 15 Å; otherwise, the pair was subjected to 30 minimization steps. During the minimization, solvent effects were included by scaling the Coulomb energy with a constant dielectric factor $\varepsilon = 8$. At the end of minimization, the interaction energy between the pair was computed in the residue-GB approximation and stored in a file. All calculations were performed with the XPLOR program,[114] using in-house scripts.

In the third stage, the AsnRS sequence and structure were optimized, using a modified version of the Proteus program.[17,67] We employed optimization protocols, based on three general criteria; (i) A maximum-stability criterion, which identifies sequences that minimize the folding free-energy of the protein (or complex); (ii) An absolute-affinity criterion, which identifies sequences that minimize the association free-energy for a specific ligand; (iii) A relative-affinity criterion, which identify sequences that minimize the association free-energy for one ligand, relative to a second ligand. These criteria are explained in detail below.

Criterion of maximum stability

The difference in folding free-energies between the native (P) and a designed (P^*) sequence is given by the following relation

$$\Delta\Delta G = (G_{P^*_N} - G_{P^*_D}) - (G_{P_N} - G_{P_D}) \quad (9)$$

where N and D denote, respectively, the folded and denatured states. Because the last term on the right-hand side of Eq. (9) is constant, the design minimizes the free-energy difference $G_{P^*_N} - G_{P^*_D}$.

Proteus uses a heuristic design procedure, first developed and validated by Wernisch *et al.*[67] Briefly, a "heuristic cycle" for stability (or rotamer) optimization proceeds as follows: An initial amino acid sequence and a set of rotamers are chosen randomly. These are improved in a step-wise way. At a given amino acid position, the best amino-acid type and rotamer are selected, with the rest of the sequence held fixed. The same is done for all subsequent positions, performing multiple passes over the amino acid sequence until the energy no longer improves or a predefined large number of passes is reached. The final sequence, rotamer set and energy are output. This protocol is used in design calculations that optimize the folding free-energy. Affinity calculations employ a modified heuristic cycle, as explained below.

Criterion of maximum absolute affinity

The association free-energy of a designed complex (P^*L), relative to the native complex (PL), is

$$\begin{aligned}\Delta\Delta G &= (G_{P^*L} - G_{P^*} - G_L) - (G_{PL} - G_P - G_L) \\ &= (G_{P^*L} - G_{P^*}) - (G_{PL} - G_P) \quad (10)\end{aligned}$$

Since the last parenthesis on the rhs is constant, the design minimizes the first parenthesis. We place all nonfixed residues in a random initial state (sidechain chemical type and rotamer orientation for active positions, rotamer orientation for inactive positions). During a heuristic cyle, we consider all (nonfixed) residues of the complex and free protein. At each active position i, we insert the same chemical type j in the complex and in the free protein and determine the rotamers $k_{P^*L}(i, j)$ and $l_{P^*}(i, j)$ that minimize, respectively, the free energies G_{P^*L} and G_{P^*}. Finally, we choose the chemical type j that minimizes the free-energy difference $G_{P^*L} - G_{P^*}$ and we place its sidechain in the optimum rotamer $k_{P^*L}(i, j)$ (complex P^*L) or $l_{P^*}(i, j)$ (free protein P^*). At each inactive position of the complex and free protein, we place the sidechain in the (possibly different) rotamer orientations that minimize the free energies G_{P^*L} and G_{P^*}.

The absolute-affinity criterion can yield sequences with low protein stability (due to a large G_{P^*} value). To design sequences with high affinity and near-native stability, we also employed a combined stability/affinity criterion. In this protocol, we optimize the weighted sum $w_a\Delta G + w_s G^*$, where $\Delta G \equiv G_{P^*L} - G_{P^*}$ is the same free-energy difference as above, and $G^* \equiv (G_P + G_{PL})/2$ is the arithmetic mean of the free-protein and complex folding free energies. For the special cases $(w_a = 0, w_s = 1)$ and $(w_a = 0.5, w_s = 0.5)$, this criterion is equivalent to a simple stability optimization of the either free protein or the complex.

Criterion of relative affinity

These calculations identify protein mutations that maximize the relative binding affinity of a protein (P) for a ligand L_1, compared with a second ligand L_2. The relative affinity $\Delta\Delta G$ is

$$\begin{aligned}\Delta\Delta G \equiv \Delta G_{PL_2} - \Delta G_{PL_1} &= (G_{PL_2} - G_P - G_{L_2}) \\ &\quad - (G_{PL_1} - G_P - G_{L_1}) \\ &= (G_{PL_2} - G_{PL_1}) - (G_{L_2} - G_{L_1}) \quad (11)\end{aligned}$$

where $G_{PL_{1/2}}$ and $G_{L_{1/2}}$ are, respectively, the free energies of the two complexes and the isolated ligands. The corresponding relative affinity for a designed protein (P^*) is

$$\Delta\Delta G^* = (G_{P^*L_2} - G_{P^*L_1}) - (G_{L_2} - G_{L_1}) \qquad (12)$$

The change in relative affinity due to the inserted protein mutation is

$$\begin{aligned}\Delta\Delta\Delta G &\equiv \Delta\Delta G^* - \Delta\Delta G \\ &= (G_{P^*L_2} - G_{P^*L_1}) - (G_{PL_2} - G_{PL_1}) \qquad (13)\end{aligned}$$

The term in the last parenthesis of Eq. (13) does not depend on the protein mutation. Thus, it is sufficient to minimize the first term, $G_{P^*L_2} - G_{P^*L_1}$.

To apply the relative-affinity criterion, we follow a procedure analogous to the absolute-affinity protocol. (i) At each active position i, we insert the chemical type j that minimizes the free-energy difference $(G_{P^*L_2} - G_{P^*L_1})$, in the rotamer conformations $k_{P^*L_2}(i, j)$ (complex P^*L_2) and $l_{P^*L_1}(i, j)$ (complex P^*L_1) that minimize, respectively, the free energies $G_{P^*L_2}$ and $G_{P^*L_1}$. (ii) At each inactive position, we choose (possibly different) rotamers that minimize the free energies $G_{P^*L_2}$ and $G_{P^*L_1}$ of the two complexes.

The relative-affinity criterion can yield sequences with low free-protein stabilities. To avoid this, we also employed a combined relative affinity/stability criterion in the following way. During the heuristic cycle, we also determine the rotamer $m_{P^*}(j)$ minimizing the free-energy of the free protein P^* and compute the combined free-energy $(G_{P^*}L_2 - G_{P^*}L_1) + G_{P^*}$, where G_{P^*} is the folding energy of the free protein. Finally, we place at active position i of the complexes P^*L_1, P^*L_2, and P^* the chemical type j that minimizes the above free-energy combination, in rotamers $k_{P^*L_2}(i, j)$, $l_{P^*L_1}(i, j)$, and $m_{P^*}(i, j)$.

Binding affinity calculations

To test the performance of the GB/HCT-based energy function, we considered several point mutants of the proteins Aspartyl-tRNA synthetase (AspRS) and Tyrosyl-ARNt synthetase (TyrRS), with experimentally measured binding affinities for their respective substrates, Asp[115] and Tyr.[116–120] The sidechain orientations were chosen in two ways. In a "minimum protocol", all invariant sidechains were placed in their crystallographic conformations; for the mutated sidechain, the optimum conformation was chosen among rotamers taken from the Tuffery library.[113] In a second protocol ("Proteus protocol"), an optimization of all protein sidechains was performed by the Proteus sequence/structure optimization program. Details of the methods used are in the Supporting Information.

Molecular dynamics simulations

Selected complexes of designed AsnRS sequences with bound AspAMP or AsnAMP were studied further by MD simulations. The methodology was the same as in Ref. 17. We employed the same protein model as in the design calculations (a 20 Å sphere centered on the adenylate ligand). The initial coordinates of protein mainchain atoms were taken from the crystallographic positions.[108] The sidechains were initially placed in the positions predicted by the design calculations. A single Mg^{+2} ion bound to the ligand α-phosphate was included in the system. The complex was immersed in a 68 Å cubic water box, and any waters overlapping the protein or ligand were removed. Enough sodium or chloride ions were added to neutralize the total charge of the system. The final model contained the ligand, about 4270 protein atoms and about 6500 water molecules. The cubic box was periodically replicated in all directions.

The CHARMM22 forcefield was employed for the protein and the AMP moiety of the ligands.[121] The linkage between the AMP and aminoacid moieties of the ligands was parameterized as in Ref. 122. The water was represented by a TIP3P model.[123] Long-range electrostatic interactions were computed without cutoff by the particle-mesh Ewald method,[124] with a parameter $k = 0.34$ for the charge screening and sixth-order splines for the mesh interpolations. The Lennard-Jones interactions between atom pairs were switched to zero at a cutoff distance of 12 Å. The temperature was kept at 300 K by a Nosé-Hoover thermostat[125,126] using a mass of 1000 kcal/mol ps^2 for the thermostat. The pressure was maintained at 1 Atm with a Langevin piston,[127] using a 500 amu mass and a 5 ps^{-1} collision frequency for the piston. The classical equations of motion were integrated by the Leap-Frog integrator, using a time step of 1 fs. Bond lengths to hydrogen atoms and the internal water geometry were constrained to standard values with the SHAKE algorithm.[128] All simulations were performed with the CHARMM program, version c35b3.[129]

The system was initially equilibrated by 60 ps of dynamics, during which all nonhydrogen protein atoms were harmonically restrained around the initial positions. The harmonic restraints were progressively varied from 5 to 1 kcal/mol/$Å^2$ (mainchain) and 0.5 kcal/mol/$Å^2$ (sidechain atoms). Subsequently, all protein atoms in the inner 17 Å-sphere were left free; the remaining protein atoms (excluding hydrogens) were harmonically restrained to their experimental positions; the employed force constants reproduced the corresponding crystallographic B factors. This production phase lasted 4 ns in most simulations.

Poisson-Boltzmann free energy calculations

The electrostatic contribution to the ligand binding free energy was computed by subtracting the electrostatic

free energy in the complex and the isolated ligand and protein. The electrostatic potential was computed by numerical solution of the Poisson-Boltzmann equation, using a finite-difference algorithm implemented in CHARMM.[130] The solvent-protein dielectric boundary was defined by a probe sphere with radius of 2.0 Å. The solvent dielectric constant was set to 80 and the protein/ligand dielectric constants were set to 4, as in Ref. 17. The ionic strength was set to 100 m*M*. PB calculations were performed for 1000 structures, sampled every 4 ps of the MD simulation.

Experimental activity measurements

We tested the activity against L-asparagine and L-aspartate of wildtype AsnRS, seven sequences designed in the present work, and the most promising sequence of our previous CPD calculation.[17] Cloning, purification and activity measurement details are given in the Supporting Information.

RESULTS

Testing the residue-GB model

Comparing residue-GB to atomic-GB for solvation energies

Previously, we showed that residue-GB reproduces Poisson-Boltzmann results at least as well as the parent, atomic-GB model.[38,111] Specifically, we considered a large set of protein solvation energies and solvation energy changes due to chemical mutations and protonation changes. Monte Carlo simulations with residue-GB also led to good agreement with experiment for a large test set of pK_a's.[38] Here, however, we use a simplification of residue-GB, and so additional testing is needed. Indeed, even though the residue-GB approximation is pairwise decomposable,[111] it is still relatively laborious: to update the GB energy following a structure or sequence change, it is necessary to reevaluate GB solvation radii [coefficients B_R in Eq. (5)] for all residues and use them to compute GB interaction energies. Therefore, for this first design application, we use a computationally simpler version of residue-GB, where the solvation radius of a particular sidechain is computed by assuming that the rest of the protein has its native sequence and conformation. This way, the solvation radii can be computed once at the beginning of the calculation. In Figures 1 and 2, we compare this implementation to the parent atomic-GB model, using a set of mutant AsnRS sequences and conformations generated by Proteus.

Figures 1(A,B) compare the residue-GB and atomic-GB interaction energies for sidechain-total backbone and sidechain-sidechain pairs of the native AsnRS sequence and conformation. Note that in this case the approximate and exact residue-GB values are identical, since the former are computed using the native protein. The RMS differences (residue-atomic) are 0.15 kcal/mol and 0.06 kcal/mol, respectively. They arise from the interaction GB terms, because the atomic and residue GB self-energies are the same. Mayo and coworkers[70] developed a residue-pairwise approximation to the Poisson-Boltzmann model for CPD calculations. For a set of 12 proteins in their native sequence and conformation, they reported similar RMS differences from exact PB [0.18 kcal/mol for sidechain-total backbone and 0.05 kcal/mol for sidechain-sidechain pairs; see table I and Figs. 3(D,F) in Ref. 70].

Figures 1(C,D) compare energies of sidechain-backbone pairs for 179 AsnRS sequences with random mutations in the five active positions (187, 190, 225, 227, 366). Each sequence is in a generally different, low-energy conformer. The RMSD between approximate or exact residue-GB and atomic-GB is 0.17 kcal/mol, slightly higher with respect to the native-AsnRS value (0.15 kcal/mol).

Figures 1(E,F) compare sidechain-sidechain interaction energies for a single, low-energy conformer of a designed AsnRS sequence, with substitutions Y187, S190, Q225, K227, S366 in the five active positions. The RMS difference from atomic GB is 0.07 kcal/mol for the exact-residue GB model and 0.27 kcal/mol for the approximate-residue GB model. Thus, the introduction of mutations increases somewhat the deviation from the atomic-GB values.

Finally, in Figure 2, we compare total solvation free-energies for a set of 1800 mutant AsnRS sequences and conformations generated by Proteus. The RMSD between exact residue-GB and atomic-GB energies is 1.3 kcal/mol. When we switch to the simplified residue-GB implementation, the RMSD increases to 4.2 kcal/mol, but the good correlation between residue-GB and atomic-GB energies is retained.

Comparing residue-GB to experimental stability mutants

As another test, we considered stability changes in three short, helical peptides, due to mutations in a single position near the middle of each helix. For each peptide, 17 amino acid types were considered (all but Cys, Gly, and Pro), giving 16 stability changes for each peptide. These mutations were used earlier (along with some others) to test the CASA implicit solvent model.[99] The peptides considered here are pepT1,[132] K2AE2,[133] and KEAKE,[134] of lengths 17, 17, and 21 amino acids. Ideal α-helical structures were built earlier; sidechains were positioned here by rotamer exploration. This led to just 1–3 distinct conformations for any given mutant. Reasonable free energy results were obtained with a peptide dielectric constant of four. The mean unsigned errors for the three peptides were 1.2, 1.1, and 1.0 kcal/mol, respectively. The overall rms error was 1.3 kcal/mol. Other dielectric values (such as 8) were not tried. A Null model (no stability changes) gives a smaller overall mean unsigned error of just 0.4 kcal/mol (the mean unsigned stability change in the experimental dataset). The earlier,

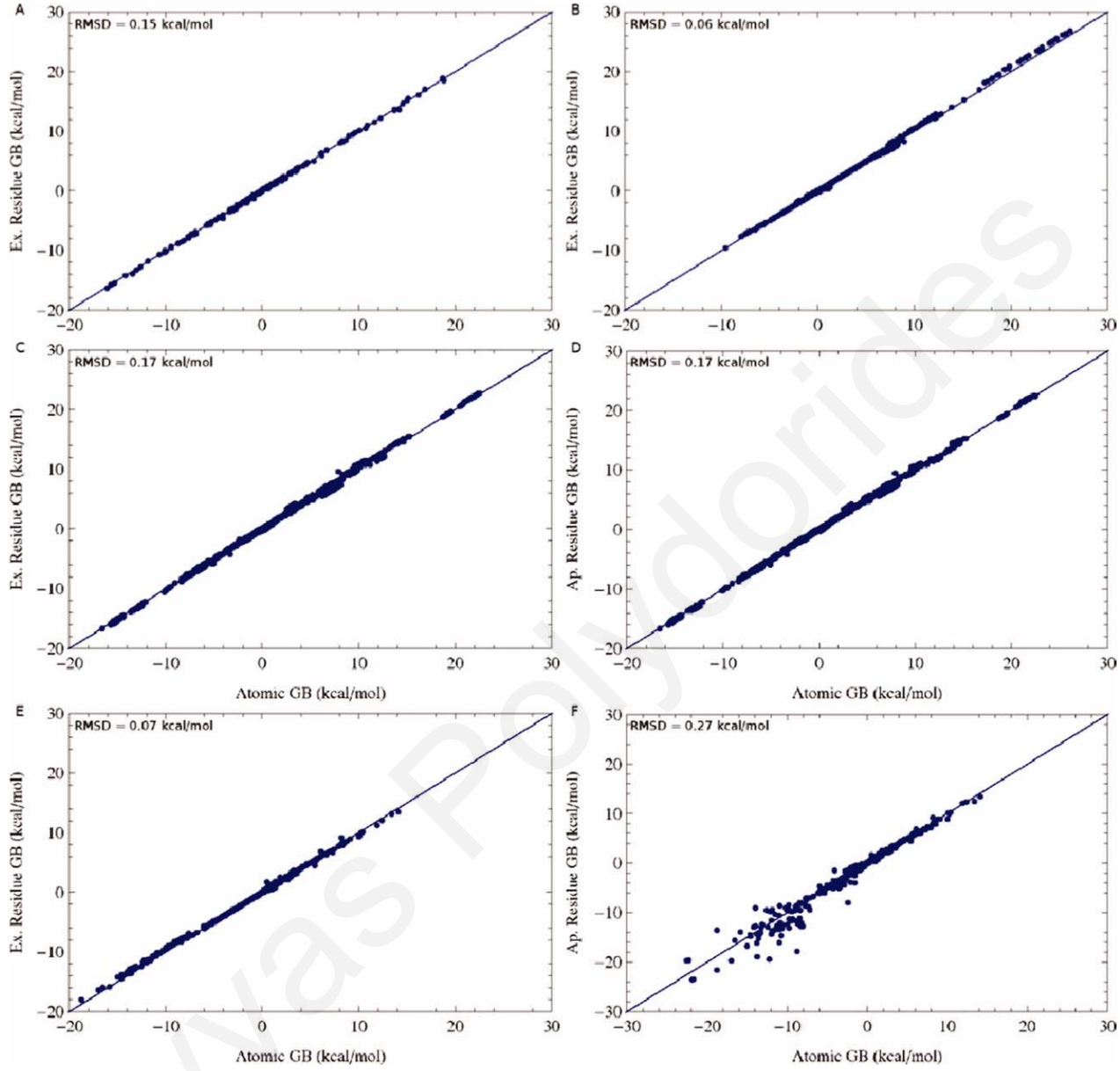


Figure 1

Accuracy of exact and approximate residue-GB, with respect to atomic-GB. (**A**) Residue-GB sidechain-total backbone and (**B**) sidechain-sidechain interaction energies for the native AsnRS sequence and conformation; (**C**) Exact residue-GB and (**D**) approximate residue-GB sidechain-total backbone energies for a set of 179 mutant AsnRS sequences; (**E**) Exact residue-GB and (**F**) approximate residue-GB sidechain-sidechain energies, for a designed AsnRS sequence (see text). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

CASA model gave a mean unsigned error of 2.2 kcal/mol, and an rms error of 2.7 kcal/mol, twice as large as the present errors. More details are given in Supporting Information.

AspRS:Asp and TyrRS:Tyr binding affinities

To test the performance of the GB/HCT-based energy function, we also considered a set of 10 AspRS point mutants and 13 TyrRS point mutants, with experimentally measured binding affinities for their respective substrates, Asp[115] and Tyr.[116–120] Tables 1–3 in Supporting Information list the individual mutations and binding free energy changes with two different protocols. When the sidechain orientations are optimized by rotamer exploration ("Proteus protocol"), the mean unsigned error is 2.1 kcal/mol for AspRS and 1.2 kcal/mol for TyrRS. With the simpler, "minimal" protocol, the mean AspRS error is slightly smaller (0.9 kcal/mol if only the rotamer conformation of the mutated side-chain is optimized and

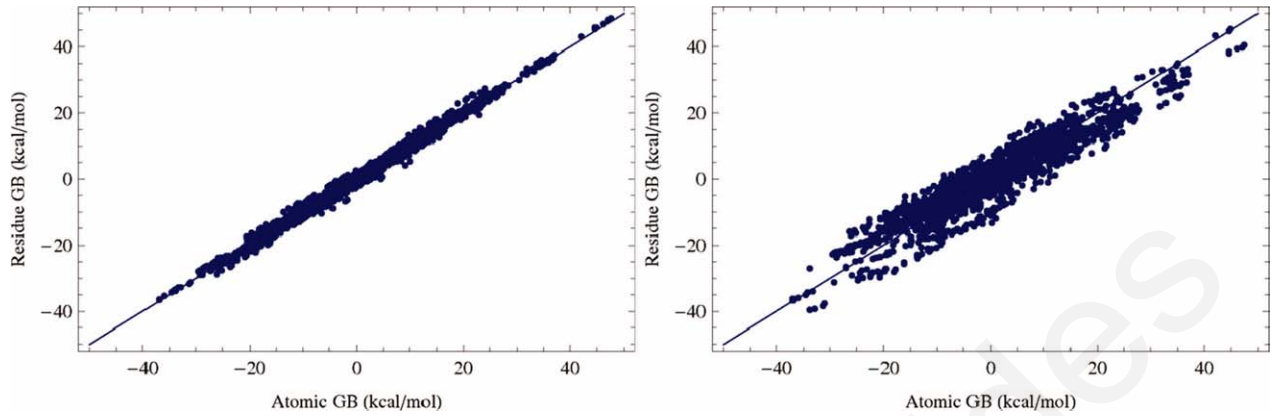


Figure 2

Residue-GB solvation energies for 1800 AsnRS mutant sequences/conformations generated by Proteus, plotted against the corresponding atomic-GB values. Left: Original residue-GB approximation.[111] Right: Simplified residue-GB approximation, used in the present design work (see text). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

1.7 kcal/mol if it is subjected to energy minimization). The best results are obtained with a protein/ligand dielectric of 8. A null model (no affinity changes upon mutation) gives a slightly smaller mean unsigned error of 1.5 kcal/mol for AspRS and a slightly larger one of 1.3 kcal/mol for TyrRS.

AsnRS CPD calculations

The goal of our design calculations is to introduce Asp binding and specificity into Asparaginyl-tRNA synthetase. Overall, the calculations were only partly successful, since the sequences that were experimentally tested were not able to catalyze the adenylation reaction of Asp (or Asn) with ATP (see below). Nevertheless, the MD simulations and PB free energy calculations do suggest that the method yields sequences that have a much better binding affinity than the ones obtained earlier with the CASA solvent model.[17] In this section, we describe the design calculations performed with different protocols and selection criteria. We will then describe the MD testing. The experimental results are mostly in Supporting Information.

Native AsnRS: rotamer exploration

The AsnRS active site is highly homologous with the active site of Aspartyl-tRNA synthetase, with differences mainly in three residues[48]: Ala190 of *Thermus thermophilus* AsnRS is replaced by Lys198 in *E. coli* AspRS; Glu225 is replaced by Asp233, and Glu191 by Gln199. Figure 4(A) shows the active site of *Th. thermophilus* AsnRS, in complex with its cognate AsnAMP ligand. The sidechain of AsnAMP interacts with Arg368 and Glu225; its ammonium group interacts with Ser185, Gln187, and Glu164. Ala190 makes a nonpolar contact with the Glu227 C_β; Glu225 forms a hydrogen-bond with Ser185 and a contact with Met223; the Glu227 sidechain forms a salt-bridge with Arg368. Figure 4(B) displays the *E. coli* AspRS active site in complex with its cognate AspAMP ligand.[112] The ligand sidechain carboxylate forms direct interactions with Arg489, Lys198, and His448. The ligand ammonium interacts with Ser193, Gln195, and Glu171. Arg489 and Lys198 form salt-bridges, respectively, with Glu235 and Asp233. Figure 4(C) shows the initial conformation of the non-native complex between *Th. thermophilus* AsnRS and AspAMP used in the design calculations. The AspAMP adenylate was placed at the same position as in the AsnRS:AsnAMP complex, and the Asp sidechain in a very similar orientation as in the active site of the AspRS:AspAMP complex [Fig. 4(B), 112),

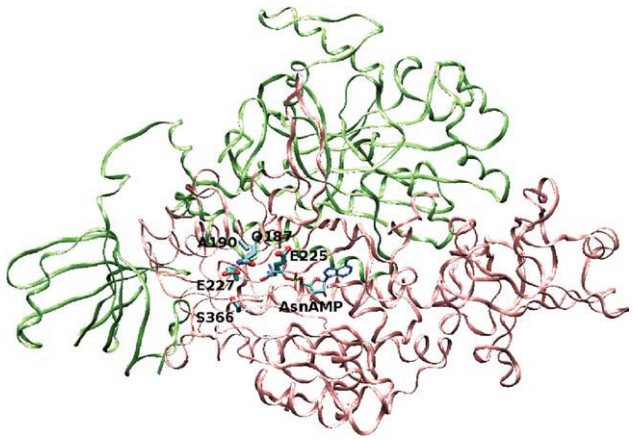


Figure 3

The Asparaginyl-tRNA synthetase homodimer, in complex with the AsnAMP ligand (shown in thin licorice). The five mutatable ("active") residues are indicated by thick licorice lines. This and subsequent figures were created with the program VMD.[131]

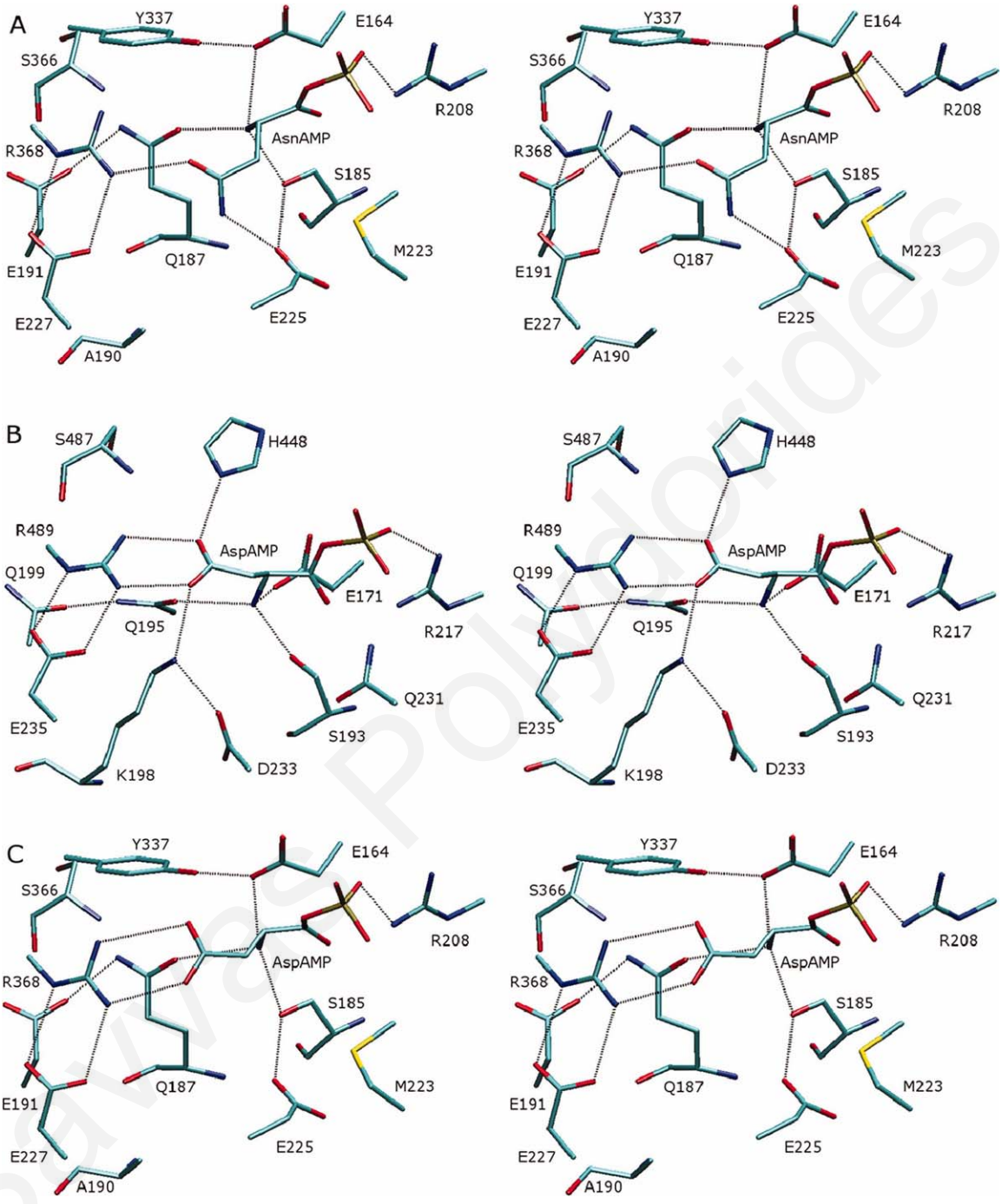


Figure 4

Stereo representations of: (**A**) the active site of the complex *Th. thermophilus* AsnRS:AsnAMP[108]; (**B**) the active site of the complex *E. coli* AspRS:AspAMP[112]; (**C**) a model of the AsnRS active site, with the AspAMP ligand positioned as in the AspRS complex. The nonaminoacid moiety of the ligand beyond the phosphate group is omitted here and in all subsequent figures. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

forming two direct interactions with Arg368. The ligand ammonium retained its interactions with Ser185 and Gln187, as in the AsnRS:AsnAMP complex.

Before doing sequence optimization, we did rotamer optimization for the wildtype AsnRS:AspAMP complex and free AsnRS, using the maximum-stability criterion of section 2.5.1 (20,000 heuristic cycles). For the complex, the optimization produced 9504 different conformations, with folding free energies between −795 and −638 kcal/mol and a mean of 763.3 ± 15.5 kcal/mol. Figure 5

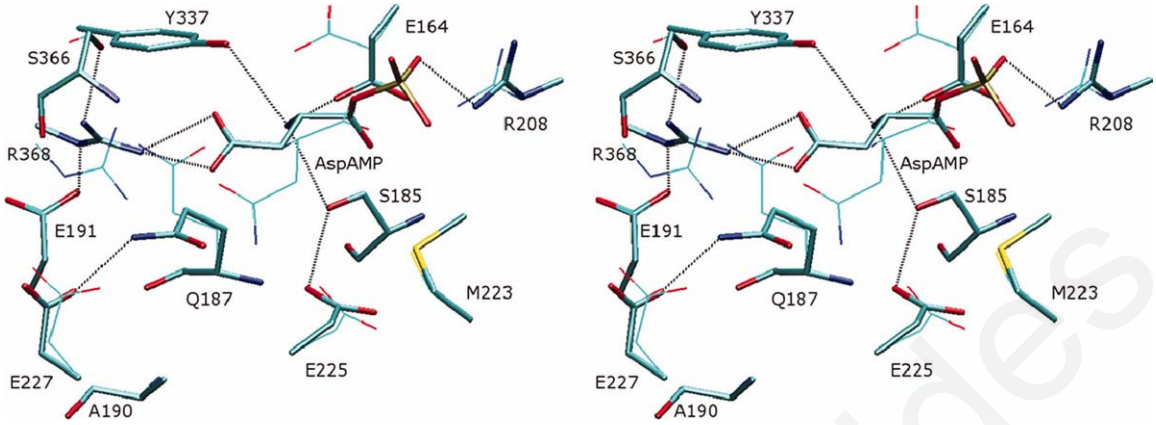


Figure 5

Stereo representation of the conformation of maximum stability (folding free energy = −794.4 kcal/mol) for the complex between native AsnRS and the AspAMP ligand. In thin lines is shown the crystallographic conformation of the AsnRS:AsnAMP complex. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

shows the most stable conformation. Arg368 forms new interactions with Glu191 and Ser366 and retains one interaction with the ligand sidechain carboxylate. Gln187 loses its interactions with the ligand ammonium and Glu191 and forms a new interaction with Glu227. The ligand ammonium compensates its lost Gln187 interaction by improved interactions with Glu164 and Tyr337. In the low free energy conformations (within 20 kcal/mol of the lowest), the active site sidechains are in their crystallographic orientation, with the exception of Glu164, Gln187, Arg208, and Arg368, which explore a few different rotamers (2–6 each); the initial ligand conformation is also maintained in these structures. In free AsnRS, the optimization produced 7156 conformations with folding free energies between −682 and −604 kcal/mol and a mean of −657.0 ± 13.2 kcal/mol. The sidechain orientations were similar to the AsnRS:AspAMP case.

Sequence optimization with the maximum-stability criterion

Next, we optimized the sequence and structure of the AsnRS:AspAMP complex and the free AsnRS (20,000 heuristic cycles each). Five sidechains were "active" (allowed to mutate): 187, 190, 225, 227, and 366. Table II lists all sequences with folding free energies below the mean values for the native protein. The total charge of the five active positions is −2 in the native sequence (QAEES). In the designed sequences, it is less negative (−1 to +1); however, the most stable sequence (YSEES) has the same total charge as the native sequence.

Two of the active positions show minimal variability: Ser487 remains unchanged and Ala190 is always converted to a serine, whose hydroxyl group makes a hydrogen bond with the main chain of residue 225. Gln187 is converted to an aromatic (Tyr), non-polar (Val), or positive (Arg) residue, or left unchanged. The Tyr187 ring is always inserted between the ligand carboxylate and the key recognition residue, Arg368. Figure 6(A) compares the most stable sequence (YSEES) to the most stable conformation of the native sequence. The Arg368 sidechain retains two interactions with Glu227 but adopts a different orientation compared with the crystal structure [Fig. 4(A)] and to the most stable native AsnRS:AspAMP structure (Fig. 5). This orientation is also seen in other sequences with a native Glu at position 227 (YSEES, YSIES, YSRES). A Tyr or Val at position 187 is often combined with a

Table II

AsnRS Sequences that Minimize the Folding Free-energies of the AsnRS:AspAMP Complex and the Free AsnRS Protein

Sequence								
187	190	225	227	366	Charge[a]	AsnRS:Asp[b]	AsnRS[b]	ΔG[c]
		Native						
Q	A	E	E	S	−2	−763.3	−657.0	−106.3
		Designed						
Y	S	E	E	S	−2	−782.7	−672.6	−110.1
Y	S	E	M	S	−1	−781.9	−670.6	−111.3
V	S	E	M	S	−1	−781.5	−670.0	−111.5
Q	S	E	S	S	−1	−780.6	−670.8	−109.8
V	S	I	M	S	0	−779.2	−667.7	−111.5
Y	S	I	E	S	−1	−778.1	−668.3	−109.8
Q	S	I	S	S	0	−777.3	−667.7	−109.6
Y	S	R	E	S	0	−777.0	−672.7	−104.3
R	S	E	S	S	0	−772.0	−673.9	−98.1
Q	S	R	S	S	+1	−771.9	−670.5	−101.4
Y	S	E	S	S	−1	−768.7	−656.2	−112.5
V	S	R	E	S	0	−767.0	−675.9	−91.1
R	S	I	S	S	+1	−763.4	−669.3	−94.1

For each sequence, the reported values are averaged over all rotamer conformations of the Proteus optimization.
[a]Net total charge of the five mutatable residues.
[b]Folding free energies for the AsnRS:AspAMP complex and the free AsnRS protein.
[c]AspAMP binding affinities, estimated by the difference of the two previous columns.
All values in kcal/mol.

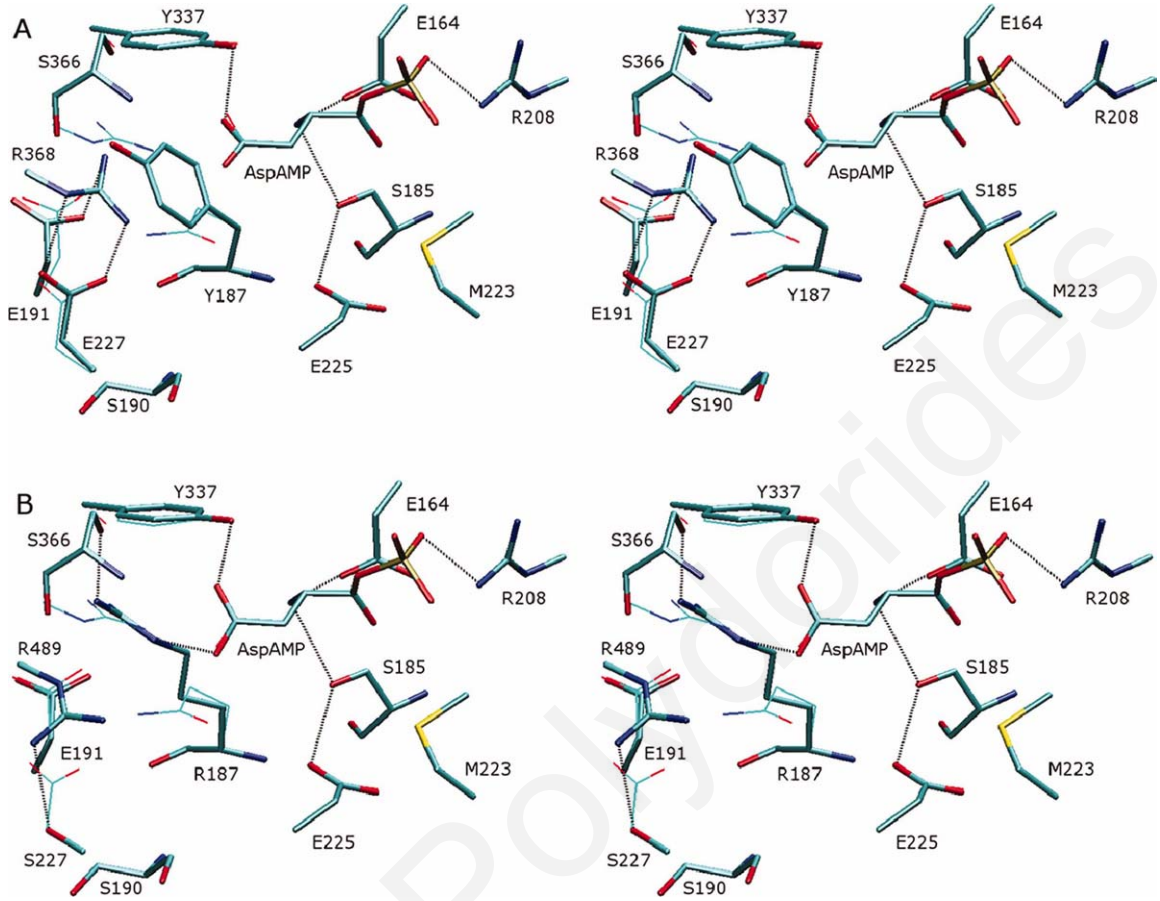


Figure 6

Stereo representations of the AsnRS:AspAMP active-site conformations, for selected sequences identified by the stability criterion (Table II). (**A**) sequence YSEES; (**B**) sequence RSESS. In thin lines is shown the lowest energy conformation of the native complex AsnRS:AspAMP (Fig. 5). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

methionine at position 227; the two sidechains form a nonpolar contact that presumably increases stability. Some sequences contain more extensive nonpolar contacts. For example, VSIMS has a chain of contacts involving Val187, Met227, Ile225, and Met223.

In two sequences, Gln187 is converted to Arg. Interestingly, the Arg187 guanidinium occupies a position similar to Arg368 in the native conformation of greatest stability. This is shown for sequence RSESS in Figure 6(B). The Arg187 guanidinium forms hydrogen bonds with the sidechains of the ligand, Glu191 and Ser366, and packs against the Tyr337 ring. Arg368 rotates toward the sidechain of His403 and makes a hydrogen bond with Ser227. Among the other active positions, Glu225 can become positive (Arg) or nonpolar (Ile). Glu227 can be replaced by Ser or Met. In this case, Arg368 maintains the orientation seen in the most stable native structure (Fig. 5).

Subtracting the folding free energies of the complex and apo-protein, we obtain the relative AspAMP binding affinities of the designed sequences. The values (averaged over sampled rotamers) are given in the last column of Table II. Interestingly, eight sequences are predicted to have better affinity for AspAMP than the native sequence, even though the design is based exclusively on a maximum stability criterion. Six sequences have a Tyr or Val residue in position 187. These sidechains make nonpolar contacts that contribute to both stability and affinity. Six have a Met or Ser substitution at position 227, which eliminates a negative charge near the AspAMP ligand; this substitution does not destabilize the apo protein because Arg368 replaces its lost Glu227 interaction by moving into the orientation of Figure 5, where it interacts with Glu191, Ser366, and the ligand. The introduction of a positive charge at position 225 is not strictly correlated with higher affinity, because the native Glu sidechain interacts with the ligand ammonium.

These results suggest that the protein can gain stability and affinity by introducing an aromatic or nonpolar residue at position 187, along with a non-negative residue at position 227. Nevertheless, the designed sequences are not satisfactory. The Tyr187 ring disrupts a key salt-bridge between Arg368 and the ligand carboxylate [Fig. 6(A)]. Furthermore, rotamer optimization of the complex

between native AsnRS and the cognate ligand, AsnAMP, yields an average folding free energy of −770.6 kcal/mol and a −113.6 kcal/mol affinity for AsnAMP. Thus, the designed sequences do not bind AspAMP as strongly as native AsnRS binds AsnAMP. In the next section, we discuss the optimization results when the affinity for AspAMP is explicitly selected for.

Sequence optimization with the absolute affinity criterion

The absolute-affinity criterion selects sequences that minimize the binding free energy, $G_{PL} - G_P$. This can be achieved by increasing the stability of the complex and/or decreasing that of the apo-protein. To obtain sequences with both a high affinity and near-native stability, we minimized the weighted sum $w_a(G_{PL} - G_P) + (1 - w_a)(G_{PL} + G_P)/2$. We did 100,000 design cycles, with affinity weights w_a = 1.0, 0.9, 0.75, 0.60, or 0.25. With a pure affinity criterion (w_a = 1), the resulting sequences were much less stable than the native complex (by at least 100 kcal/mol). Adding a small stability weight ($w_s \equiv 1 - w_a$ = 0.1 to 0.25) was sufficient to yield sequences with both a high affinity and high stability. At small w_a values, the optimization yielded the same high-stability sequences as in the previous section.

We filtered the resulting sequences in several steps, to identify solutions with high affinity and stability. We first discarded any sequence having average (over rotamers) free energies $\langle G_{PL} - G_P \rangle$ or $\langle (G_{PL} + G_P)/2 \rangle$ larger than the corresponding values for native AsnRS (−106.3 kcal/mol and −710.2 kcal/mol) and subjected the remaining sequences to 10,000 rotamer-optimization cycles. For each sequence, we minimized the 100 lowest-energy rotameric combinations by 90 Powell conjugate-gradient steps. We continued with an additional 10 minimization steps in the presence or absence of the ligand. We computed the final energies with the atomic-GB model and used them to estimate the binding free energy; this result was averaged over the 100 rotamer conformations of the sequence. Repeating the same calculation with the 100 lowest-energy conformations of the native AsnRS:AspAMP complex, we estimated an AspAMP binding affinity of −111.2 ± 1.2 kcal/mol and −61.5 ± 9.5 kcal/mol, before and after minimization. We employed these values as a final selection filter, and discarded any designed sequences with less negative binding affinities.

Sequences produced with weights w_a = 0.75 and 0.9 are reported in Table III. For all designed sequences, the total net charge of the five active positions is less negative than the native case (−2). Positions 190 and 366 are occupied by Ser, as before. The three sequences with the largest affinity (KSTES, KSEES, KSIES) contain a Lys at position 187, conserve Glu227, and differ only at position 225. Figure 7(A) shows the conformation of sequence KSEES, together with the lowest-energy conformation of the native sequence. Lys187 makes two direct interactions with the sidechains of the ligand moiety and of Glu227. Interestingly, the position and orientation of this lysine is similar to that of Lys198 in *E. coli* AspRS, one of two key Asp recognition residues. The ligand sidechain remains in its initial position and interacts with Arg368, Lys187, and Tyr338. Arg368 rotates to a position similar to the lowest-energy native conformation (Figure 5). It loses its salt-bridge with Glu227, maintains one interaction with the ligand, and forms interactions with Glu191 and Ser366. The ligand ammonium interacts with Ser185, Tyr338, and Glu164. The Glu225 sidechain interacts with the Ser185 sidechain and the Lys187 mainchain carbonyl. The same, Lys198-like orientation is observed in the sequences KSTES and KSIES. In KSTES, Thr225 forms a hydrogen bond with the mainchain carbonyl of Met223.

Table III
AsnRS Sequences Optimized by a Combined Criterion of Stability and Affinity for Aspamp

Sequence								
187	190	225	227	366	Charge[a]	G*[b]	ΔG[c]	ΔG[d]
		Native						
Q	A	E	E	S	−2	−736.5	−111.2	−61.5
		Designed						
K	S	T	E	S	0	−737.8	−112.1	−66.3
K	S	E	E	S	−1	−746.5	−111.7	−64.8
K	S	I	E	S	0	−741.2	−112.0	−63.3
H	S	T	E	S	−1	−738.6	−112.3	−63.1
H	S	E	K	S	0	−740.0	−112.0	−63.1
H	S	T	M	S	0	−736.9	−111.8	−62.6
H	S	V	M	S	0	−744.5	−111.3	−62.5
Y	S	Q	K	S	+1	−749.0	−111.7	−62.5
Y	S	E	M	S	−1	−739.1	−111.7	−62.3
Y	S	E	K	S	0	−739.4	−111.7	−61.7
K	S	E	D	S	−1	−737.0	−113.1	−61.6

Affinity/stability weights w_a = 0.7–0.9/w_s = 0.3–0.1 are used in the optimization (see text).
[a]Net total charge of the five mutatable residues.
[b]Arithmetic mean of the folding free energies for the AspRS:Asp complex and the free AspRS protein [$G^* \equiv (G_{PL}+G_P)/2$].
[c]AspAMP binding affinities. The reported values correspond to averages over the 100 lowest-energy rotamer conformations, determined for each sequence by Proteus.
[d]AspAMP binding affinities, resulting after subjecting the 100 lowest-energy rotamer conformations of each sequence to minimization (see text).
All values in kcal/mol.

The last group of sequences contains a His or Tyr at position 187, combined with a Met (as in the stability calculations) or Lys at position 187. His187 (or Tyr187), Arg368, and Tyr337 form a ladder of pi-stacking interactions. Lys227 interacts with the Tyr187 and His403 sidechains and makes a more distant interaction with Glu191.

The designed sequences have marginally better AspAMP affinities than native AsnRS (Table III). This reflects the difficulty to increase AspAMP affinity and maintain at the same time native-like protein stability. The affinities are improved upon minimization (column 9). Overall, introducing a positive (Lys) or polar (His) residue at position 187 increases Asp binding to the same level as that of the native ligand Asn. In the next section,

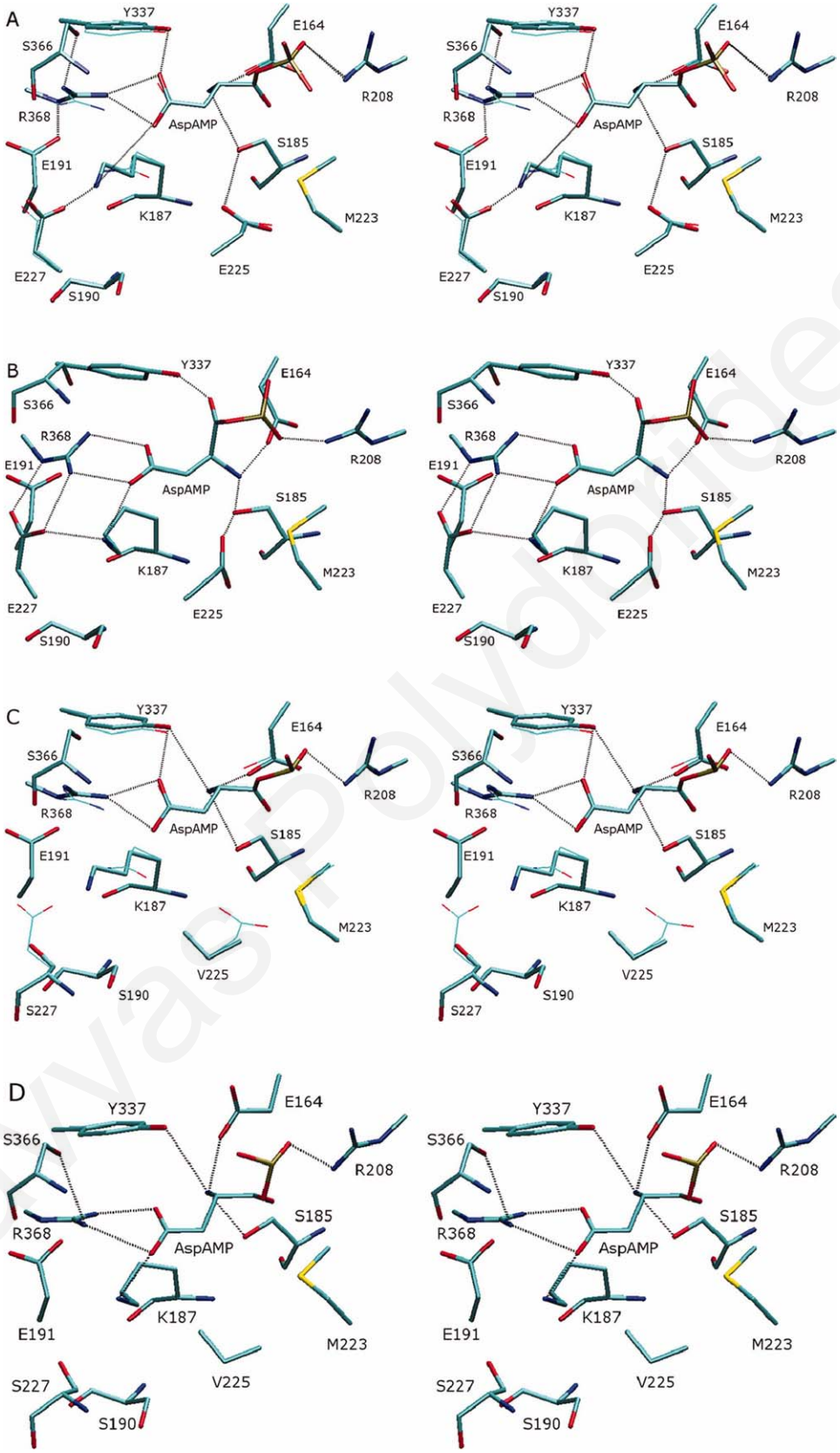


Figure 7

Stereo representations of the AsnRS:AspAMP active-site conformations for the most promising sequence of the stability/absolute affinity design, KSEES (shown in **A**) and stability/relative affinity design, KSVSS (in **C**). In thin lines is shown the native-AsnRS:AspAMP conformation of maximum stability (Fig. 5). **(B)** and **(D)** show the conformations of the same AsnRS:AspAMP complexes (respectively, AsnRS sequences KSEES and KSVSS), at the end of 4-ns MD simulations in explicit water. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table IV
AsnRS Sequences Optimized by a Combined Criterion of Stability and Relative Affinity (AspAMP – AsnAMP)

Sequence									
187	190	225	227	366	Charge[a]	AsnRS[b]	ΔG[c]	ΔG[d]	ΔΔG[d]
Native									
Q	A	E	E	S	−2	−680.9	−111.2[e] −116.4[e]	−61.4[e] −65.3[e]	3.9 (5.2[f])
Designed									
K	S	V	E	S	0	−688.9	−111.7	−66.3	−13.8
K	S	V	S	S	+1	−688.9	−109.0	−62.9	−9.6
H	S	V	M	S	0	−689.2	−111.0	−61.6	−9.7
M	S	V	S	S	0	−687.7	−110.5	−62.0	−9.0
H	S	K	M	S	+1	−689.5	−109.1	−61.5	−8.2
I	S	V	M	S	0	−689.0	−110.1	−62.9	−8.0
V	S	I	M	S	0	−690.5	−110.6	−63.5	−7.9
V	S	V	M	S	0	−691.6	−110.5	−63.1	−5.7
M	S	K	M	S	+1	−689.5	−106.0	−59.0	−5.7
Q	S	K	S	S	+1	−692.2	−108.5	−57.5	−5.0
V	S	K	M	S	+1	−689.2	−111.0	−62.4	−4.9
K	S	K	S	S	+2	−687.9	−107.5	−60.3	−3.8

[a]Net total charge of the five mutatable residues.
[b]Free AsnRS folding free energies.
[c]AspAMP affinity free-energies. For each sequence, the values are averaged over the 100 lowest-energy rotamer conformations generated by the Proteus optimization.
[d]Columns 9 and 10 report, respectively, the average AspAMP affinity and the relative (AspAMP – AsnAMP) affinity, resulting after subjecting the 100 lowest-energy conformations of each designed sequence to minimization (see text).
[e]Affinity of native AsnRS for AspAMP (top line) and AsnAMP (bottom line), before and after minimization.
[f]Relative affinity of native AsnRS for Asp before minimization.
All values in kcal/mol.

we examine whether a relative-affinity criterion, which discriminates against Asn binding, can produce sequences with inverted specificity.

Sequence optimization with the relative affinity criterion

Preliminary calculations showed that a relative-affinity criterion alone (without considering stability) yielded unstable sequences. We thus employed a combined relative-affinity/stability criterion (section "criterion of relative affinity"). We did 50,000 design cycles and postprocessed the resulting sequences as above. In particular, we retained sequences whose relative binding free energy $\langle G_{PL_2} - G_{PL_1} \rangle$ and average folding free energy $\langle G_P \rangle$ were lower than the native AsnRS values (7.3 kcal/mol and −657.2 kcal/mol, respectively). The results are summarized in Table IV. The relative (AspAMP − AsnAMP) binding free energy of the native sequence is +5.2 kcal/mol before and +3.9 kcal/mol after minimization, reflecting optimization of native AsnRS for Asn recognition. All the designed sequences in Table IV have negative relative affinities, i.e., they prefer to bind AspAMP. Furthermore, they are more stable than the native protein and have a somewhat stronger affinity for AspAMP after minimization, except for MSKMS, QSKSS, and KSKSS. The AspAMP affinity of the best sequence (−66.3 kcal/mol) is stronger than the affinity of native AsnRS for AsnAMP (−65.3 kcal/mol). The same is true for sequence KSTES, obtained by the absolute affinity criterion in the previous section (Table III).

For all tabulated sequences, the total net charge of the five active positions is zero or positive; thus, the relative affinity criterion facilitates the introduction of positive charges (or the elimination of negative charges). Some of the sequences/conformations were obtained earlier with the absolute-affinity and stability criteria. The sequences with the best relative and absolute affinities, KSVES and KSVSS [shown in Fig. 7(B)], have the same conformations as the optimum structure KSEES of the absolute-affinity calculation (Fig. 7(A)], except for residue 225. The sequence HSVMS was also found, with the same conformation, in the absolute-affinity calculation. VSIMS was found in the stability optimization of the AsnRS:AsnAMP complex. In all the sequences, Arg368 interacts with the ligand sidechain. When position 187 is occupied by a nonpolar residue (VSKMS, VSIMS, MSVSS), Arg368 is slightly rotated toward the 187 sidechain. In the other sequences (KSVSS, KSVES, HSVMS, and HSKMS), Arg368 is rotated slightly upward, parallel to the Tyr337 ring.

Exploring an alternate active position

The sequences identified by CPD obviously depend on the choice of mutatable ("active") positions. The five positions considered above (187, 190, 225, 227, and 366) were chosen for their proximity to the ligand sidechain and their homology to important recognition residues in AspRS (195, 198, 233, 235, and 487 in *E. coli* AspRS). The design calculations then predicted several mutations that increase the similarity between the AsnRS and AspRS active sites, especially the Q187K mutation. Three key interactions seen in AspRS, involving the AspAMP carboxylate, AspRS-Arg489, and AspRS-Lys198, are also present in the Q187K-AsnRS complexes. However, the AspRS active site contains a fourth key interaction, between the ligand carboxylate and His448 [see Fig. 4(B)]. This interaction does not exist in the AsnRS:Asn complex, and is not inserted by our CPD calculations, because there is no active homologue of AspRS-His448. We therefore examined whether a sixth position could be identified, homologous to the AspRS-His448, which might allow additional mutations stabilizing AspAMP in AsnRS. Sequence alignment and visual inspection of the AspRS and AsnRS active sites show that the His448/His449 pair in *E. coli* AspRS does not have a close homologue in AsnRS. His448 is replaced by a nonpolar residue in AsnRS, and His449 is replaced by Lys334, but the Lys334 sidechain points directly away from the ligand. In fact, the sidechain that appears most likely to accept a His and form a strong interaction with the ligand is Tyr337. This residue is the nearest (after Arg368) to the ligand sidechain within the Proteus-optimized native complex; its C_β atom is about 8 Å from the AspAMP

Table V
Binding Affinities of Selected Designed AsnRS Sequences for AspAMP and AsnAMP

Sequence					AspAMP binding		AsnAMP binding		Asp – Asn binding	
187	190	225	227	366	GB/HCT[a]	PBFE	GB/HCT[a]	PBFE	GB/HCT[a]	PBFE
Q	A	E	E	S	−111.2	−23.7	−116.4	−25.7	5.2	+2.0
K	S	T	E	S	−112.1	−22.7				
H	S	V	M	S	−111.0	−22.4	−109.5	−11.9	−1.5	−10.5
K	S	V	S	S	−109.0	−22.3	−108.8	−17.1	−0.2	−5.2
K	S	E	E	S	−111.7	−21.4				
K	S	V	E	S	−111.7	−19.9	−111.2	−16.6	−0.5	−3.3
H	S	E	K	S	−112.0	−19.3				
K	S	I	E	S	−112.0	−17.9				
Y	S	Q	K	S	−111.7	−15.9				
H	S	T	E	S	−112.3	−14.0				
H	S	T	M	S	−111.8	−13.8	−103.3	−9.2	−8.6	−4.6

[a]Values at the end of Proteus optimization (before minimization) from Tables III and IV.
The affinities were computed by PBFE calculations on equilibrium conformations, obtained by explicit-solvent MD runs. All values in kcal/mol.

carboxylate. Other positions are either too far away, poorly oriented to form an interaction with AspAMP, or both. A Tyr337His mutation was therefore introduced with a "minimum" protocol (see "Methodology" Section). The His337 sidechain did indeed interact closely with the ligand carboxylate, just 2.6 Å away, but it interfered sterically and electrostatically with the ligand ammonium group (which hydrogen-bonds to native Tyr337). As a result, the protein-ligand interaction energy actually increased (disfavoring binding) by 7 kcal/mol (charged His337) or 3 kcal/mol (neutral His337). Thus, there is no obvious, additional, active position, that is consistent with the AsnRS backbone fold and appears likely to allow a strong interaction, homologous to the His448-ligand interaction in AspRS. Allowing for structural rearrangements of the AsnRS backbone during the design might facilitate the insertion of such an interaction. This is beyond the scope of the present fixed-backbone design study but is worth exploring more systematically in the future.

Molecular dynamics calculations with selected sequences

Selected mutant sequences were studied further by MD simulations and PB free-energy calculations. We chose the native *Thermus thermophilus* AsnRS protein (QAEES in Table V), seven mutants determined by the absolute-affinity criterion (KSEES, KSTES, KSIES, HSEKS, HSTES, HSTMS, YSQKS), and three mutants determined by the relative-affinity criterion (KSVSS, KSVES, HSVMS). All 11 sequences were simulated in complex with the target ligand AspAMP. The native sequence and four mutants were also simulated in complex with the native ligand, AsnAMP. The simulations were conducted in explicit water and lasted 4 ns (see "Methodology" Section). A total of 1000 snapshots, extracted at 4-ps intervals, were used to compute the AspAMP or AsnAMP binding affinities by Poisson calculations. We did not include nonpolar contributions to the affinities; in our earlier work, we showed that these contributions were small, not very sensitive to the mutations, and within the PBFE uncertainty.[17]

The conformations of the various complexes are quite stable in the MD simulations. The rmsd from the starting conformation is 0.5–0.7 Å for mainchain and 1.2 Å for sidechain atoms in the inner (unrestrained) sphere of the complexes. In the AspAMP complexes of the Lys187-containing sequences, the active sites maintain the Asp ligand recognition mode of AspRS, shown in Figure 4(B). This is illustrated in Figures 7(C,D), which display the final MD conformations of the two most promising sequences from the absolute-affinity (KSEES) and relative-affinity design (KSVSS). The ligand AspAMP carboxylate retains two hydrogen-bonding interactions with Arg388 and one hydrogen bond with Lys187. Furthermore, in KSEES, the orientation of Arg388 is further stabilized by two interactions with Glu227. In the His187 complexes, the ligand carboxylate retains two interactions with Arg388 during the simulations but loses its interactions with the His187 sidechain (not shown).

The PBFE affinity results are summarized in Table V. The calculations predict that native AsnRS has a PBFE binding affinity of −25.7 kcal/mol for the native ligand (AsnAMP) and an AspAMP affinity that is weaker by 2 kcal/mol. With a zero ionic strength, the corresponding numbers are −26.0 and −29.8 kcal/mol, resulting in a relative affinity of 3.8 kcal/mol in favor of AsnAMP. The four AsnRS mutants simulated with bound AsnAMP have inverted binding affinities, favoring AspAMP by −3.3 kcal/mol to −10.5 kcal/mol. Three of these sequences (KSVSS, KSVES, HSVMS) were indeed determined by the stability/relative affinity criterion. Thus, our design was successful in decreasing Asn binding compared with Asp binding. At the same time, according to PBFE, none of the sequences binds Asp more strongly than the native sequence. This is consistent with the marginally better

affinity of the stability/absolute-affinity design sequences (Table III). What is more, the AspAMP affinities are smaller, by several kcal/mol, than the native affinity for AsnAMP.

Even though the PBFE analysis employed here is less accurate with respect to alchemical free energy calculations,[47,50,57,135] in the past we have employed it successfully to compute protein-ligand affinities in various systems, including the proteins AspRS and AsnRS.[45,47,49,136] The PBFE affinity estimates obtained here are consistent with the experimental activity measurements on selected designed sequences, as explained below.

EXPERIMENTAL RESULTS

To further test the success of our design, we measured the activity against L-asparagine and L-aspartate of wild-type AsnRS and several mutant sequences. Specifically, we chose the sequences KSTES, KSEES, KSIES, HSEKS, and HSTES (stability/absolute-affinity criterion) and KSVES, KSVSS (stability/relative-affinity criterion). Furthermore, we tested the most promising sequence, DKMMD, from our earlier CPD.[17] Sequence DKMMD was predicted earlier (with CASA) to have a native-like binding mode and an inverted affinity, favoring Asp by 11.8 kcal/mol.

The experiments failed to show a detectable Asp or Asn activity for any of the designed sequences. The initial rate of L-asparagine-dependent isotopic ATP-PP_i exchange was 0.83 s^{-1} for wild type AsnRS, and less than 1×10^{-2} s^{-1} for all the mutants. The corresponding initial rate of L-aspartate-dependent isotopic ATP-PP_i exchange was 6×10^{-3} s^{-1} for wildtype AsnRS. This last value is likely to reflect contamination of the experimental aspartate sample by asparagine. The initial rates of exchange were less than 5×10^{-3} s^{-1} for the mutants.

DISCUSSION AND CONCLUSIONS

In the present work, we have improved our earlier CPD methodology and used it in an attempt to engineer Asp specificity into the protein Aspartyl-tRNA synthetase. Compared with a previous CPD study of the same system,[17] our present work has two methodological innovations. In the previous study,[17] we computed protein interactions by a polar-hydrogen energy function [97] and solvent effects by a CASA approximation.[57,98] Here, we use an all-atom energy function[109] for the protein and ligand interactions, and a continuum electrostatics generalized Born model based on the GB-HCT formalism[110] for solvent effects. We use a parameterized version of the GB-HCT model, shown to yield accurate protein solvation free energies and free-energy changes due to mutations in fully or partly buried positions.[57] We deal with the many-body nature of the GB model in two steps. (i) We use a "residue-GB" approximation,[38,111] in which all atoms of a residue are assigned a common, "residue Born" radius, defined as a harmonic average over the Born radii of the individual atoms within the residue; (ii) We compute and tabulate prior to the design calculations the residue Born radii, assuming that the environment of each residue corresponds to the chemical structure and geometry of the native state. An analogous approximation has been used in conjunction with continuum electrostatics treatments in protein design.[65,68–70]

Second, we compared the performance of three optimization criteria, which target, respectively, the folding free energy (stability), the affinity for the noncognate ligand (AspAMP), or the affinity relative to the cognate ligand, AsnAMP. In contrast, our earlier study only used the stability criterion.

The all-atom/GB (this work) and polar-atom/CASA[17] free-energy functions yield different optimized AspRS sequences. We first compare the predictions of the stability criterion, which is common to both studies. Sequence alignment of AspRS and AsnRS active sites[48] shows that the five AspRS active-site residues corresponding to the designed AsnRS positions Gln187, Ala190, Glu225, Glu227, Ser366 are, respectively, Gln195, Lys198, Asp233, Glu235, and Ser487 (in *E. coli* numbering). Thus, the chemical identity is preserved at positions 187 (Gln), 227 (Glu), and 366 (Ser); the negative charge is retained at position 225, and a positive charge is inserted at position 190.

The majority of sequences designed with the polar-hydrogen/CASA free energy function have negatively charged residues (Asp, Glu) at positions 187 and 366[17]; position 190 either retains its native identity (Ala), or was set manually to be a Lys (in analogy with Lys198 in AspRS), and positions 225, 227 mostly become hydrophobic (Ala, Met) or polar (Ser, Asn). The sequences designed with the GB-HCT/all-atom energy function and the stability criterion (Table II) are more consistent with the properties of the AspRS active site. Position 366 remains invariant (Ser) and positions 225, 227 often retain a negatively charged residue (Glu). Position 187 can remain invariant (Gln), but mostly accepts an aromatic (Tyr) or non polar (Val) residue which disrupts the ligand carboxylate-Arg388 interactions.

With a combined stability/absolute-affinity criterion (Table III), position 187 is changed to a charged (Lys) or aromatic (His, Tyr) sidechain. The Lys insertion at this position constitutes a major improvement, compared with the stability/CASA prediction (mostly Asp or Glu): The Lys187 ammonium group forms an interaction with the ligand sidechain carboxylate, analogous to the Lys198-ligand interaction in the AspRS:Asp active site. The resulting conformations are very stable during the MD simulations, retaining the AspRS-like geometry [Fig. 4(B)]. In contrast, the CASA-derived sequences experienced significant distortions with respect to the native

AsnAMP complex in the MD runs of Ref. 17, due to repulsions between the ligand sidechain carboxylate and the negatively charged residue inserted at position 187 (and sometimes 366). The CASA design did not insert a lysine residue at position 187 spontaneously; rather, this Lys was inserted by hand.[17]

The good conformational stability and AspRS-like protein-ligand interactions of the simulated complexes indicate that the mutations introduced by the present model were physically reasonable. Stability optimization of the native complex AsnRS:Asn with the same five active positions produced sequences with aminoacid frequencies listed in Table 5 of the Supporting Material. With the exception of position 190 (100% serine), the native aminoacids are observed with the highest frequency in all other positions. Sequence alignment shows that the four active positions Q187, E225, E227, and S366 are conserved across AsnRS from various species; position 190 is more variable, containing methionine(45%), alanine(15%), glycine(12%), leucine(11%), and valine(6%). Thus, Ala190 is replaced by much bulkier residues (Leu, Val); presumably such an insertion is not favored by our design due to the fixed backbone. Our model predicts a serine at position 190 with 100% probability; at the same time, it predicts no serine at positions 187 and 225, suggesting that there is no consistent bias for this residue.

The Proteus design results with combined stability/affinity criteria (Tables III and IV) and the PBFE analysis of the trajectories (Table V) suggested that the obtained sequences had inverted specificities, but their Asp binding was not as strong as Asn binding by native AsnRS. This is consistent with the inability of the experimentally tested sequences to adenylate Asp or (Asn) with ATP. A pure absolute-affinity or relative-affinity criterion yielded sequences with strong Asp affinity; these sequences had decreased stability by 100 kcal/mol or more with respect to native AsnRS and were not considered further.

Several factors may have contributed to the failure to obtain sequences able to adenylate Asp. The computationally simpler approximation to the residue-GB/HCT model is less accurate than the original atomic- or residue-GB/HCT model; the heuristic algorithm searches a small subset of the conformational space: side-chains are placed into a small set of rotamer states[113] and the protein backbone is retained into the conformation of native AsnRS. Using a much more extended rotamer library,[137] and/or introducing backbone flexibility[8] could yield additional sequences, missed by the present design. Finally, the inactivity of the mutant proteins could be due to a disruption of transition state stabilization. It could also be that structural rearrangements due to the design interfere with ATP binding, necessary for the initial adenylation reaction to occur. Further testing is needed to determine the importance of these factors.

In conclusion, we have implemented and tested a CPD methodology in which solvent effects are modelled with the generalized-Born approximation and sequences are selected according to both stability and affinity. Using this methodology, we have engineered Asp specificity into AsnRS. The engineered AsnRS active sites have some of the structural features seen in the Asp-specific protein Aspartyl-tRNA synthetase; their conformations and interactions are maintained in MD simulations. Compared with the earlier CPD study,[17] which treated solvent effects by a Coulomb/Accesssible Surface Area approxination, the present, GB treatment appears promising.

REFERENCES

1. Floudas CA, Fung HK, Mcallister SR, Monnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: a review. Chem Engin Sci 2006;61:966–988.
2. Lippow SM, Tidor B. Progress in computational protein design. Curr Opin Biotechn 2007;18:305–311.
3. Boas FE, Harbury PB. Potential energy functions for protein design. Curr Opin Struct Biol 2007;17:199–204.
4. Das R, Baker D. Macromolecular modelling with Rosetta. Ann Rev Biochem 2008;77:363–382.
5. Schmidt am Busch M, Mignon D, Simonson T. Computational protein design as a tool for fold recognition. Proteins 2009;77:139–158.
6. Karanicolas J, Kuhlman B. Computational design of affinity and specificity at protein-protein interfaces. Curr Opin Struct Biol 2009;19:458–463.
7. Dambrovsky J, Brezovsky J. Computational tools for designing and engineering biocatalysts. Curr Opin Chem Biol 2009;13:26–34.
8. Mandell DJ, Kortemme T. Backbone flexibility in computational protein design. Curr Opin Biotechn 2009;20:420–428.
9. Suarez M, Jaramillo A. Challenges in the computational design of proteins. J R Soc Interface 2009;6:5477–5491.
10. Mashiach E, Nussinov R, Wolfson HJ. Flexible induced-fit backbone refinement in molecular docking. Proteins 2009;78:1503–1519.
11. Schmidt am Busch M, Sedano A, Simonson T. Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. PloS One 2010;5:e10410.
12. Saven JG. Computational protein design: Advances in the design and redesign of biomolecular nanostructures. Curr Opin Colloid Interf Sci 2010;15:13–17.
13. Bellows ML, Floudas CA. Computational methods for de novo protein design and its applications to the Human Immunodeficiency Virus 1, Purine Nucleoside Phosphorylase, Ubiquitin Specific Protease 7, and Histone Demethylases. Current drug targets 2010;11:264–278.
14. Bolon DN, Grant RA, Baker TA, Sauer RT. Specificity versus stability in computational protein design. Proc Natl Acad Sci USA 2005;102:12724–12729.
15. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Jr, Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 2006;441:655–659.
16. Green DF, Dennis AT, Fam PS, Tidor B, Jasanoff A. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. Biochemistry 2006;45:12547–12559.
17. Lopes A, Schmidt am Busch M, Simonson T. Computational design of protein-ligand binding: modifying the specificity of asparaginyl-tRNA synthetase. J Comp Chem 2010;31:1273–1286.
18. Klepeis JL, Floudas CA, Morikis D, Tsokos CG, Argyropoulos E, Spruce L, Lambris JD. Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity. J Am Chem Soc 2003;125:8422–8423.
19. Shifman JM, Mayo SL. Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc Natl Acad Sci USA 2003;100:13274–13279.

20. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond *in vivo* maturation. Nat Biotechnol 2007;25:1171–1176.
21. Altman MD, Nalivaika EA, Prabu-Jeyabalan M, Schiffer CA, Tidor B. Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. Proteins 2008;70:678–694.
22. Reynolds KA, Hanes MS, Thomson JM, Antczak AJ, Berger JM, Bonomo RA, Kirsch JF, Handel TM. Computational redesign of the SHV-1 beta-lactamase/beta-lactamase inhibitor protein interface. J Mol Biol 2008;382:1265–1275.
23. Haidar JN, Pierce B, Yu Y, Tong WW, Li M, Weng ZP. Structure-based design of a T-cell receptor leads to nearly 100-fold improvement in binding affinity for pepMHC. Proteins 2009;74: 948–960.
24. Bellows ML, Fung HK, Taylor MS, Floudas CA, de Victoria AL, Morikis D. New compstatin variants through two de novo protein design frameworks. Biophys J 2010;98:2337–2346.
25. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. Science 2005;308:857–860.
26. Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL. Full-sequence computational design and solution structure of a thermostable protein variant. J Mol Biol 2007;372:1–6.
27. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
28. Ambroggio XI, Kuhlman B. Computational design of a single aminoacid sequence that can switch between two distinct protein folds. J Am Chem Soc 2006;128:1154–1161.
29. Calhoun JR, Kono H, Lahr S, Wang W, deGrado WF, Saven JG. Computational design and characterization of a monomeric helical dinuclear metalloprotein. J Mol Biol 2003;334:1101–1115.
30. Calhoun JR, Liu W, Spiegel K, Peraro MD, Klein ML, Valentine KG, Wand JA, deGrado WF. Solution NMR structure of a designed metalloprotein and complementary molecular dynamics refinement. Structure 2008;16:210–215.
31. Tynan-Conolly BM, Nielsen JE. Redesigning protein pK_a values. Prot Sci 2007;16:239–249.
32. Röthlisberger D, Khersonsky O, Wollacott AM, Ziang L, Dechncie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature 2008;453:190–195.
33. Chaloupkova R, Sykorova J, Prokop Z, Jesenska A, Monincova M, Pavlova M, Tsuda M, Nagata Y, Dambrovsky J. Modification of activity and specificity of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26 by engineering of its entrance channel. J Biol Chem 2003;278:52622–52628.
34. Slovic AM, Kono H, Lear JD, Saven JG, deGrado WF. Computational design of water-soluble analogues of the potassium channel KcsA. Proc Natl Acad Sci USA 2004;101:1828–1833.
35. Joachimiak LA, Kortemme T, Stoddard BL, Baker D. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J Mol Biol 2006;361:195–208.
36. Grigoryan G, Reinke AW, Keating AE. Design of protein interaction specificity gives selective bZIP-binding peptides. Nature 2009; 458:859–864.
37. der Sloot AMV, Kiel C, Serrano L, Stricher F. Protein design in biological networks: from manipulating the input to modifying the output. Prot Eng Des Sel 2009;22:537–542.
38. Aleksandrov A, Polydorides S, Archontis G, Simonson T. Predicting the acid/base behavior of proteins: a constant-pH Monte Carlo approach with generalized Born solvent. J Phys Chem B 2010;114: 10634–10648.
39. Schaefer M, Vlijmen Hv, Karplus M. Electrostatic contributions to molecular free energies in solution. Adv Prot Chem 1998;51: 1–57.
40. Warshel A, Parson W. Dynamics of biochemical and biophysical reactions: insight from computer simulations. Q Rev Biophys 2001;34:563–679.
41. Simonson T. Electrostatics and dynamics of proteins. Rep Prog Phys 2003;66:737–787.
42. Koehl P. Electrostatic calculations: latest methodological advances. Curr Opin Struct Biol 2005;16:142–151.
43. Roca M, Vardi-Kilshtain A, Warshel A. Toward accurate screening in computer-aided enzyme design. Biochemistry 2009;48:3046–3056.
44. Ofran Y, Rost B. Analyzing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.
45. Thompson D, Plateau P, Simonson T. Free energy simulations reveal long-range electrostatic interactions and substrate-assisted specificity in an aminoacyl-tRNA synthetase. ChemBioChem 2006; 7:337–344.
46. Thompson D, Simonson T. Molecular dynamics simulations show that bound Mg^{2+} contributes to amino acid aminoacyl adenylate binding specificity in aspartyl-tRNA synthetase through long range electrostatic interactions. J Biol Chem 2006;281:23792–23803.
47. Aleksandrov A, Thompson D, Simonson T. Alchemical free energy simulations for biological complexes: powerful but temperamental. . .. J Mol Rec 2010;23:117–127.
48. Archontis G, Simonson T, Moras D, Karplus M. Specific amino acid recognition by Aspartyl-tRNA synthetase studied by free energy simulations. J Mol Biol 1998;275:823–846.
49. Archontis G, Simonson T, Karplus M. Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. J Mol Biol 2001;306:307–327.
50. Simonson T, Archontis G, Karplus M. Free-energy Simulations come of age: protein-ligand recognition. Acc Chem Res 2002;35: 430–437.
51. Jiang L, Kuhlman B, Kortemme TA, Baker DA. 'solvated rotamer' approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. Proteins 2005;58:893–904.
52. Roux B, Simonson T. Implicit solvent models. Biophys Chem 1999;78:1–20.
53. Simonson T. Macromolecular electrostatics: continuum models and their growing pains. Curr Opin Struct Biol 2001;11:243–252.
54. Eisenberg D, McClachlan A. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
55. Ooi T, Oobatake M, Nemethy G, Scheraga H. Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
56. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1: 227–235.
57. Lopes A, Alexandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins 2007;67:853–867.
58. Dahiyat B, Mayo S. Protein design automation. Protein Sci 1996;5: 895–903.
59. Ferrara P, Apostolakis J, Caflisch A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins 2002; 46:24–33.
60. Ogata K, Jaramillo A, Cohen W, Briand J, Conana F, Wodak S. Automatic sequence design of MHC class-I binding peptides impairing CD8+ T-cell recognition. J Biol Chem 2003;278:1281–1290.
61. Davis ME, McCammon JA. Electrostatics in biomolecular structure and dynamics. Chem Rev 1990;90:509–521.
62. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
63. Gohlke H, Kiel C, Case D. Insight into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. J Mol Biol 2003;330:891–913.
64. David L, Luo R, Gilson M. Comparison of generalized Born and Poisson models: energetics and dynamics of HIV protease. J Comp Chem 2000;21:295–309.

65. Pokala N, Handel TM. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Protien Sci 2004;13:925–936.
66. Koehl P, Levitt M. De novo protein design. I. In search of stability and specificity. J Mol Biol 1999;293:1161–.
67. Wernisch L, Henry S, Wodak S. Automatic protein design with all-atom force-fields by exact and heuristic optimization. J Mol Biol 2001;301:713–736.
68. Marshall SA, Vizcarra CL, Mayo SL. One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. Protein Sci 2005;14:1293–1304.
69. Vizcarra CL, Mayo SL. Electrostatics in computational protein design. Curr Opin Chem Biol 2005;9:622–626.
70. Vizcarra CL, Zhang N, Marshall SA, Wingreen NS, Zeng C, Mayo SL. An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design. J Comp Chem 2008;29:1153–1162.
71. Still WC, Tempczyk A, Hawley R, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
72. Bashford D, Case D. Generalized Born models of macromolecular solvation effects. Ann Rev Phys Chem 2000;51:129–152.
73. Feig M, Brooks CL, III. Recent advances in the development and application of implicit solvent models in biomolecular simulations. Curr Opin Struct Biol 2004;14:217–224.
74. Chen J, Brooks CL, III, Khandogin J. Recent advances in implicit-solvent based methods for biomolecular simulations. Curr Opin Struct Biol 2008;18:140–148.
75. Schaefer M, Karplus M. A comprehensive analytical treatment of continuum electrostatics. J Phys Chem 1996;100:1578–1599.
76. Qiu D, Shenkin P, Hollinger F, Still W. A fast analytical method for hte calculation of approximate Born radii. J Phys Chem A 1997;101:3005–3014.
77. Gosh A, Rapp C, Friesner RA. Generalized Born model based on a surface-area formulation. J Phys Chem B 1998;102:10983–10990.
78. Wagner F, Simonson T. Implicit solvent models: combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect. J Comp Chem 1999;20:322–335.
79. Tamamis P, Kasotakis E, Mitraki A, Archontis G. Amyloid-like self-assembly of peptide sequences from the adenovirus fiber shaft: insights from molecular dynamics simulations. J Phys Chem B 2009;113:15639–15647.
80. Tamamis P, Adler-Abramovich L, Reiches M, Marshall K, Sikorski P, Serpell L, Gazit E, Archontis G. Self-assembly of phenylalanine oligopeptides: insights from experiments and simulations. Biophys J 2009;96:5020–5029.
81. Onufriev A, Bashford D, Case D. Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 2000;104:3712–3720.
82. Lee M, Salsbury F, Brooks CL, III. Constant pH molecular dynamics using continuous titration coordinates. Proteins 2004;56:738–752.
83. Mongan J, Case DA, McCammon JA. Constant pH molecular dynamics in generalized Born implicit solvent. J Comp Chem 2004; 25:2038–2048.
84. Khandogin J, Brooks CL, III. Toward the accurate first-principles prediction of ionization equilibria in proteins. Biochemistry 2006; 45:9363–9373.
85. Calimet N, Schaefer M, Simonson T. Protein molecular dynamics with the Generalized Born/ACE solvent model. Proteins 2001;45:144–158.
86. Simmerling C, Strockbine B, Roitberg A. All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 2002;124:11258–11259.
87. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Exhaustive docking of molecular fragments with electrostatic solvation. Proteins 1999;38:88–105.
88. Liu HY, Zou X. Electrostatics of ligand binding: parameterization of the generalized Born model and comparison with the Poisson-Boltzmann approach. J Phys Chem B 2006;110:9304–9313.
89. Simmerling C, Strockbine B, Roitberg A. All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 2002;124:11258–11259.
90. Chen J, Im W, Brooks CL, III. Balancing solvation and intramolecular interactions: towards a consistent generalized Born force field. J Am Chem Soc 2006;128:3728–3736.
91. Jang S, Kim E, Pak Y. Direct folding simulation of alpha-helices and beta-hairpins based on a single all-atom force field with an implicit solvation model. Proteins 2007;66:53–60.
92. Lei H, Wu C, Liu H, Duan Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc Natl Acad Sci USA 2007;104:4925–4930.
93. Moulinier L, Case DA, Simonson T. Xray structure refinement of proteins with the generalized Born solvent model. Acta Cryst D 2003;59:2094–2103.
94. Chen J, Im W, Brooks CL, III. Refinement of NMR structures using implicit solvent and advanced sampling techniques. J Am Chem Soc 2004;126:16038–16047.
95. Lee MS, Olson MA. Assessment of detection and refinement strategies for de novo protein structures using force field and statistical potentials. J Chem Theory Comput 2007;3:312–324.
96. Boas FE, Harbury PB. Design of protein-ligand binding based on the molecular-mechanics energy model. J Mol Biol 2008;380:415–424.
97. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
98. am Buch MS, Lopes A, Mignon D, Simonson T. Computational protein design: software implementation, parameter optimization and performance of a simple model. J Comp Chem 2008;29:1092–1102.
99. am Buch MS, Lopes A, Amara N, Bathelt C, Simonson T. Testing the Coulomb/Accessible Surface Area solvent model for protein stability, ligand binding and protein design. BMC Bioinformatics 2008;9:148–163.
100. Cusack S. Eleven down and nine to go. Nat Struct Biol 1995;2: 824–831.
101. Arnez J, Moras D. Structural and functional consideration of the aminoacylation reaction. Trends Biochem Sci 1997;22:211–216.
102. Meinnel T, Mechulam Y, Blanquet S. Aminoacyl-tRNA synthetases: occurrence, structure and function. In tRNA: structure, biosynthesis and function, ASM Press: Washington, D.C.; 1995.
103. de Prat Gay G, Duckworth H, Fersht A. Modification of the amino acid specificity of tyr-tRNA synthetase by protein engineering. FEBS Lett 1993;318:167–171.
104. Agou F, Quevillon S, Kerjan P, Mirande M. Switching the amino acid specificity of an aminoacyl-tRNA synthetase. Biochemistry 1998;37:11309–11314.
105. Hendrickson TL, de Crecy-Lagard V, Schimmel P. Incorporation of nonnatural amino acids into proteins. Ann Rev Biochem 2004;73: 147–176.
106. Xie J, Schultz P. Adding amino acids to the genetic repertoire. Curr Opin Chem Biol 2005;9:548–554.
107. Thompson D, Lazennec C, Plateau P, Simonson T. Probing electrostatic interactions and ligand binding in aspartyl-tRNA synthetase through site-directed mutagenesis and computer simulations. Proteins 2008;71:1450–1460.
108. Berthet-Colominas C, Seignovert L, Hartlein M, Grotli M, Cusack S. The crystal structure of asparaginyl-tRNA synthetase from *thermus thermophilus* and its complexes with ATP and asparaginyl-adenylate: the mechanism of discrimination between asparagine and aspartic acid. EMBO J 1998;17:2947–2960.
109. Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc 1995;117:5179–5197.
110. Hawkins G, Cramer C, Truhlar D. Pairwise descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.

111. Archontis G, Simonson T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. J Phys Chem B 2005;109:22667–22673.
112. Eiler S, Dock-Bregeon A, Moulinier L, Thierry J, Moras D. Synthesis of aspartyl-tRNA(asp) in *Escherichia coli*: a snapshot of the second step. EMBO J 1999;18:6532–6541.
113. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267.
114. Brünger AT. X-PLOR version 3.1, A system for X-ray crystallography and NMR. New Haven: Yale University Press; 1992.
115. Cavarelli J, Eriani G, Rees B, Ruff M, Boeglin M, Mitschler A, Martin F, Gangloff J, Thierry J, Moras D.. The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. EMBO J 1994;13:327–337.
116. Ho CK, Fersht AR. Internal thermodynamics of position 51 mutants and natural variants of tyrosyl-tRNA synthetase. Biochemistry 1986;25:1891–1897.
117. Wells TN, Fersht AR. Use of binding energy in catalysis analyzed by mutagenesis of the tyrosyl-tRNA synthetase. Biochemistry 1986; 25:1881–1886.
118. Fersht AR, Leatherbarrow RJ, Wells TN. Structure-reactivity relationships in engineered proteins:analysis of use of binding energy by linear free energy relationships. Biochemistry 1987;26:6030–6038.
119. de Prat Gay G, Duckworth HW, Fersht AR. Modification of the amino acid specificity of tyrosyl-tRNA synthetase by protein engineering. FEBS Letters 1993;318:167–171.
120. First EA, Fersht AR. Mutational and kinetic analysis of a mobile loop in tyrosyl-tRNA synthetase. Biochemistry 1993;32:13658–13663.
121. Mackerell Jr.,AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FKT, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III, WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkewicz-Kuczera J, Yin D, Karplus M. An all-atom empirical potential for molecular modelling and dynamics study of proteins. J Phys Chem B 1998; 102:3586–3616.
122. Arnez J, Flanagan K, Moras D, Simonson T. Engineering a Mg^{2+} site to replace a structurally conserved arginine in the catalytic center of histidyl-tRNA synthetase by computer experiments. Proteins 1998;32:362–380.
123. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926–35.
124. Darden T, York D, Pedersen L. Particle Mesh Ewald: an N log(N) method for Ewald sums in large systems. J Chem Phys 1993;98: 10089–10092.
125. Nose S. A unified formulation of the constant temperature molecular dynamics method. J Chem Phys 1984;81:511–519.
126. Hoover W. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A 1985;31:1695–1697.
127. Feller S, Zhang Y, Pastor RW, Brooks B. Constant-pressure molecular-dynamics simulation: the Langevin piston method. J Chem Phys 1995;103:4613–4621.
128. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23: 327–341.
129. Brooks BR, Brooks CL, III, Mackerell AD, Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor R, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. J Comp Chem 2009;30: 1545–1614.
130. Im W, Beglov D, Roux B. Continuum solvation model: a computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. Comp Phys Comm 1998;111:59–75.
131. Humphrey W, Dalke A, Schulten K. Visual molecular dynamics. J Molec Graphics 1996;14:33–38.
132. Myers JK, Pace CN, Scholtz JM. Helix propensities are identical in proteins and peptides. Biochemistry 1997;36:10923–10929.
133. Yang J, Spek EJ, Gong Y, Zhou H, Kallenbach NR. The role of context on alpha-helix stabilization: host-guest analysis in a mixed background peptide model. Protein Sci 1997;6:1–9.
134. Park SH, Shalongo W, Stellwagen E. Residue helix parameters obtained from dichroic analysis of peptides of defined sequence. Biochemistry 1993;32:7048–7053.
135. Gilson MK, Zhou HX. Calculation of protein-ligand binding affinities. Ann Rev Biophys Biomol Struct 2007;36:21–42.
136. Polydoridis S, Leonidas DD, Oikonomakos NG, Archontis G. Recognition of Ribonuclease A by 3′–5′-Pyrophosphate-linked dinucleotide inhibitors: a molecular dynamics/continuum electrostatics analysis. Biophys J 2007;92:1659–1672.
137. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 2004;13:735–751.

Appendix C

Predicting the Acid/Base Behavior of Proteins

Predicting the Acid/Base Behavior of Proteins: A Constant-pH Monte Carlo Approach with Generalized Born Solvent

Alexey Aleksandrov,[†] Savvas Polydorides,[‡] Georgios Archontis,*[,‡] and Thomas Simonson*[,†]

Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France, and Department of Physics, University of Cyprus, PO20537, CY1678, Nicosia, Cyprus

Received: May 14, 2010; Revised Manuscript Received: July 7, 2010

The acid/base properties of proteins are essential in biochemistry, and proton binding is a valuable reporter on electrostatic interactions. We propose a constant-pH Monte Carlo strategy to compute protonation free energies and pK_a's. The solvent is described implicitly, through a generalized Born model. The electronic polarizability and backbone motions of the protein are included through the protein dielectric constant. Side chain motions are described explicitly, by the Monte Carlo scheme. An efficient computational algorithm is described, which allows us to treat the fluctuating shape of the protein/solvent boundary in a way that is numerically exact (within the GB framework); this contrasts with several previous constant-pH approaches. For a test set of six proteins and 78 titratable groups, the model performs well, with an rms error of 1.2 pH units. While this is slightly greater than a simple Null model (rms error of 1.1) and a fully empirical model (rms error of 0.9), it is obtained using physically meaningful model parameters, including a low protein dielectric of four. Importantly, similar performance is obtained for side chains with large and small pK_a shifts (relative to a standard model compound). The titration curve slopes and the conformations sampled are reasonable. Several directions to improve the method further are discussed.

1. Introduction

The acid/base properties of proteins are essential in biochemistry and have been studied for almost a century.[1−3] Proton transfer is central to both respiration and photosynthesis;[4] many enzyme reactions include proton binding steps,[5] while protein folding, supramolecular assembly, adsorption, and binding to ligands are all sensitive to pH.[6,7] In addition, protons can be valuable reporters on electrostatic interactions and dielectric relaxation,[8−10] which have even broader significance.[9−13] Experimental methods, especially high-resolution structure determination, have opened the way to a thorough understanding of acid/base biochemistry and electrostatic interactions in general. However, the complexity of proteins and their aqueous environment are such that biophysical models are almost always needed to complement and interpret the experiments: X-ray structures do not reveal electric fields.

Simple, empirical models have been developed in recent years that often give good accuracy (around one pK_a unit) and can help quantify the main effects.[14,15] For example, the popular PROPKA program[14] takes into account hydrogen bonds involving the titratable side chains, the extent of solvation/desolvation of these side chains, and their proximity to ionized groups. However, detailed physical theories obviously provide a deeper understanding. For example, Hush and Marcus used a two-step thought experiment to introduce a new charge, such as a proton or electron, allowing them to isolate the free energy contributions of electronic and dipolar reorganization.[16−19] Similarly, dielectric continuum theory quantifies the desolvation of a charge through its "self-energy",[20−23] which has a simple relation to the reorganization free energy for introducing the charge.[24] Thus, considerable efforts have been made to develop physically meaningful simulation models that give reasonable accuracy for pK_a's and can at the same time increase our understanding of protein electrostatics.

The most detailed and accurate models use an atomistic treatment of both the protein and the surrounding solvent, combined with molecular dynamics (MD) or Monte Carlo (MC) sampling of conformational space.[25−28] Fully quantum mechanical treatments are still not practical for most problems, but mixed quantum/classical treatments are common for studying enzyme reactions.[27,28,13,29] While fully atomistic models have been applied to proton binding and transfer,[30−36] there are many problems where they cannot be routinely used. In particular, when a protein is studied over a broad pH range, where many groups can bind or release protons, additional approximations are necessary. The oldest and simplest approach treats the protein and its surroundings as two homogeneous, isotropic, dielectric media and solves the Poisson−Boltzmann equation to obtain free energies, sometimes analytically,[20,37,38] but usually numerically.[39,23,40,41] The simple, two-dielectric approach has very severe limitations, discussed in the literature and in the Discussion. More recent models use a hybrid approach, where part of the system and/or some of its degrees of freedom are treated explicitly and atomistically, while other parts or degrees of freedom are "integrated out" and treated implicitly.[42] The implicit treatment often relies on a dielectric continuum model.[20,39,23,43,41,42,44,45,10]

Two of the most important, recent, hybrid approaches are the "multiconformation" Poisson−Boltzmann, or MCPB, methods[46−49] and the so-called "constant-pH" MD methods.[50−53] With MCPB, the protein and solvent are treated as two dielectric media, while protein side chain conformations are explored with Monte Carlo. The protein backbone is held fixed, and its conformational flexibility is absorbed into the protein dielectric constant.[54] With constant-pH MD, the protein and its motions

* Corresponding author. E-mail: thomas.simonson@polytechnique.fr; archonti@ucy.ac.cy.
[†] Ecole Polytechnique.
[‡] University of Cyprus.

10.1021/jp104406x
Published on Web 07/26/2010

are described in full atomistic detail, while the solvent is modeled as a dielectric continuum. The motions are sampled by MD simulation, and proton binding/release is treated through an extended ensemble that mimics a semigrand canonical ensemble (see Theory section). The atomic charges in the protein usually have fixed magnitudes, which implies that the electronic polarizability of the protein is absorbed into the protein dielectric constant. Since MD simulations with a PB solvent are rather expensive,[43,55,56] constant-pH MD is often done with a generalized Born (GB) solvent model,[57−62] which contains similar physics but is more efficient.[52,53]

Here, we present a novel hybrid approach. As with MCPB, we use a Monte Carlo method to explore the conformations of all protein side chains, and we hold the protein backbone fixed; its motions are treated implicitly, through the protein dielectric constant. However, our method has two distinctive features. First, we use a generalized Born solvent, as opposed to the PB solvent used with MCPB.[46−49] Second, we accurately account for changes in the shape of the protein volume (and the remaining solvent volume), which occur constantly as a result of the protein side chain fluctuations. Until now, only the constant-pH approach allowed an exact description of these volume changes. All other hybrid approaches require an additional approximation; usually, an average protein−solvent boundary is used, possibly with additional, ad hoc corrections.[46−49,63−65] Even though these corrections appear quite successful in several published applications,[63−65] it is of great interest to derive an essentially exact scheme and eliminate one source of empiricism from the model. This is done here in an efficient way by use of a recent, "residue-GB" variant, which has a "pairwise decomposable" property:[66] side chain−side chain and side chain−backbone interactions can be precomputed and tabulated, following a method introduced for computational protein design.[66,67] We show in the Theory section how this is possible without loss of accuracy, even though the generalized Born interaction energy is a many-body function.[57−59]

The method is tested on six small proteins, including a total of 78 titratable groups. The accuracy is reasonable: only slightly poorer than the best empirical and "physics-based" methods and comparable to several other physics-based methods. The best results are obtained with a model parametrization that is physically reasonable. The performance is comparable for side chains with unshifted and with strongly shifted pK_a values (shifted compared to standard model compounds). There are several clear directions to improve the method systematically. It is straightforward to extend it to compute redox potentials and to treat processes where proton and electron binding are coupled. It should also be possible to apply the method with a PB solvent instead of a GB solvent.

The article is organized as follows. In the next, Theory, section, we present the semigrand canonical ensemble, the classical mechanical treatment of proton binding, the residue-GB model, and the algorithm that reduces the computational complexity to a quadratic dependence on protein size. Computational details are described next. In the Results, we begin by presenting a systematic comparison between the residue-GB model and a standard PB model. This is appropriate since this work is the first application of residue-GB. Next, we present pK_a calculations for our test set of proteins. The last section is a Discussion. In particular, we discuss the limitations of continuum models, and the meaning of the protein dielectric constant in the context of hybrid models, where some of the relaxation channels are modeled explicitly and others are modeled implicitly.

2. Theory: Constant-pH Monte Carlo with Generalized Born Solvent

2.1. Statistical Mechanical Framework. Several authors have described the framework for constant-pH simulations using MD or MC.[50−53,48] A similar framework has been described and used for simulations with variable numbers of water molecules.[68−70] Following Baptista et al.,[50] we consider a dilute solution of the protein of interest, with a constant volume V and temperature T, closed with respect to protein and water but open with respect to protons. The corresponding thermodynamic ensemble is a slight variation of the grand canonical ensemble;[71] the partition function has the form (e.g., see eq 1-60 in ref 71)

$$\Xi(V,T,\mu) = \sum_{j,N} \exp(-\beta E_j(N,V))\exp(\beta N\mu) \qquad (1)$$

where μ is the chemical potential of the proton; the sum is over the number of protons N and the different states j of the system, which are characterized by their energies E_j; $\beta = 1/kT$; and k is Boltzmann's constant. For convenience, we have assumed the energy states are discrete; this would be the case for a finite quantum mechanical system, and it will also be the case in the applications below (since we will use a fixed-backbone, side chain rotamer description of the protein's conformational space, along with an implicit solvent model). The statistical probability $P_j(N; V, T, \mu)$ of a particular state can be written

$$P_j(N;V,T,\mu) = \exp(-\beta E_j(N,V))\exp(\beta N\mu)/\Xi(V,T,\mu) \qquad (2)$$

We will not consider the solvent explicitly; rather, we integrate over the solvent degrees of freedom.[72,42,73] The probability function keeps the same form, but we must replace the energy E_j by a potential of mean force, or PMF, denoted W_J[42]

$$P_J(N;V,T,\mu) = \frac{\exp(-\beta W_J(N;V))\exp(\beta N\mu)}{\sum_{J,N} \exp(-\beta W_J(N;V))\exp(\beta N\mu)} \qquad (3)$$

The index J now refers to an energy state of the protein alone (bathed in implicit solvent). Below, to approximate the PMF, we will use a continuum dielectric treatment, specifically, a generalized Born variant.[57,42,61,44] For the Monte Carlo simulations, below, we do not need to sample the protein velocities; therefore, we also integrate them out. This does not change the form of the probability $P_J(N; V, T, \mu)$; however, J now refers to a *conformational* state of the protein. Notice that J also identifies the specific location of all the protons bound to the protein. If we compare two states that differ by the addition of a proton to a specific titratable side chain, with the protein in a given conformational state J, the ratio of Boltzmann probabilities has the form

$$\frac{P_J(N+1)}{P_J(N)} = e^{-\beta(W_J(N+1)-W_J(N))}e^{\beta\mu} = e^{-\beta(W_J(N+1)-W_J(N))-2.303\mathrm{pH}+\beta\mu^0} \qquad (4)$$

The last equality uses the relation between μ and the pH: $\beta\mu = \beta\mu^0 - 2.303\mathrm{pH}$, where μ^0 is the proton chemical potential in aqueous solution in the standard state. For simplicity, in eq 4,

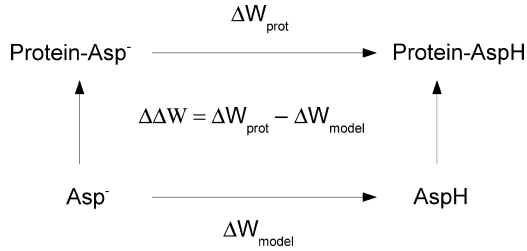


Figure 1. Thermodynamic cycle to analyze pK_a shifts: example of an Asp side chain. The upper leg represents proton binding to an Asp side chain in a protein. The lower leg represents proton binding to a model compound in solution. The double PMF difference between the two legs is $\Delta\Delta W = \Delta W_{prot} - \Delta W_{model}$. All four states are assumed to be at standard state concentrations (ideal solutions with 1 M concentrations).

we have kept the same index J for the two states, even though the $N + 1$ state has an extra proton; this slightly abusive notation helps emphasize that the protein conformation is otherwise unchanged, and only a proton has been added to a particular location. The Boltzmann probability distribution in eqs 3 and 4 will be sampled using the standard, Metropolis, Monte Carlo algorithm[74,75] (see below).

2.2. Classical Mechanical Treatment of Proton Binding. Proton binding is described within a classical mechanical, molecular mechanics framework, first proposed by Warshel and co-workers,[30] and abundantly used ever since.[23,40,41,76,33] The titrating protons, like the other atoms in the system, are treated as classical mechanical particles, bearing a partial charge and interacting with the other atoms through Coulombic terms, stereochemical terms, and van der Waals terms. The force field used here[77] employs fixed partial charges; i.e., electronic polarizability is treated in a simple, mean-field way.

This approach would not be appropriate to compute absolute protonation free energies; rather, one uses it to compute the difference between the PMF change for protonation of a protein side chain and that of a suitable model compound in solution.[30] The model compound is typically an analogue of the corresponding side chain; for example, aspartate with neutral backbone blocking groups (2*N*-acetyl-1*N*-methylaspartic acid-1-amide) could serve as a model compound for Asp side chain protonation (see below). We consider the thermodynamic cycle in Figure 1: subtracting the PMF change for the upper and lower legs yields a double PMF difference, $\Delta\Delta W$. We expect that taking this difference leads to a cancellation of most of the error associated with the simple, molecular mechanical treatment employed.[30,23,40,41,76,33] To obtain an absolute PMF change for protonation/deprotonation in the protein, $W_J(N + 1) - W_J(N)$ (eq 4), we must then add back the contribution of the model compound in solution. This is straightforward since the model compound's standard protonation free energy ΔG_{model} at a given pH has a simple relation to its pK_a and to the PMF change for the model compound

$$\beta\Delta G_{model} = \beta\Delta W_{model} - \beta\mu^0 = -2.303\mathrm{p}K_{a,model} \quad (5)$$

where $pK_{a,model}$ is the pK_a of the model compound in aqueous solution. The μ^0 term appears because with our definition of the PMF ΔW_{model} does not include any contributions from the solvated proton. In eq 5, we assume the states linked by the thermodynamic cycle of Figure 1 correspond to standard state concentrations.

2.3. Reducing the Computational Complexity: Residue-GB. To make the Monte Carlo calculations efficient, we use a strategy borrowed from the field of computational protein design.[67,78,79] The interaction energies are computed ahead of time for all side chain pairs in the protein, allowing for all possible rotamers and protonation states. The Monte Carlo exploration of rotamers and protonation states can then be done very efficiently, with the interaction energies obtained from lookup tables.

At first glance, this strategy appears impossible with a continuum dielectric solvent model. Indeed, in continuum electrostatics, the effective interaction between two residues depends on the entire protein's shape and the complementary volume occupied by high dielectric solvent.[22] Therefore, continuum electrostatic energies are many-body quantities that cannot ordinarily be expressed as a sum over residue or atom pairs.[10,63] Here, we overcome this difficulty thanks to a novel generalized Born model that is residue-pairwise and can be used efficiently for constant pH simulations, as well as for protein design.[66] Two steps make the scheme pairwise. (i) First, we adopt an expression for the interaction energy between two residues R and R′ that depends on the product $B = B_R B_{R'}$ of their *residue* Born solvation radii. These radii reflect the desolvation, or burial, within the protein of each residue. With most GB models, they are readily obtained from residue-pairwise quantities. (ii) Second, we fit the RR′ interaction energy by a simple function of B; the fitting coefficients depend only on the pair RR′, *not* on its environment. In effect, the quantity B captures all the information that is relevant about the pair's dielectric environment. The numerical accuracy of the fitting scheme can be made arbitrarily high; the fitting scheme chosen below can be considered essentially exact.[66] Below, we describe steps (i) and (ii) in detail.

2.3.1. Step (i): Residue Generalized Born. With GB, the electrostatic energy includes both a direct, Coulomb term and a contribution from the solvent, polarized by the solute charges. Treating the solvent as a linear, homogeneous, dielectric medium, the total electrostatic energy has the form

$$\begin{aligned} E^{elec} &= E^{Coul} + \Delta G^{solv} \\ &= \frac{1}{2}\sum_{i\neq j}\frac{q_i q_j}{\epsilon_p r_{ij}} + \frac{1}{2}\sum_{ij} g_{ij} \end{aligned} \quad (6)$$

where the sums are over all pairs of protein charges and the second sum includes diagonal terms, $i = j$. In eq 6, ε_p is the protein dielectric constant, which is usually set to one in constant-pH studies but will be set to 4 or 8 in the application below (to account for the backbone and electronic degrees of freedom). The second sum, ΔG^{solv}, represents the electrostatic solvation free energy of the protein (in the given conformation).[42] The term g_{ij} represents the interaction between a protein charge q_i and the solvent polarization induced by another charge, q_j. We refer to it as a GB interaction or screening energy. In the standard, Atomic GB model,[57] this term is approximated by

$$g_{ij} = g(\underline{r}_i, \underline{r}_j) = \frac{\tau q_i q_j}{(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j])^{1/2}} \quad (7)$$

where $r_{ij} = |\underline{r}_i - \underline{r}_j|$; $\tau = 1/\varepsilon_w - 1/\varepsilon_p$; ε_w is the solvent dielectric constant (80 at room temperature); and b_i and b_j are

effective, *atomic*, "solvation radii" of the charges i, j. The interaction between two residues, R and R′, can then be written

$$g_{RR'} = \sum_{i\in R, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j])^{1/2}} \quad (8)$$

Earlier, we obtained a Residue GB model by replacing the solvation radii b_i within a particular residue R by an average value b_R.[66] The interaction between two residues R and R′ can then be written

$$g_{RR'} = \sum_{i\in R, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + B_R B_{R'} \exp[-r_{ij}^2/4B_R B_{R'}])^{1/2}} \quad (9)$$

The average solvation radius B_R is related to the GB self-energy of residue R[58,59]

$$E_R^{self} = \sum_{i\in R} E_i^{self} = \frac{\tau}{2}\sum_{i\in R}\frac{q_i^2}{b_i} \stackrel{def}{=} \frac{\tau}{2}\sum_{i\in R}\frac{q_i^2}{B_R} \quad (10)$$

Equivalently

$$(\sum_{i\in R} q_i^2)\frac{1}{B_R} = \sum_{i\in R}\frac{q_i^2}{b_i} \quad (11)$$

Thus, B_R is a harmonic average over the b_i, $i \in R$, weighted by the squared charges.[66]

2.3.2. Step (ii): The Residue Self-Energies Are Sufficient to Parametrize the Pair Screening Energies. A residue-pairwise scheme can now be devised. We note that, for fixed interatomic distances r_{ij}, $g_{RR'}$ is a slowly varying function of $B = B_R B_{R'}$. This dependency can be approximated by a low-order polynomial

$$g(B;r) = (r^2 + B\exp[-r^2/4B])^{-1/2} \approx c_1(r) + c_2(r)B + c_3(r)B^2 + O(B^3) \quad (12)$$

The interaction energy $g_{RR'}$ then takes the form

$$g_{RR'}(B) \approx c_1^{RR'} + c_2^{RR'}B + c_3^{RR'}B^2 \quad (13)$$

Although this approximation usually holds for a large range of $B = B_R B_{R'}$ values, we actually prefer to use a more complex fitting function,[66] which provides very high accuracy for all residue pairs, all rotamer combinations, and all relevant B values

$$g_{RR'}(B) \approx c_1^{RR'} + c_2^{RR'}B + c_3^{RR'}B^2 + c_4^{RR'}B^{-1/2} + c_5^{RR'}B^{-3/2} \quad (14)$$

The two rightmost terms improve the fit accuracy for large B values. The coefficients $c_i^{RR'}$, $i = 1, ..., 5$, will be precomputed for all residue pairs, allowing for all combinations of rotamers and protonation states. During the Monte Carlo simulation, we can then obtain $B_R B_{R'}$ and hence $g_{RR'}$ very efficiently, on-the-fly.

3. Computational Details

3.1. Computation of Energy Matrices. We first consider individual side chains and compute their interactions with the backbone, including all possible titration states and rotamers (see below). Backbone atoms are held fixed in their experimental positions. Pro, Ala, and Gly are treated as a part of the backbone, as are cysteines involved in disulfide bridges. The side chain is positioned in a particular rotamer, then the atomic positions are slightly optimized, through 25 steps of Powell energy minimization, only considering interactions with the protein backbone. The side chain−backbone interaction energy is then computed and stored, and the side chain's "minimized-rotamer" coordinates are saved for future use.

For a side chain pair, (R, R′), we again consider all choices of titration states and rotamers. We use the minimized-rotamer coordinates obtained above; the pair of side chains is further minimized through 10 steps of Powell minimization, to further alleviate the steric overlap that can result from the rotamer approximation. Interactions between the pair and with the backbone are included in the minimization, and the backbone is held fixed as before. The side chain interaction energy is computed and stored. Individual energy terms are stored separately; for example, the contribution of side chain R to the GB self-energy of residue R′ and the contribution of R′ to the GB self-energy of R are stored in a square, asymmetric, self-energy matrix. All the rotamer construction and energy calculations are done with the Xplor program.[80]

3.2. Five-Point Fitting Procedure for the GB Interaction Energies. For a particular pair of side chains, R and R′, and choice of rotamers, the GB interaction energy $g_{RR'}$ was computed for 20 B values, evenly spaced between 1 and 150 Å^2. These energies were fit by the five-point function given in eq 14. The fitting was done with a Fortran program based on the general linear fit subroutine LFIT from Numerical Recipes.[81,66]

3.3. Monte Carlo Simulation Protocol. The Monte Carlo simulation explores the space of side chain rotamers and titration states. Asp, Glu, His, Lys, and Tyr are all considered titratable; there are no cysteines in our data set, except those involved in disulfide bonds. Histidines have three possible protonation states: doubly protonated and singly protonated on either Nε or Nδ. The backbone N- and C-termini are also titratable.

Each Monte Carlo move $(J, N) \rightarrow (J', N')$ changes either the protonation state or the rotamer of either a single group or a pair of groups. The first group is chosen randomly; half of the time, a second group is chosen randomly from among those that have a strong interaction with the first one (±2 kcal/mol or more). Moves are accepted or rejected according to the Metropolis algorithm. If the new state (J', N') has an increased Boltzmann probability, the move is accepted. If it has a decreased probability, the move is accepted with the probability

$$p_{(J,N)\rightarrow(J',N')} = \exp(-\beta\Delta W + \beta(N' - N)\mu) \quad (15)$$

where $\beta = 1/kT$; k is Boltzmann's constant; T is the temperature; and $\Delta W = W_{J'}(N') - W_J(N)$ is the PMF change. For a move that adds a proton (eq 4), the PMF change is computed with the help of the thermodynamic cycle in Figure 1, and the acceptance probability has the form

$$p_{(J,N)\rightarrow(J,N+1)} = e^{-\beta(W_J(N+1)-W_J(N))-2.303pH+\beta\mu^0} = e^{-\beta\Delta\Delta W+2.303(pK_{a,model}-pH)} \quad (16)$$

TABLE 1: Model pK_a Values Used in This Work

titrating group	model pK_a
Asp	4.0
Glu	4.4
Tyr	10.3
Lys	10.4
C-Ter	3.2
N-Ter	9.1
His_δ	6.6
His_ε	7.0

The last equality has made use of eqs 3 and 5. For a move that removes a proton or adds or removes two protons, the acceptance probability is obtained similarly.

Simulations are done at a fixed pH, at a temperature of $T = 300$ K. Each simulation is started from a set of randomly chosen rotamers and titration states. Equilibration is done for 1 million steps. At this point, rotamers that are never accessed are removed, and a further 5 million steps are done. The pH is varied from 0 to 14 in steps of 0.5 units. Thus, a complete pH scan involves 174 million MC steps.

3.4. Model Compounds. The protonation free energy of a residue in the protein is computed as the sum of the protonation free energy of a model compound in solvent and a contribution from the protein, $\Delta\Delta W$, defined in Figure 1.[23,30] The protonation free energy of the model compound in solvent at a given pH is given by eq 5. The model compounds and their pK_a's are listed in Table 1. For Asp, Glu, His, Lys, and Tyr, the model compound is represented by the side chain and backbone atoms of the particular residue. For the N-terminus (respectively, C-terminus), we use the backbone of the first (last) two residues.

3.5. Rotamer Library. Side chain dihedral angles were taken from the Tuffery rotamer library,[82] with the following changes. For C-ter, Asp, and Glu, the number of rotamers was doubled, to ensure that alternate orientations of protonated carboxylic acids were present. The number of Tyr rotamers was also doubled, so that for each Tuffery rotamer we allow both possible orientations of the hydroxyl hydrogen in-plane with the phenyl ring. Similarly, for Ser and Thr, the number of rotamers was tripled to allow, for each Tuffery rotamer, three orientations of the hydroxyl hydrogen. For each side chain, we also added between one and three "native" rotamers, which combine the experimental position of the side chain heavy atoms with the hydrogen orientations just described (two per Asp, Glu, C-ter, and Tyr, three per Ser and Thr).

3.6. Force Field and GB Parameters. We use the Amber all-atom energy function, version parm11.[77] For the titratable residues, the charges of backbone atoms (N, HN, C_α, H_α, C, and O) were set to have average Amber values for these atoms. Charges for side chain atoms of titratable residues were adopted with slight modifications from the Amber force field. No cutoff was used for the nonbonded interactions.

For the solvent, we use a generalized Born variant developed to be consistent with the Amber force field.[60,83] It uses the self-energy treatment of Hawkins et al.,[58] so we refer to it as the GB/HCT model. Recently, we optimized the atomic volumes and other GB parameters to maximize agreement with a large set of Poisson−Boltzmann reference calculations, including mutation free energies and conformational free energies, all done with the same Amber force field, version parm11.[84] The PB reference calculations used atomic radii specifically optimized for use with the Amber charges.[85] The force field and GB variant are implemented in the Xplor program,[80,86,83] which was used for all the energy calculations.

TABLE 2: Protein Structures Used in This Work

protein	PDB structure	X-ray resolution	experimental pH
bovine pancreatic trypsin inhibitor (BPTI)	4PTI[104]	1.50 Å	NA[a]
streptococcal protein G, B1 Ig-binding domain	1PGA[144]	2.07 Å	4.5
turkey ovomucoid third domain (OMTKY3)	2GKR[145]	1.16 Å	7.5
chicken lysozyme	2LZT[146]	1.97 Å	4.5
Bacillus amyloliquefaciens barnase	1A2P[147]	1.50 Å	7.5
Escherichia coli thioredoxin, oxidized	2TRX[148]	1.68 Å	7.5

[a] Not available.

3.7. Protein Set and Atomic Coordinates. The pK_a calculations were done for six proteins, listed in Table 2. Their chain lengths are between 56 and 129 amino acids. Crystal structures were used, with a resolution of 2.1 Å or better. Crystal waters and ligands were removed. Hydrogens of protonated carboxylic acids (aspartates, glutamates, C-termini) were added in the most common, syn orientation. Other hydrogens were first positioned with the HBUILD facility in XPlor;[87] all hydrogen positions were then adjusted by full-energy minimization, with all non-hydrogen atoms fixed and all titratable groups kept in their protonated forms.

3.8. Poisson−Boltzmann pK_a Calculations. For comparison to the GB method, single-conformation Poisson−Boltzmann (SC-PB) calculations were done, using the MEAD program[88,89] to solve the linearized PB equation. We used the slightly modified Amber atomic charges described above, along with atomic radii specifically optimized for PB calculations with the Amber charges.[85] The protein dielectric constant was set to 4 or 8, and the solvent dielectric constant was set to 80. The boundary between the protein and solvent regions was taken to be the molecular surface, constructed with a probe radius of 1.4 Å. Calculations were done at 300 K with an ionic strength of 0.150 M. The PB equation was solved using a three-step focusing procedure, and the spacings between grid points at the three steps were 4, 1, and 0.5 Å. To determine the protonation pattern at a given pH, we used the Monte Carlo procedure implemented in the MCTI program,[90] obtaining the occupancies of the deprotonated and protonated states of all the titratable residues. The pH was varied from −10 to 30 in steps of 0.1 pH unit. The pK_a of each titratable group was taken to be the pH where the occupancies of the protonated and deprotonated forms were equal.

3.9. PROPKA Calculations. We used the PROPKA interactive web interface, http://propka.ki.ku.dk/, which takes a PDB code as input and returns the pK_a values for titratable amino acids.

3.10. Comparing Residue-GB to Atomic-GB and to the Poisson Model. To test the residue-GB method, we report a series of tests, comparing it to atomic-GB and the Poisson model. We compute free energy changes associated with conformational, protonation, and chemical modifications in four medium-sized proteins. The chemical modifications are described in the Supporting Information.

3.10.1. Rotamer Calculations. For each protein, we constructed a set of 3000 structures by randomizing the side chain rotamers of all residues except prolines, alanines, and cysteines. Small voids in the interior of the rotameric structures were filled by dummy atoms, to prevent the occurrence of artificial, high-

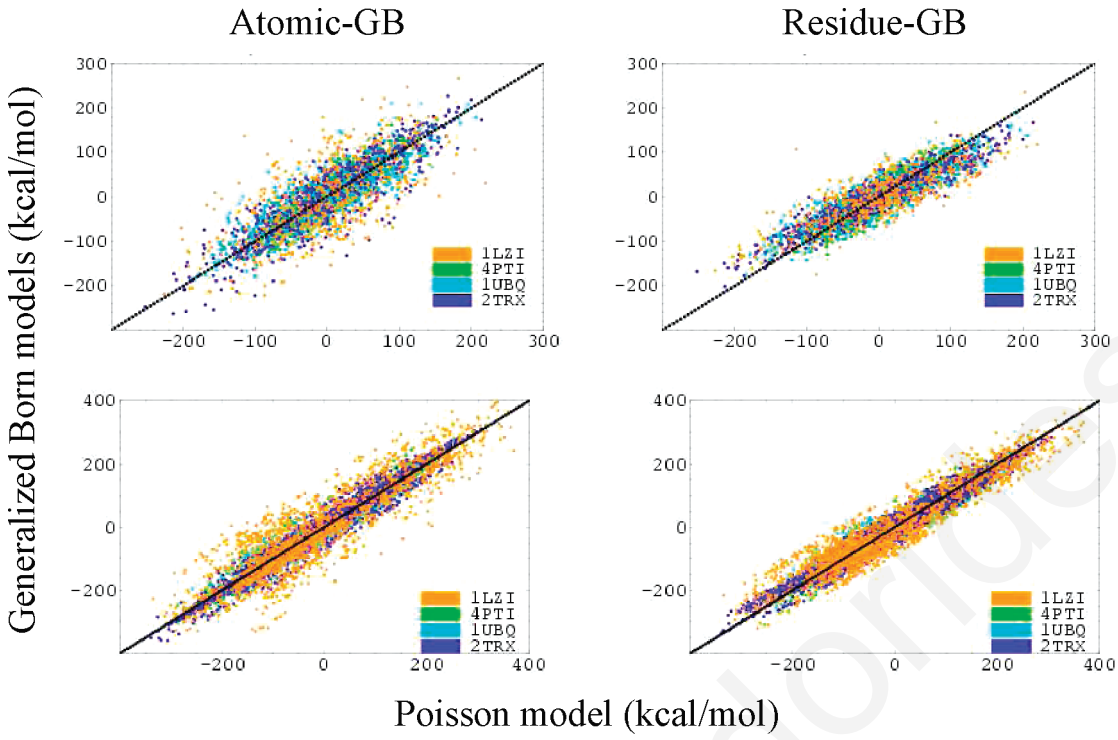


Figure 2. Solvation free energies of four proteins for multiple rotamer combinations (upper panels) and multiple protonation states (lower panels). Atomic-GB (left, vertical axis) and residue-GB (right) are compared to the Poisson equation (horizontal axis). The proteins are identified by their PDB codes and a color scheme.

dielectric internal cavities. The protein backbones were held fixed in their X-ray conformations, taken from the following four PDB structures: 1UBQ for ubiquitin; 4PTI for BPTI; 1LZ1 for lysozyme; 2TRX for thioredoxin.

3.10.2. Mutations in the Protonation State of Titratable Residues. In each of the four proteins, the titratable residues Asp, Glu, His, Cys, and Lys were initially assigned their most common protonation states at physiological pH. Subsequently, the charge state of one titratable residue at a time was modified, and the solvation free energy change was computed for 100 protein conformations, created by randomizing the rotameric state of all side chains. The chemical types, charges, and radii for the different charge states of titratable residues were taken from the AMBER force field, as for the Monte Carlo simulations (see above).

3.10.3. Reference Calculations with the Poisson Model. For these calculations, where residue-GB is compared to the Poisson model, we used a slightly different Poisson protocol, compared to the pK_a calculations, above. For a given protein structure, we solved the Poisson equation with the finite-difference program UHBD[91] (instead of MEAD, above). The protein/solvent dielectric boundary was defined by the molecular surface of the protein. The solution employed a two-step focusing procedure and a cubic grid with spacings of 0.8 and 0.4 Å. All calculations were done with a protein dielectric constant $\varepsilon_p = 1$ and a solvent dielectric $\varepsilon_s = 80$. The same protein dielectric is used for the GB model (unlike the pK_a case, above). Indeed, for these GB/PB comparisons, what is essential is the consistency between the various treatments, rather than the precise choice of protein dielectric constant. The molecular surface was constructed with 2000 points per atom, using a probe sphere of radius 2 Å and the boundary smoothing method in UHBD.[91] The atomic charges corresponded to the AMBER, all-atom force field (see above). The atomic radii were those specifically optimized for Poisson calculations with the AMBER charges.[85]

4. Results

4.1. Residue-GB Compares Well with the Poisson Model. This work represents the first application of residue-GB. Therefore, before describing the pK_a results, we report a series of tests that specifically compare residue-GB to the more traditional, atomic-GB and to the Poisson method. We introduced a large number of conformational, protonation, and chemical modifications into four, medium-sized proteins, ubiquitin, BPTI, lysozyme, and thioredoxin, and computed the corresponding free energy changes. The Poisson model is considered the benchmark for accuracy, following common practice with GB model development. Below, we describe the conformational and titration changes, and the chemical mutations are described in Supporting Information.

4.1.1. Conformational Energies. The GB electrostatic solvation free energies for the rotameric structures of the four proteins lysozyme, thioredoxin, ubiquitin, and BPTI are plotted against the corresponding PE values in Figure 2. Results are shown for both atomic- and residue-GB. The free energies vary over a 500 kcal/mol range. The present variant of atomic GB was optimized earlier by comparison to PE[84] and agrees very well with the PE model across the whole energy range. The rms differences between the GB and PE electrostatic solvation free energies are given in Table 3. The values for the four proteins vary between 21 and 45 kcal/mol for atomic-GB and between 21 and 40 kcal/mol for residue-GB. Overall, residue-GB performs as well as atomic-GB for three of the proteins and slightly better for lysozyme.

4.1.2. Protonation of Titratable Residues. Here, we systematically change the protonation state of titratable amino acids in the four proteins considered above. The resulting GB solvation energies are plotted against the corresponding PE energies in Figure 2. The corresponding rms differences are given in Table 3. The GB values correlate well with PE, with

TABLE 3: Comparing Atomic-GB and Residue-GB to the Poisson Model[a]

	rotamers		protonation states	
protein	atomic-GB	residue-GB	atomic-GB	residue-GB
1LZ1	45.3	39.4	52.0	42.8
2TRX	30.2	28.9	24.5	24.9
1UBQ	26.0	27.7	25.4	24.9
4PTI	20.7	20.9	22.2	22.7

[a] RMS deviation (kcal/mol) between the solvation free energies from GB and the Poisson model (PE) for random rotamer combinations and protonation states of titratable side chains in four proteins (indicated by their PDB code; see text).

rms differences of between 22 and 52 kcal/mol with atomic-GB for the four proteins and between 23 and 43 kcal/mol for residue-GB. Residue-GB is somewhat better than atomic-GB for lysozyme and comparable for the other three proteins. All the GB pK_a results, below, were obtained with residue-GB.

4.2. Comparison between Computed and Experimental pK_a's. ***4.2.1. Overview of the Results.*** We now consider the titration behavior of 78 titratable groups in six proteins. Our MC simulations yield the distribution of protonation states for each group as a function of pH, in other words, the titration curves. Experiments usually yield the same information[3] (typically deduced from a change in the NMR chemical shifts with pH). However, the titration curves are not usually reported in the experimental papers; rather, a single pK_a value is given for each group, corresponding to the inflection point. For proteins, the relation between the titration curves and the individual pK_a's is not completely straightforward, as recognized by Tanford, Ullmann, and others.[49,96,92−95] Indeed, coupling between titrating groups can affect the shape of the curves, and one must distinguish carefully between alternate definitions of the proton binding reaction. For example, the binding of a proton to a group of interest can be studied with the other groups held in a fixed protonation state or free to adapt (as here, both in the computations and the cited experiments). The experimental pK_a values are normally the inflection points of the titration curves, so that they correspond to the following "Henderson−Hasselbach" pK_a definition[92,95]

$$pK_{a,i} = pH + \log \frac{x_i}{1 - x_i} \qquad (17)$$

where i represents a titrating group and x_i is the mean population of its protonated form at the given pH. By comparing our computed inflection points with the experimental ones, we are consistent with the experimental definition, and we refer to these inflection points simply as pK_a's.

Our test set includes eight residues for which only experimental upper or lower bounds are available. These residues are Asp54, Asp93, Asp101, and Glu73 in Barnase and Asp7, Asp27, Tyr31, and the C-terminus in OMTKY3. Since precise experimental pK_a's for these residues were not known, they were treated in a special way:[14] an experimental value was assumed to be equal to the computed, if the computed pK_a was in the experimental range; otherwise, it was assumed to be the experimental lower or upper bound. Several theoretical methods were compared to each other and to experiment: a "standard" Poisson approach (denoted SC-PB), using a single protein conformation; residue-GB with a single protein conformation (SC-GB); and our constant-pH, Monte Carlo approach. With this last approach, the pK_a calculations involve three general stages: (1) computing the residue−residue interaction energy matrices; (2) fitting the residue−residue GB interaction energies with a simple function (leading to five matrices of fitting coefficients; see Methods); and (3) the Monte Carlo simulations themselves, which explore the space of side chain conformations and protonation states and yield the average protonation states as a function of pH. This three-stage workflow is schematized in Figure 3. We also consider the Null model and the empirical, PROPKA method.[14]

Table 4 and Figure 4 summarize the agreement between experiment and the various computational methods. Good agreement is obtained with the simplest, Null model, as already shown by many authors.[40,97,32,63] The rms deviation is just 1.07 pH units, and the maximum error is 3.5. The empirical, PROPKA method[14] gives even better results, with an rms error of just 0.88 units, a maximum error of 4.4, and a correlation between PROPKA and experiment of 0.74. With PROPKA, the pK_a shift is expressed as a sum of physicochemical and empirical contributions, which take into account desolvation effects, the number and geometry of hydrogen bonds, and long-range electrostatic interactions. The different terms were fitted to reproduce experimental pK_a's. In fact, three of the six proteins considered here (BPTI, OMTYK3, and HEWL) were also part of the PROPKA parametrization set, which included just two other proteins (RNase A and RNase H). Thus, there may be a favorable bias; however, PROPKA gave results nearly as good, earlier, for proteins not included in the parametrization set.[14]

The "standard" SC-PB method with a protein dielectric value of four gives an rms deviation from experiment of 2.3 pH units, with two large errors: Y53 in 2LZT (error 10.8) and D75 in 1A2P (error 9.7). Excluding these residues, the mean error is 1.8. SC-GB (with the same dielectric) gives a lower deviation of 1.6 units. The maximum error is just 4.3 with SC-GB. The correlation between theory and experiment is also higher with SC-GB, 0.71 compared to 0.67. Increasing the protein dielectric to eight gives significantly lower errors of 1.36 (SC-PB) and 1.30 (3.9), with similar correlations. This is consistent with earlier studies, where a higher protein dielectric gives lower average errors,[40,97] with values as large as 20−80 giving the best results. Notice that this does not necessarily mean that a large protein dielectric is the more physically correct value. Indeed, continuum models contain systematic errors that can make the interpretation less straightforward (see Discussion). Computer simulations and experimental data on dry protein powders indicate that the interior of globular proteins is best represented with a low dielectric value of around 4−8.[10]

The multiconformation GB method gives the best results among the continuum electrostatic approaches considered here. With a protein dielectric of four, the rms error is 1.22, only slightly greater than the Null model; the maximum error is 3.9, slightly lower than PROPKA; and the theory/experiment correlation is 0.77, slightly higher than PROPKA. If the protein dielectric is increased to eight, the quality of the results changes only slightly: the rms error decreases slightly (to 1.16), while the maximum error increases slightly and the correlation decreases slightly. It is satisfactory that good results are obtained with a rather low protein dielectric of four since the side chain degrees of freedom are explicitly represented at the atomic level (through the MC simulations); they should not also be represented implicitly through a high protein dielectric (see Discussion). Finally, the performance is equally good for highly shifted and weakly shifted pK_a's, as reported later on.

For MC-GB, the Null model, and PROPKA, the rms deviations from experiment are fairly close (1.2, 1.1, and 0.9

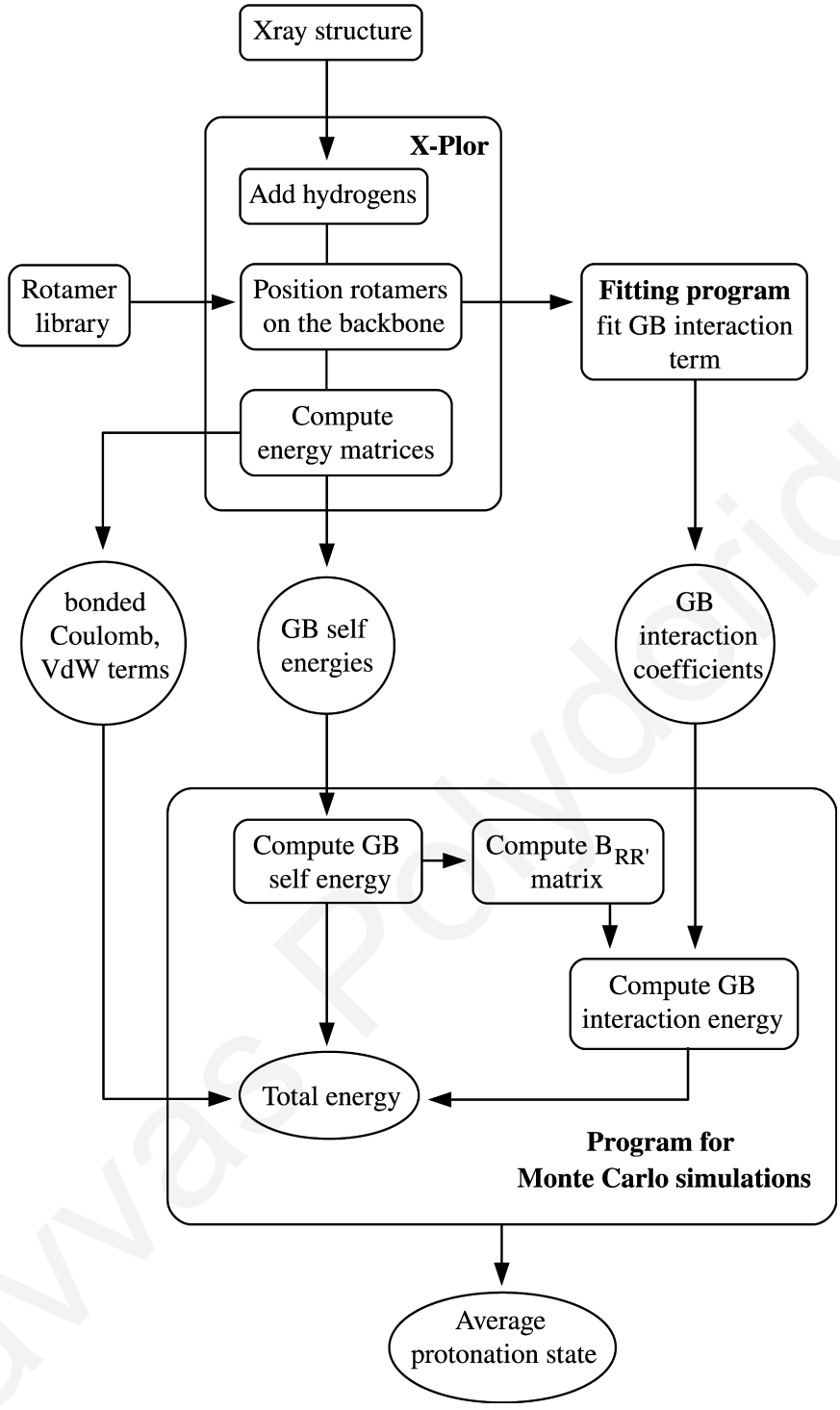


Figure 3. Chart of the computations performed with the multiple-conformation GB method. The input is an X-ray structure from the PDB. The output includes the computed pK_a values for titratable groups.

pK_a units), and so we have done statistical tests to assess the significance of the differences between models. For each model, we considered the square deviations of the 78 computed pK_a's from experiment, say $\{\delta X_i^2\}_{GB}$, $\{\delta X_i^2\}_{Null}$, and $\{\delta X_i^2\}_{PROPKA}$, as random quantities. We used an F-test to establish that their variances are significantly different.[81] Then we used two-tailed Student's tests to test the null hypothesis that the three data sets are all instances of the same random variable δX^2, given the observed differences between their means.[81] The data correspond to about 130 degrees of freedom in the Student's sense, and the t values are 1.56 for GB vs the Null model and 2.24 for GB vs PROPKA. Thus, we may assert with a 99% confidence level that the GB and PROPKA rms deviations are significantly different. Comparing GB and the Null model, the confidence level is only 88%, so that the difference in performance between the two models is only moderately significant.

Table 5 summarizes the results obtained by several other groups for proteins included in our test set. Constant-pH MD simulations by Brooks and co-workers led to better agreement for lysozyme but similar or slightly worse agreement for BPTI, OMTKY3, and barnase.[52,98] Nielsen and Vriend obtained an rms error of about 0.9 for five of our proteins, using a single conformation PB approach with a large protein dielectric of 16.

TABLE 4: Comparison of Different pK_a Methods

method	protein dielectric	rms deviation[a] (maximum)[b]	correlation[c]
single conformation PB	$\varepsilon_p = 4$	2.34 (10.8)	0.67
single conformation GB	$\varepsilon_p = 4$	1.62 (4.3)	0.71
multiconformation GB	$\varepsilon_p = 4$	1.22 (3.9)	0.77
single conformation PB	$\varepsilon_p = 8$	1.36 (5.7)	0.67
single conformation GB	$\varepsilon_p = 8$	1.30 (3.9)	0.70
multiconformation GB	$\varepsilon_p = 8$	1.16 (4.2)	0.71
Null model	–	1.07 (3.5)	–
PROPKA	–	0.88 (4.4)	0.74

[a] rms deviation between computed and experimental pK_a's. [b] In parentheses: maximum error. [c] Pearson correlation coefficient between computed and experimental pK_a's.

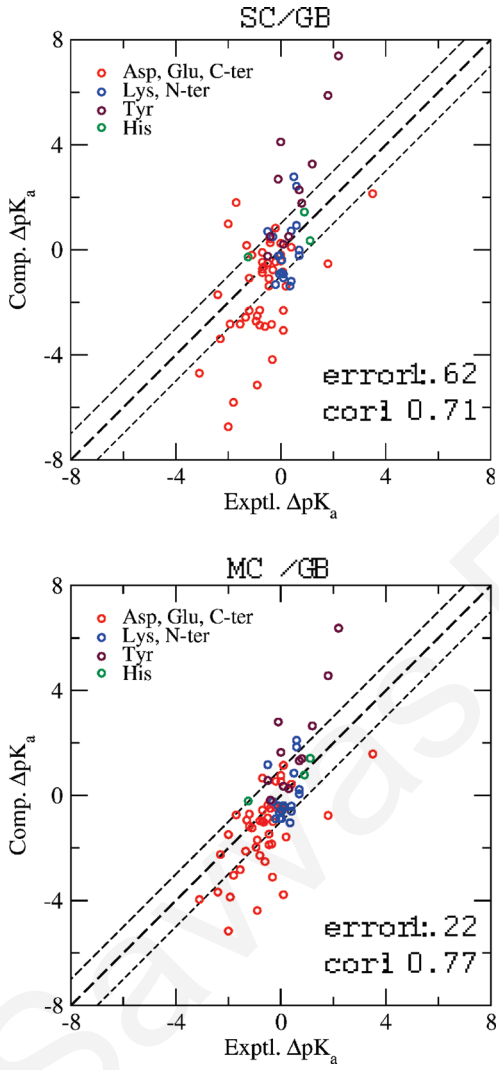


Figure 4. Comparison of the computed and experimental pK_a shifts from the single-conformation method (SC-GB, upper panel) and the multiple-conformation method (MC-GB, bottom panel). The thick dashed line in each inset represents perfect agreement between predictions and experiments; the thin dashed lines are 1 pH unit above and below. The titrating groups are identified by colors. The rms error and the correlation between computed and experimental pK_a shifts are indicated.

A large protein dielectric is expected to give poor results for buried groups and highly shifted pK_a's; these authors' data set included 8 buried groups out of 127, and their rms error was 1.25. The multiconformation PB method of Gunner and coworkers gave an rms error of 0.9 for four proteins.[48] Wisz and Hellinga obtained an rms error of 0.8 for five proteins,[63] using a rather sophisticated, semiempirical method. Warwicker obtained an improved rms error of 0.6 with a hybrid method that treats surface groups with a Debye−Huckel approach and more buried groups with a traditional, SC-PB approach.[99] Finally, very good results were obtained by Spassov and Yan, using a GB solvent and a physiological ionic strength, with an rms deviation of 0.4 for lysozyme, BPTI, and protein G.[100] The use of ionic strength may have helped reduce the error, as noted by Brooks and co-workers.[98] This rapid survey is obviously very incomplete. It does show that improved agreement can be obtained, especially if one uses a higher protein dielectric and/or additional parametrization. Nevertheless, the present MC-GB approach is not much poorer overall, has a good ability to treat highly shifted groups (see below), and can be further improved (see Discussion).

With the MC-GB method, the total CPU time needed for a small protein (1A2P) is about 123 h, including 72 h for the matrix calculations, 12 h for the fitting coefficients, and 39 h for the Monte Carlo simulations. Using a single, eight-core computer, the entire calculation takes under 16 h. Obviously, the empirical models are much faster; for example, the PROPKA calculations take just a few minutes.

We have also analyzed the titration curves obtained with MC-GB and their slopes. The maximum slope of each curve can be interpreted as Hill's coefficient n, which measures the influence of other titrating groups on the group of interest.[5] The titration curves for BPTI are shown in Figure 5. Overall, out of 78 titrating sites, 68 (87%) have n values below 0.85; the rest (13%) have n values between 1.1 and 0.85. In a previous study,[48] the computed (respectively, experimental) slopes had the following distribution: 38% (50%) below 0.85; 60% (48%) between 0.85 and 1.1; and 2% (2%) above 1.1. Thus, our slopes tend to be somewhat too low. For 33 groups with published experimental slopes, our slopes had an rms error of 0.23 and a mean error of −0.13. We found that the computed slopes were sensitive to details of the Monte Carlo procedure, such as the number of conformational relaxation steps following each protonation change. More work is needed to investigate this point.

4.2.2. Large and Small pK_a Shifts. It is important to examine separately groups that have significant pK_a shifts (compared to the usual model compound).[32] Indeed, these are the groups that are hardest to predict and for which the Null model is not satisfactory. In addition, they are often functionally important.[101,49] Out of 78 sites considered in this work, 57 have weak experimental pK_a shifts, $|\Delta pK_a| < 1$ pH unit; 21 have noticeable shifts, $|\Delta pK_a| \geq 1$; and 8 have large shifts, $|\Delta pK_a| \geq 2$. Table 6 summarizes the prediction results.

With SC-PB and a protein dielectric of four, the intermediate group has the largest rms error: 3.0. However, this large value is entirely due to two sites, Y53 in 2LZT and D75 in 1A2P, which give errors of 10.8 and 9.7. For the 20 other sites, the errors are similar to the low and high shift cases. For the highest shifts (eight sites), the predictions are actually a little better than for the low shifts (1.91 vs 2.06). With SC-GB (and a protein dielectric of four), the intermediate group also gives the largest rms error, 1.9, compared to 1.5 for the low and large shifts. With PROPKA, the errors are similar across the whole range of shifts (0.9 units); the largest error is in the low shift group.

Finally, with MC-GB, the performance actually improves slightly as the shifts become larger: the intermediate group has the same rms error as the low shift group (1.2) but a smaller maximum error (2.8), while the larger shifts are reproduced within one pH unit (rms error of 0.93), comparable to PROPKA (but without any empirical parameter adjustment). Thus, by

TABLE 5: RMS Deviation between Experimental and Computed pK_a's: Comparison to Some Earlier Studies[a]

authors	overall	lysozyme	BPTI	OMTKY3	protein G	barnase
Khandogin, Brooks[98b]		−(0.7)	−(0.9)	−(0.6)	−	−(1.0)
Nielsen, Vriend[149]	0.87	0.66	0.60	1.19	0.87	0.90
Georgescu et al.[48]	0.93	0.81	0.67	1.26	0.63	−
Wisz, Hellinga[63]	0.76	0.69	0.45	0.80	0.53	1.17
Warwicker[99]	0.64	0.47	0.35	0.77	0.80	0.76
this work[b]	1.21 (0.86)	1.47 (1.28)	0.77 (0.65)	0.95 (0.52)	0.95 (0.79)	1.55 (0.90)

[a] rms deviations between the computed and experimental pK_a's. [b] Mean unsigned error in parentheses.

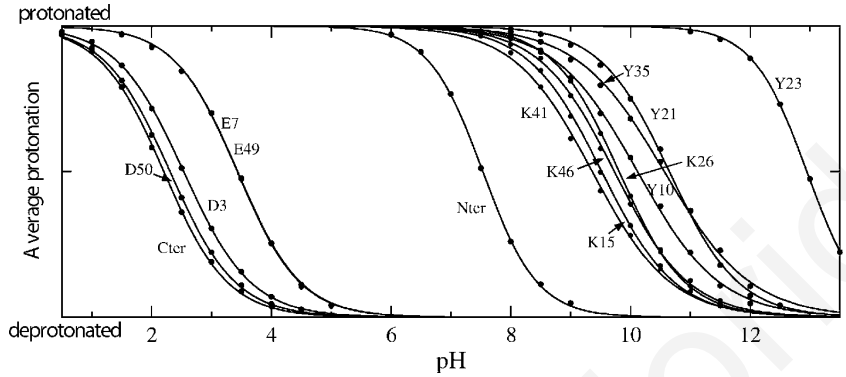


Figure 5. Titration of the 14 amino acids in BPTI. The dots are the populations obtained from the Monte Carlo simulations with MC-GB, and the smooth curves correspond to fits using Hill's expression. Each curve is labeled by the titrating group.

TABLE 6: Performance for Weakly and Strongly Shifted pK_a's[a]

		method			
experimental shift	number of groups	MC-GB	SC-GB	SC-PB	PROPKA
$\|\Delta pK_a\| < 1$	57	1.22 (3.9)	1.52 (4.3)	2.06 (9.7)	0.86 (4.4)
$\|\Delta pK_a\| > 1$	21	1.22 (2.8)	1.85 (4.1)	2.96 (10.8)	0.95 (2.3)
$\|\Delta pK_a\| > 2$	8	0.93 (1.9)	1.47 (3.5)	1.91 (4.4)	0.85 (2.3)
all pK_a's	78	1.22 (3.9)	1.62 (4.3)	2.34 (10.8)	0.88 (4.4)

[a] rms deviations between the computed and experimental pK_a's (maximum deviation in parentheses). Results are shown separately for groups with small and large shifts (relative to the corresponding model compounds).

explicitly representing the conformational reorganization of side chains in response to protonation changes, we improve the prediction quality, as observed by earlier workers.[48,49,65,102,52,103] This is another indication that the MC-GB approach is physically meaningful.

4.2.3. Behavior of Individual Groups. In this section, we describe selected groups and the effect of conformational mobility on their predicted pK_a's. This will help understand the limitations of the current approach and possible ways to improve it.

BPTI Glu49 and Asp50. Using the single conformation GB approach, we compute the pK_a of Glu49 to be 2.1 in moderate agreement with the experimental value of 3.6. Including side chain flexibility gives a pK_a of 3.4, in close agreement with experiment. The structures observed in the Monte Carlo simulations are illustrated in Figure 6. Upon ionization, Glu49 reorients to make a salt bridge with Arg53. The Oδ (Glu49)···NH (Arg53) distance is 4.6 Å at pH = 2 and 3.1 Å at pH = 4, after Glu49 ionization. In the X-ray structure used in this work (PDB entry 4PTI[104]), Glu49 is positioned to interact with its backbone amino group. In the MC simulations, in contrast, Glu49 prefers to be solvent-exposed when protonated or to interact with Arg53 when ionized. Interestingly, in the solution structure (PDB entry 1PIT[105]), determined by NMR, Glu49 also makes a salt bridge with Arg53 in some of the reported conformers.

For Asp50, the MC-GB approach gives a pK_a of 2.3, in good agreement with the experimental value of 3.1; SC-GB gives a pK_a of 1.5, in poorer agreement. In the X-ray structure, Asp50 makes a salt bridge with Arg53 and a hydrogen bond with nearby Ser47, which can explain the experimental pK_a downshift. In the MC simulations, protonated Asp50 does not interact with Arg53 (Oδ(Asp50)···NH_2(Arg53) distance of 5.3 Å), while at a higher pH, deprotonated Asp50 makes a salt bridge with Arg53. The corresponding conformations are illustrated in Figure 6.

Lysozyme Asp66. The experimental pK_a of Asp66 in Lysozyme is 0.9, which corresponds to a very large downshift. The ionized side chain form is stabilized by hydrogen bonds with the hydroxyl groups of Tyr53, Thr69, and Ser60 and with the backbone amino groups of Thr69, Arg68, and Asp66. SC-GB gives a pK_a of −0.7, 1.6 units below the experimental value. With MC-GB, the computed pK_a increases to 0.0, in better agreement with experiment. This improvement is due to a reorientation of the Thr69 hydroxyl upon protonation/deprotonation of Asp66. When Asp66 is protonated, Thr69 donates a hydrogen bond to Gly49. When Asp66 is ionized, Thr69 makes a hydrogen bond to the Asp66 side chain. Since the X-ray structure was determined at a high pH, where Asp66 is always ionized, the positions of the backbone groups are optimized to favor interactions with deprotonated Asp66. With our current implementation of MC-GB, these backbone groups cannot reorganize explicitly upon Asp66 protonation/deprotonation; their reorganization is treated implicitly, through the protein dielectric constant.

This residue was also studied recently by another method that takes into account side chain flexibility (in a more limited way).[102] With a single conformation method, a pK_a of −2.4 was obtained, while the multiconformation method gave a pK_a of 0.9, in perfect agreement with experiment. The improvement was attributed to reorientation of the same, Thr68 hydroxyl group.

Lysozyme Tyr53. The experimental pK_a of this residue is 12.1, upshifted by 1.8 units, compared to the model compound. The

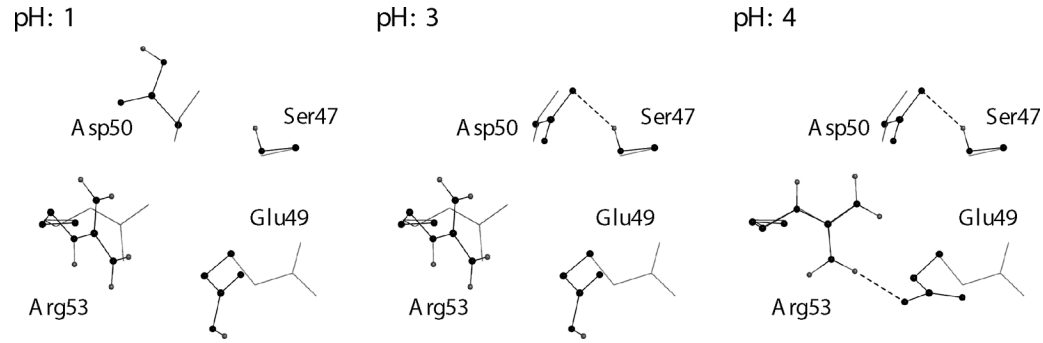


Figure 6. BPTI conformations observed in the Monte Carlo simulations at three selected pH's. The experimental X-ray structure[104] is shown in gray. For clarity, backbone atoms and nonpolar hydrogens are not shown. Polar hydrogens are gray.

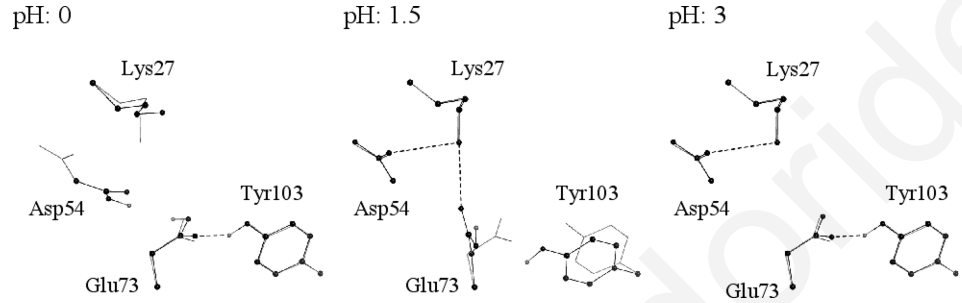


Figure 7. Barnase conformations observed in the Monte Carlo simulations at three selected pH's. The experimental X-ray structure[147] is shown in gray. For clarity, backbone atoms and nonpolar hydrogens are not shown. Polar hydrogens are gray.

upshift can be attributed to the interaction of Tyr53 with Asp66, discussed above. With the experimental structure and SC-GB, we compute the pK_a of this group to be 16.2. Taking into account side chain flexibility, through MC-GB, we obtain 14.9, which is an improvement, but still an overestimate. This could occur because our rotamer library is too limited for tyrosine (only three rotamers differ by more than 30°). Indeed, Tyr53 is partly buried in the X-ray structure, and reorientation toward the solvent may be needed to stabilize the ionized state. In the MC simulations, Tyr53 remained in its crystallographic conformation, possibly because there were not any suitable, more solvent-exposed rotamers in the library. Notice also that at a pH above 12 the protein may begin to unfold, which would then favor Tyr53 ionization. Unfolding is not possible with our rigid-backbone MC approach.

Lysozyme Glu35. The experimental pK_a of this site is 6.2, representing a 1.8 unit upshift. SC-PB with the experimental structure gives a pK_a of 4.6, barely upshifted. With SC-GB and MC-GB, we obtain 3.6 and 3.9, respectively, slightly downshifted. In the X-ray structure, one of the Glu35 side chain oxygens interacts with the Ala10 backbone NH, while the other oxygen is solvent-exposed. Thus, the source of the experimental upshift is not evident.

Barnase Asp54 and Glu73. The experimental pK_a of Asp54 is not known precisely but is less than 2.2. With SC-GB, we predict a very low value of −1.8; with side chain flexibility (MC-GB), we predict a larger value of 1.0, still well within the experimental range. In the X-ray structure (determined at a pH of 7.5, where Asp54 is ionized), Asp54 makes a salt bridge with Lys27 and a hydrogen bond with the backbone NH of the same Lys27. These interactions can explain the downshifted pK_a. In the MC simulations, Asp54 rotates away from Lys27 at low pH, with a distance increasing from 3.0 Å (X-ray structure) to 4.1 Å, and loses its hydrogen bond to the Lys27 backbone; it then makes a weak hydrogen bond to Glu73, with a Oδ(Asp54)···Oε(Glu73) distance of 4.1 Å. This reorganization is illustrated in Figure 7. At pH = 2, where Asp54 becomes ionized, it rotates back to interact with Lys27. The predicted Nζ(Lys27)···Oδ(Asp54) and NH(Lys27)···Oδ(Asp54) distances are 3.3 and 2.9 Å, respectively, close to the X-ray values.

As for Glu73, also shown in Figure 7, we find that it reorients upon protonation, losing a hydrogen bond to Tyr103; this leads to a smaller computed pK_a shift when side chain flexibility is included.

OMTKY3 Asp7 and Asp27. The Asp7 pK_a is not known precisely, but it is less than 2.2. SC-GB predicts 4.2, which is too high. MC-GB predicts 3.1, which is an improvement, though still higher than the experimental upper bound. In the MC simulations at pH = 2, where Asp7 is ionized, one of the Asp7 side chain oxygens interacts with the Ser9 backbone NH; the other one makes a hydrogen bond with the Ser9 side chain hydroxyl.

Another case where SC-GB and MC-GB differ is Asp27, whose experimental pK_a is 2.3. With SC-GB and the X-ray structure, we compute a pK_a of 5.8, so that the protonated form is overstabilized. In the crystal structure, Asp27 interacts with Tyr31, with an Oδ(Asp27)···OH(Tyr31) distance of 2.6 Å. In the MC simulations with deprotonated Asp27, the Tyr31 hydroxyl reorients to donate a hydrogen bond to Asp27. The MC-GB pK_a prediction is 3.3, in much better agreement with experiment.

Thioredoxin Asp20 and Asp26. In oxidized *Escherichia coli* thioredoxin, Asp26 is partly buried and has a large pK_a of 8.1, while solvent-exposed Asp20 has an unshifted pK_a of 4.[106] The titration of these two residues was studied previously by MD free energy simulations in explicit solvent and also by MD simulations with a GB solvent.[33] Here, we obtain an Asp20 pK_a of 3.6, with either SC-GB or MC-GB, in good agreement with experiment. This result also agrees with the previous GB-solvent MD simulations, where the computed pK_a shift was zero[33] (notice that the explicit-solvent MD simulations underestimated the pK_a[33]).

For Asp26, SC-GB predicts a pK_a of 6.1, while MC-GB gives 5.6, slightly farther from the experimental result. This is one of

the largest errors obtained with MC-GB. We do not observe any conformational change when Asp26 becomes protonated, which explains why the SC-GB and MC-GB results are similar. In the earlier MD simulations, when Asp26 became ionized, Lys57 reoriented to form a salt bridge with it; in contrast, protonated Asp26 did not interact closely with Lys (consistent with the X-ray structure, where Asp26 is presumably protonated). This behavior was seen with both explicit and implicit (GB) solvent.[33] Here, in contrast, Lys57 remains 4.6 Å away from ionized Asp26 (Nζ(Lys57)···Oδ(Asp26) distance). Nevertheless, the pK_a shift computed here with MC-GB (+1.6 units) is close to the value obtained from the earlier MD study (pK_a shift of +2.3 units with the GB MD simulations).[33]

5. Discussion

5.1. Hybrid Models: Modeling Dielectric Relaxation. We begin by considering the nature of the hybrid dielectric model used here, with protein backbone, protein electronic, and solvent motions all treated implicitly, while protein side chain motions were treated explicitly. The relation between the explicit and implicit treatments was analyzed in a profound way by Kirkwood and Fröhlich,[107,54,108] and it is instructive to make a brief detour and consider their analysis in the case of a spherical molecule embedded in a uniform solvent.[109−112] The spherical molecule can be thought of as a protein caricature. Using linear response theory, Kirkwood and Fröhlich related the fluctuations of the molecule's dipole moment ΔM, treated explicitly, to the protein and solvent dielectric constants that would be used in an implicit treatment, ε_p and ε_w

$$\frac{\langle \Delta M^2 \rangle}{kTR^3} = \frac{(\varepsilon_p - 1)(1 + 2\varepsilon_w)}{\varepsilon_p + 2\varepsilon_w} \tag{18}$$

Here, R is the protein radius, and the brackets $\langle\rangle$ represent thermal averaging. Fröhlich went further, introducing the first hybrid model, with the fast degrees of freedom treated as a dielectric continuum (dielectric constants ε_p^∞, ε_w^∞ for the protein and solvent) and the slow degrees of freedom treated explicitly. The slow degrees of freedom then obey a relation analogous to eq 18[54,112,113]

$$\frac{\langle \Delta M_{slow}^2 \rangle}{kTR^3} = \frac{(\varepsilon_p - 1)(\varepsilon_p^\infty + 2\varepsilon_w)}{\varepsilon_p + 2\varepsilon_w} - \frac{\varepsilon_w^\infty(\varepsilon_p^\infty - 1)(\varepsilon_p^\infty + 2\varepsilon_w)}{\varepsilon_w(\varepsilon_p^\infty + 2\varepsilon_w^\infty)} \tag{19}$$

In this relation, the separation between explicit (slow) and implicit (fast) degrees of freedom is obtained by a rigorous method, which assumes only that the slow degrees of freedom do not respond to a high frequency electric field.[54] The explicit degrees of freedom are represented by a single quantity, $\langle \Delta M_{slow}^2 \rangle$, which is an average over many atoms and many conformations. Nevertheless, eqs 18 and 19 show how, in a simple case, a thermodynamic property, $\langle \Delta M^2 \rangle$, can be obtained rigorously from a hybrid model.

For protonation free energies, which are of interest here, an even simpler relation exists if the system geometry is simple enough. Specifically, let us view the fast degrees of freedom of protein and solvent as a single, uniform, dielectric continuum.[113,114] This should be a rather accurate approximation for the electronic degrees of freedom of protein and solvent. Indeed, these ideas have been tested since the early days of electron transfer theory.[17,9] In that case, we have the following relation between the free energy ΔG_{MC} computed from the hybrid model and the true free energy, ΔG[113]

$$\Delta G = \Delta G_{MC} + \Delta G^{el} \tag{20}$$

ΔG_{MC} may be obtained (as in Results) from an MC simulation with a uniform dielectric constant ($\varepsilon^\infty = \varepsilon_p^\infty = \varepsilon_w^\infty$), and ΔG^{el} is the free energy when only the fast degrees of freedom are modeled (implicitly, through the continuum model[17,54]). In our calculations and most other pK_a calculations with hybrid models, the ΔG^{el} term can be considered to cancel when the protein and model compound are compared. Thus, with a hybrid model where only the electronic degrees of freedom are treated implicitly, the protonation free energies are obtained from a well-defined approximation.

The next step is to treat *all* the solvent degrees of freedom implicitly, as in most (though not all[115]) hybrid models used to date. This step is actually straightforward since it consists of replacing the potential energy of the protein by a potential of mean force (PMF), as discussed above (Theory).[42] Many studies have shown that a good GB variant is a reasonable approximation to the PMF. Thus, integrating out the solvent degrees of freedom is a well-tested approximation, which can be used in combination with the implicit treatment of the electronic degrees of freedom just discussed.

Finally, we want to integrate out the backbone motions and take them into account through an increased protein dielectric value. While this is a common method, both in pK_a calculations and protein design,[48,67,116,117] the exact nature of the approximation is less clear than the previous two. Compared to the "bath" of valence electrons, it is a bit harder to view the protein backbone as an isotropic, homogeneous medium, uniformly packed throughout the protein interior, with simple, Gaussian fluctuations. Also, any backbone motions displace the side chains as well, so that the dielectric relaxation of the backbone continuum is, in fact, a mixed response that also involves side chain relaxation. Despite its extensive use, there has been little specific testing of this hybrid, "fixed-backbone + continuum" description for its ability to accurately capture the details of a protein's dielectric relaxation. The dynamics of protein backbones have been shown to have a rather harmonic character, if the longer surface loops are excluded,[25] and this is consistent with simple, Gaussian, polarization fluctuations and a linear dielectric response. Furthermore, one early study did analyze the dipolar fluctuations of the backbone of two proteins and compared them to the predictions of dielectric continuum theory, finding a reasonable agreement.[111] The magnitude of the backbone fluctuations was consistent with a dielectric constant of 2−3. In contrast, the fluctuations of ionized protein side chains have a much larger magnitude, consistent with a higher dielectric value of 20 or more.[112] More work is needed to fully test the fixed backbone description; for example, the protein response to perturbing charges could be computed from the hybrid model and compared to detailed, atomistic simulations.[118] For now, we simply view this hybrid treatment as empirically validated by the good agreement with experiment seen for computed pK_a's here and in related studies.[48,49,65,102]

5.2. Limitations of Continuum Models. Continuum models have specific limitations that are worth reviewing briefly, especially since they may or may not be alleviated by the use of a hybrid model. The first limitation is the use of a linear response approximation, which assumes that the electrostatic potential at the proton insertion site has simple, Gaussian

fluctuations.[119,120] This assumption has been tested indirectly for many systems, by comparing computed and experimental pK_a shifts, and it has been tested more directly for a few systems by analyzing electrostatic potentials from MD simulations.[119,33,50,121] Deviations from linear response were observed for a titratable site in thioredoxin[33,122] and for water close to multivalent ions, for example.[123] However, the contribution of a protein's backbone, which is of interest in our hybrid model, has never been analyzed separately. As noted above, flexible loops can have distinctly anharmonic and nonGaussian motions; however, for titratable positions close to such loops, the solvent will often dominate the protonation free energy, so that the backbone behavior is not too important.

A second limitation in most continuum models is the assumption of a homogeneous, isotropic dielectric behavior. Studies of dielectric relaxation in proteins using experiments[124,125] and simulations[110,118,126−130] have provided evidence for spatial inhomogeneity and anisotropy. However, in several cases, these effects could be mostly attributed to the side chains, especially the ionized side chains.[112,118,126,127] The electronic polarizability of the protein was shown to have a rather isotropic and homogeneous behavior,[118] although the early analyses should be repeated with current, superior, polarizability models.[131,132] Another indication that inhomogeneity arises largely from the side chains is the improved pK_a's obtained here and elsewhere[48,49,65,102] when side chain motions are modeled explicitly (or implicitly, but with an improved continuum scheme[99]).

A third, more fundamental limitation is the use of a smoothed, spatially averaged polarization density in the continuum model,[54,133] neglecting the atomic, granular structure of protein and solvent. The effects of granularity have been studied with more detailed, dipole lattice models.[134,135] They are illustrated by the effect of individual, buried waters on pK_a's and by the effect of large, local, conformational rearrangements of the protein.[136,137] Again, by modeling side chain motions explicitly, we expect that local effects can be treated more accurately than by a pure continuum scheme.

Finally, a fourth, serious limitation of continuum models involves the choice of the source charges, which are typically atomic point charges, often taken directly from a molecular mechanics force field. Thus, the smoothing just discussed is applied only to the polarization density and not to the source charges, unlike textbook continuum electrostatics models.[54,133] This choice of charges leads to a consistency problem, which has been analyzed in detail for several systems.[24,138,122,139] Indeed, the charges are usually optimized for use in combination with a low protein dielectric value, ε_p; for example, a molecular mechanics set is optimized for use with $\varepsilon_p = 1$. A low protein dielectric value is then appropriate to describe equilibrium potentials and fields.[139,140] Unfortunately, a low dielectric value is not always appropriate to describe the reorganization that occurs in response to a new charge, such as a titrating proton. Frequently, a larger value of $\varepsilon_p = 4$ or more is required. Thus, in cases where structural reorganization is sizable, continuum calculations using a single protein dielectric value can lead to very large errors for either the equilibrium potentials, the reorganization free energies, or both,[24,138,122,140] resulting in inaccurate pK_a's. This problem has also been analyzed at length by Warshel, Krishtalik, and co-workers, with conclusions that are consistent with ours.[140,32] One way to avoid the problem is to divide the protonation reaction into two distinct steps (following Hush, Marcus, and others), with one step corresponding to proton insertion into a fixed environment and the second step corresponding to relaxation of the environment.[16−19] The two steps can employ different protein dielectric constants, giving good results for several systems.[24,138,122] In the present work, instead, we use a hybrid model, where most of the protein reorganization is described explicitly (through side chain MC exploration); this should alleviate or eliminate the consistency problem.

5.3. Hybrid Model with Residue-GB. This paper has established a method to predict the acid/base behavior of proteins with a hybrid model that treats side chain motions explicitly. Monte Carlo simulations in a semigrand canonical ensemble yield ensembles of structures and protonation states as a function of pH. A large side chain rotamer library is used, allowing an accurate description of side chain rotamer conformational space.

A distinctive feature of our method is the use of a recent, residue-GB variant.[66] An obvious advantage is the increased speed of GB calculations, compared to PB. The initial, atomic-GB variant was parametrized earlier to match a large set of PB calculations.[84] We showed above that the residue-GB variant reproduces PB solvation and conformational free energies at least as well. This good performance is a bit surprising since no reparameterization was done relative to the original, atomic-GB. With residue-GB, the atomic radii are averaged (harmonically) over each residue's side chain or backbone. For the more exposed atoms, with the smaller atomic B's, the residue B will be somewhat increased; for the more buried atoms, the residue B will be somewhat decreased (relative to the atomic B). Earlier workers, using a very similar atomic-GB variant, noticed that the atomic B's were underestimated, compared to an "exact" Poisson calculation; increasing the B values empirically led to better agreement with PB.[141] It may be that the residue B values have the same effect. The more exposed atoms behave better, although the more buried atoms may behave worse; since the exposed atoms make the largest contribution to the solvation energy, the overall effect is an improvement. This remains to be analyzed in more detail.

The residue-pairwise character of the present GB variant allowed us to derive an efficient computational algorithm, inspired by protein design methodology, where matrices of interaction energies and of fitting coefficients are precomputed. As a result, during the Monte Carlo simulations, the fluctuating shape of the protein/solvent boundary is accounted for in a way that is essentially exact, within the GB framework. In particular, there is no need to treat the side chain fluctuations with a mean field approximation; rather, they are sampled rigorously according to the correct, Boltzmann distribution. Obviously, reproducing exactly the GB boundary treatment does not mean that one reproduces the true physical effects of the protein−solvent boundary. Nevertheless, our treatment contrasts with several other recent continuum electrostatic methods that use an effective, average boundary.[48,49,65,102,63,116,117] In fact, until now, the only method that gave an exact treatment of the protein/solvent boundary was constant-pH MD.

As a test set, we computed the pK_a's of 78 residues in 6 proteins for which experimental data are available. The multi-conformation GB approach (MC-GB) gives improved results compared to single-conformation PB and GB: modeling side chain flexibility explicitly improves the pK_a prediction. The good results were obtained using a physically reasonable dielectric constant of four for the protein. This is only slightly larger than the dielectric constant (around 2) estimated from MD simulations for the interior region of several proteins.[112,127,142,143] For these regions, the electronic polarizability and the motions of the backbone polar groups account for most of the dielectric constant. Furthermore, the MC-GB performance does not depend

much on the precise dielectric value, with $\varepsilon_p = 4$ and 8 giving similar results. Importantly, the performance is equally good for pK_a's with large and small shifts. In contrast, for single-conformation PB, the best performance is obtained with a large (and questionable) protein dielectric of 20 or even 80, and the performance for highly shifted pK_a's is poor, as shown by many authors.[40,32]

There are several directions to systematically improve the method, so that further studies are needed. One aspect that should be explored is the use of a larger test set of proteins. Another is to use a still more detailed rotamer library. A third aspect that affects performance is the precise division of the protein atoms into groups, for the calculation of the residue solvation radii. Here, each side chain was treated as a single group, but other choices are possible. Indeed, the residue solvation radius is a harmonic average over the group, so that the more exposed atoms (with the smallest atomic solvation radii) tend to dominate the mean. In the case of a tyrosine, for example, a partly exposed hydroxyl will lead to a small residue solvation radius, even though most of the side chain is buried. This explains why the GB interaction energies are somewhat smaller with residue-GB than with atomic-GB, as shown in the original paper (see Figures 1 and 3 in ref 66). In the future, we will explore the possibility of splitting some of the large side chains into two groups, such as the benzene ring and the hydroxyl group in the tyrosine case.

A significant limitation of the model is the use of a single backbone conformation, with the backbone flexibility modeled implicitly. Using multiple backbone conformations would rapidly increase the computational cost. Indeed, for each backbone conformation, the side chain rotamers have a different meaning, and separate matrices must be computed. A simple implementation would lead to a linear increase with respect to the number N of backbone conformations, both in time and in memory requirements. On the other hand, allowing multiple backbone conformations for limited regions of the protein, such as flexible loops, would be less expensive. Another shortcoming that may be important is the electrostatic model with fixed atomic charges, where electronic polarizability is treated implicitly. More work is needed to determine whether this is a significant source of error. Finally, while nonpolar interactions are included within the protein, we have largely ignored nonpolar effects related to the solvent. In the future, these could be added to the implicit solvent model through surface area terms or other treatments.[42,84]

Overall, the method has a clear physical basis, and there are several directions to systematically improve it. In its present form, it already gives reasonable accuracy for pK_a's and should be a useful tool to help increase our understanding of protein electrostatics.

Acknowledgment. G.A. and S.P. acknowledge financial support through a PENEK reinforcement 0505/04 grant "Modification of the specificity of synthetase family proteins with biomolecular simulations". G.A. and T.S. acknowledge support through Cyprus-France ZENON bilateral cooperation funds (project "Protein design simulations with application to the redesign of the genetic code").

Supporting Information Available: Additional methods and results, and additional tables and figures. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Sorensen, S. P. L.; Hoyrup, M.; Hempel, J.; Palitzsch, S. *C. R. Trav. Lab. Carls.* **1917**, *12*, 68–163.
(2) Linderstrøm-Lang, K. *C. R. Trav. Lab. Carls.* **1924**, *15*, 1–2.
(3) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. *J. Biol. Chem.* **2009**, *284*, 13285–13289.
(4) Lehninger, A.; Cox, M.; Nelson, D. L. *Principles of Biochemistry*; Freeman: New York, 2008.
(5) Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*; Freeman: New York, 1999.
(6) Perutz, M. *Mechanisms of cooperativity and allosteric regulation in proteins*; Cambridge University Press: Cambridge, 1990.
(7) Kyte, J. *Structure in protein chemistry*; Garland Publishing: New York, 1995.
(8) Varadarajan, R.; Zewert, T. E.; Gray, H. B.; Boxer, S. G. *Science* **1989**, *243*, 69–72.
(9) Bendall, D. S., Ed. *Protein electron transfer*; BIOS Scientific Publishers, 1996.
(10) Simonson, T. *Rep. Prog. Phys.* **2003**, *66*, 737–787.
(11) Marcus, R. *Annu. Rev. Phys. Chem.* **1964**, *15*, 155–196.
(12) Warshel, A.; Parson, W. *Q. Rev. Biophys.* **2001**, *34*, 563–679.
(13) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. *Science* **2004**, *303*, 186–195.
(14) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.
(15) Huang, R. B.; Du, Q. S.; Wang, C. H.; Liao, S. M.; Chou, K. C. *Protein Eng.* **2010**, *23*, 35–42.
(16) Hush, N. *Trans. Faraday Soc.* **1961**, *57*, 557–580.
(17) Marcus, R. *J. Chem. Phys.* **1956**, *24*, 979–989.
(18) Warshel, A. *J. Phys. Chem.* **1982**, *86*, 2218–2224.
(19) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430–437.
(20) Tanford, C.; Kirkwood, J. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.
(21) Warshel, A.; Russell, S. *Q. Rev. Biophys.* **1984**, *17*, 283–342.
(22) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045–1066.
(23) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219–10225.
(24) Simonson, T.; Archontis, G.; Karplus, M. *J. Phys. Chem. B* **1999**, *103*, 6142–6156.
(25) Brooks, C. L.; Karplus, M.; Pettitt, M. *Adv. Chem. Phys.* **1987**, *71*, 1–259.
(26) Becker, O., Mackerell, A., Jr., Roux, B., Watanabe, M., Eds. *Computational Biochemistry & Biophysics*; Marcel Dekker: New York, 2001.
(27) Warshel, A. *Computer modelling of chemical reactions in enzymes and solutions*; John Wiley: New York, 1991.
(28) Field, M. J. *A Practical Introduction to the Simulation of Molecular Systems*; Cambridge University Press: New York, 2007.
(29) Mulholland, A. *Drug Discovery Today* **2005**, *10*, 1393–13402.
(30) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368–8372.
(31) Sham, Y.; Chu, Z.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458–4472.
(32) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44*, 400–417.
(33) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.
(34) Ghosh, N.; Cui, Q. *J. Phys. Chem. B* **2008**, *112*, 8387–8397.
(35) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20227–20232.
(36) Ji, C. G.; Mei, Y.; Zhang, J. Z. H. *Biophys. J.* **2008**, *95*, 1080–1088.
(37) Delepierre, M.; Dobson, C.; Karplus, M.; Poulsen, F.; States, D.; Wedin, R. *J. Mol. Biol.* **1987**, *197*, 111–130.
(38) Havranek, J.; Harbury, P. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11145–11150.
(39) Warwicker, J.; Watson, H. *J. Mol. Biol.* **1982**, *157*, 671–679.
(40) Antosiewicz, J.; McCammon, J.; Gilson, M. *J. Mol. Biol.* **1994**, *238*, 415–436.
(41) Schaefer, M.; Vlijmen, H. W. T. v.; Karplus, M. *Adv. Protein Chem.* **1998**, *51*, 1–57.
(42) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
(43) Gilson, M.; Davis, M.; Luty, B.; McCammon, J. *J. Phys. Chem.* **1993**, *97*, 3591–3600.
(44) Cramer, C.; Truhlar, D. *Chem. Rev.* **1999**, *99*, 2161–2200.
(45) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.
(46) You, T.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721–1733.
(47) Beroza, P.; Case, D. A. *J. Phys. Chem.* **1996**, *100*, 20156–20163.
(48) Georgescu, E. R.; Alexov, E.; Gunner, M. *Biophys. J.* **2002**, *83*, 1731–1748.
(49) Kim, J.; Mao, J.; Gunner, M. *J. Mol. Biol.* **2005**, *348*, 1283–1298.
(50) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523–544.
(51) Börjesson, U.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.
(52) Lee, M.; Salsbury, F., Jr.; Brooks, C., III *Proteins* **2004**, *56*, 738–752.

(53) Mongan, J.; Case, D. A.; McCammon, J. *J. Comput. Chem.* **2004**, *25*, 2038–2048.
(54) Fröhlich, H. *Theory of Dielectrics*; Clarendon Press: Oxford, 1949.
(55) Roux, B.; Beglov, D.; Im, W. *Simulation and theory of electrostatic interactions in solution*; Pratt, L., Hummer, G., Eds.; American Institute of Physics, 1999; pp 473−491.
(56) David, L.; Luo, R.; Gilson, M. *J. Comput. Chem.* **2000**, *21*, 295–309.
(57) Still, W. C.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
(58) Hawkins, G. D.; Cramer, C.; Truhlar, D. *Chem. Phys. Lett.* **1995**, *246*, 122–129.
(59) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.
(60) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.
(61) Bashford, D.; Case, D. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
(62) Feig, M.; Brooks, C. L., III *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
(63) Wisz, M. S.; Hellinga, H. *Proteins* **2003**, *51*, 360–377.
(64) Zollars, E. S.; Marshall, S. A.; Mayo, S. L. *Protein Sci.* **2006**, *15*, 2014–2018.
(65) Barth, P.; Alber, T.; Harbury, P. B. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4898–4903.
(66) Archontis, G.; Simonson, T. *J. Phys. Chem. B* **2005**, *109*, 22667–22673.
(67) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82–87.
(68) Resat, H.; Mezei, M. *J. Am. Chem. Soc.* **1994**, *116*, 7451–7452.
(69) Woo, H. J.; Dinner, A.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392–6400.
(70) Collins, M. D.; Hummer, G.; Quillin, M. L.; Matthews, B. W.; Gruner, S. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16668–16671.
(71) Hill, T. *Introduction to Statistical Thermodynamics*; Addison-Wesley: Reading, Massachussetts, 1962.
(72) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.
(73) Simonson, T. *J. Phys. Chem. B* **2000**, *104*, 6509–6513.
(74) Allen, M.; Tildesley, D. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1991.
(75) Frenkel, D.; Smit, B. *Understanding molecular simulation*; Academic Press: New York, 1996.
(76) Sham, Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744–1753.
(77) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(78) Guérois, R., Lopez de la Paz, M., Eds. *Protein Design: Methods And Applications*; Humana Press: Totowa, NJ, 2007.
(79) Schmidt am Busch, M.; Lopes, A.; Mignon, D.; Simonson, T. *J. Comput. Chem.* **2008**, *29*, 1092–1102.
(80) Brünger, A. T. *X-plor version 3.1, A System for X-ray crystallography and NMR*; Yale University Press: New Haven, 1992.
(81) Press, W.; Flannery, B.; Teukolsky, S.; Vetterling, W. *Numerical Recipes*; Cambridge University Press: Cambridge, 1986.
(82) Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1267.
(83) Moulinier, L.; Case, D. A.; Simonson, T. *Acta Cryst. D* **2003**, *59*, 2094–2103.
(84) Lopes, A.; Aleksandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. *Proteins* **2007**, *67*, 853–867.
(85) Swanson, J.; Adcock, S.; McCammon, J. *J. Chem. Theory Comput.* **2005**, *1*, 484–493.
(86) Wagner, F.; Simonson, T. *J. Comput. Chem.* **1999**, *20*, 322–335.
(87) Brünger, A. T.; Karplus, M. *Proteins* **1988**, *4*, 148–156.
(88) Bashford, D. *Scientific Computing in Object-Oriented Parallel Environments, volume 1343 of Lecture Notes in Computer Science*; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer: Berlin, 1997; pp 233−240.
(89) Bashford, D. *Front. Biosci.* **2004**, *9*, 1082–1099.
(90) Beroza, P.; Case, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804–5808.
(91) Madura, J.; Briggs, J.; Wade, R.; Davis, M.; Luty, B.; Ilin, A.; Antosiewicz, J.; Gilson, M.; Baheri, B.; Scott, L.; McCammon, J. *Comput. Phys. Commun.* **1995**, *91*, 57–95.
(92) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192–2198.
(93) Onufriev, A.; Case, D. A.; Ullmann, M. *Biochemistry* **2001**, *40*, 3413–3419.
(94) Ullmann, M. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.
(95) Bombarda, E.; Ullmann, M. *J. Phys. Chem. B* **2010**, *114*, 1994–2003.
(96) Sondergaard, C. R.; McIntosh, L. P.; Pollastri, G.; Nielsen, J. E. *J. Mol. Biol.* **2008**, *376*, 269–287.
(97) Warwicker, J. *Protein Sci.* **1999**, *8*, 418–425.
(98) Khandogin, J.; Brooks, C. L. *Biochemistry* **2006**, *45*, 9363–9373.
(99) Warwicker, J. *Protein Sci.* **2004**, *13*, 2793–2805.
(100) Spassov, V.; Yan, L. *Protein Sci.* **2008**, *17*, 1955–1970.
(101) Harris, T. K.; Turner, G. J. *IUBMB Life* **2002**, *53*, 85–98.
(102) Kieseritzky, G.; Knapp, E. W. *Proteins* **2008**, *71*, 1335–1348.
(103) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289–298.
(104) Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. *Acta Crystallogr.* **1983**, *B39*, 480.
(105) Berndt, K. D.; Guntert, P.; Orbons, L. P.; Wüthrich, K. *J. Mol. Biol.* **1992**, *227*, 757–775.
(106) Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. *Proteins* **2002**, *48*, 388–403.
(107) Kirkwood, J. *J. Chem. Phys.* **1939**, *7*, 911–919.
(108) Simonson, T. *Int. J. Quantum Chem.* **1999**, *73*, 45–57.
(109) Simonson, T.; Perahia, D.; Brünger, A. T. *Biophys. J.* **1991**, *59*, 670–90.
(110) King, G.; Lee, F.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366–4377.
(111) Smith, P.; Brunne, R.; Mark, A.; van Gunsteren, W. *J. Phys. Chem.* **1993**, *97*, 2009–2014.
(112) Simonson, T.; Perahia, D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 1082–1086.
(113) Lleontyev, I. V.; Stuchebrukov, A. A. *J. Chem. Phys.* **2009**, *130*, 085102.
(114) Leontyev, I. V.; Rostov, M. V. V.; Basilevsky, M. V.; Newton, M. D. *J. Chem. Phys.* **2003**, *119*, 8024.
(115) Xin, W. D.; Juffer, A. H. *J. Comput. Phys.* **2007**, *223*, 416–435.
(116) Vizcarra, C. L.; Mayo, S. L. *Curr. Opin. Chem. Biol.* **2005**, *9*, 622–626.
(117) Vizcarra, C. L.; Zhang, N. G.; Marshall, S. A.; Wingreen, N. S.; Zeng, C.; Mayo, S. L. *J. Comput. Chem.* **2008**, *29*, 1153–1162.
(118) Simonson, T.; Perahia, D. *J. Am. Chem. Soc.* **1995**, *117*, 7987–8000.
(119) Simonson, T. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6544–6549.
(120) Song, X.; Chandler, D.; Marcus, R. A. *J. Phys. Chem.* **1996**, *100*, 11954–11959.
(121) Nilsson, L.; Halle, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13867–13872.
(122) Archontis, G.; Simonson, T. *Biophys. J.* **2005**, *88*, 3888–3904.
(123) Sandberg, L.; Edholm, O. *J. Chem. Phys.* **2002**, *116*, 2936–2944.
(124) Hass, M.; Jensen, M.; Led, J. *Proteins* **2010**, *48*, 6482–6494.
(125) Hernandez, G.; Anderson, J. S.; LeMaster, D. M. *Biochemistry* **2009**, *48*, 6482–6494.
(126) Simonson, T.; Perahia, D. *Faraday Discuss.* **1996**, *103*, 71–90.
(127) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118*, 8452–8458.
(128) Voges, D.; Karshikoff, A. *J. Chem. Phys.* **1998**, *108*, 2219–2227.
(129) Song, X. *J. Chem. Phys.* **2002**, *116*, 9359–9363.
(130) Golosov, A.; Karplus, M. *J. Phys. Chem. B* **2007**, *111*, 1482–1490.
(131) Ponder, J.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27.
(132) Friesner, R. A. *Adv. Protein Chem.* **2006**, *72*, 79–104.
(133) Jackson, J. *Classical electrodynamics*; Wiley: New York, 1975.
(134) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211–217.
(135) Sandberg, L.; Edholm, O. *Biophys. J.* **2010**, *98*, 470–477.
(136) Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; Garcia-Moreno, B. E. *Biophys. J.* **2000**, *79*, 1610–1620.
(137) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; Garcia-Moreno, B. *Biophys. J.* **2007**, *92*, 2041–2053.
(138) Archontis, G.; Simonson, T. *J. Am. Chem. Soc.* **2001**, *123*, 11047–11056.
(139) Simonson, T. *Photosynth. Res.* **2008**, *97*, 21–32.
(140) Krishtalik, L.; Kuznetsov, A.; Mertz, E. *Proteins* **1997**, *28*, 174–182.
(141) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.
(142) Pitera, J.; Falta, M.; van Gunsteren, W. *Biophys. J.* **2001**, *80*, 2546–2555.
(143) Park, H.; Jeon, Y. H. *Phys. Rev. E* **2007**, *75*, 021916.
(144) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721–4729.
(145) Lee, T. W.; Qasim, M. A.; Laskowski, M.; James, M. N. *J. Mol. Biol.* **2007**, *367*, 527–546.
(146) Ramanadham, M.; Sieker, L. C.; Jensen, L. H. *Acta Crystallogr.* **1990**, *B46*, 63–69.
(147) Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y. *Acta Crystallogr.* **1999**, *D55*, 386–398.
(148) Katti, S.; Lemaster, D.; Eklund, H. *J. Mol. Biol.* **1990**, *212*, 167.
(149) Nielsen, J. E.; Vriend, G. *Proteins* **2001**, *43*, 403–412.

JP104406X

Appendix D

Recognition of Ribonuclease A by Dinucleotide Inhibitors: An MD / Continuum Electrostatics Analysis

Recognition of Ribonuclease A by 3′–5′-Pyrophosphate-Linked Dinucleotide Inhibitors: A Molecular Dynamics/Continuum Electrostatics Analysis

Savvas Polydoridis,* Demetres D. Leonidas,† Nikos G. Oikonomakos,†‡ and Georgios Archontis*‡
*Department of Physics, University of Cyprus, Nicosia, Cyprus; and †Institute of Organic and Pharmaceutical Chemistry, ‡Institute of Biological Research and Biotechnology, the National Hellenic Research Foundation, Athens, Greece

ABSTRACT The proteins of the pancreatic ribonuclease A (RNase A) family catalyze the cleavage of the RNA polymer chain. The development of RNase inhibitors is of significant interest, as some of these compounds may have a therapeutic effect in pathological conditions associated with these proteins. The most potent low molecular weight inhibitor of RNase reported to date is the compound 5′-phospho-2′-deoxyuridine-3-pyrophosphate (P→5)-adenosine-3-phosphate (pdUppA-3′-p). The 3′,5′-pyrophosphate group of this compound increases its affinity and introduces structural features which seem to be unique in pyrophosphate-containing ligands bound to RNase A, such as the adoption of a syn conformation by the adenosine base at RNase subsite B_2 and the placement of the 5′-β-phosphate of the adenylate (instead of the α-phosphate) at subsite P_1 where the phosphodiester bond cleavage occurs. In this work, we study by multi-ns molecular dynamics simulations the structural properties of RNase A complexes with the ligand pdUppA-3′-p and the related weaker inhibitor dUppA, which lacks the 3′ and 5′ terminal phosphate groups of pdUppA-3′-p. The simulations show that the adenylate 5′-β-phosphate binding position and the adenosine syn orientation constitute robust structural features in both complexes, stabilized by persistent interactions with specific active-site residues of subsites P_1 and B_2. The simulation structures are used in conjunction with a continuum-electrostatics (Poisson-Boltzmann) model, to evaluate the relative binding affinity of the two complexes. The computed relative affinity of pdUppA-3′-p varies between −7.9 kcal/mol and −2.8 kcal/mol for a range of protein/ligand dielectric constants (ϵ_p) 2–20, in good agreement with the experimental value (−3.6 kcal/mol); the agreement becomes exact with $\epsilon_p = 8$. The success of the continuum-electrostatics model suggests that the differences in affinity of the two ligands originate mainly from electrostatic interactions. A residue decomposition of the electrostatic free energies shows that the terminal phosphate groups of pdUppA-3′-p make increased interactions with residues Lys^7 and Lys^{66} of the more remote sites P_2 and P_0, and His^{119} of site P_1.

INTRODUCTION

The pancreatic ribonuclease A (RNase A) family contains proteins, which decompose the RNA polymer chain (1,2). Many members of the family display pathological side effects. For example, human angiogenin (3) is implicated in cancer and in vascular and rheumatoid diseases (4); eosinophil-derived neurotoxin and eosinophil cationic protein are neurotoxic in vivo and are involved in hypereosinophilic syndromes and allergy (5); and bovine seminal RNase has antispermatogenic and immunosuppressive activity (6). The activity of these proteins is critically affected by mutations of residues involved in ribonucleolysis, or by ribonucleolytic inhibitors (3,7,8). Thus, the development of RNase inhibitors is of significant interest, as some of these compounds may act as therapeutic agents in pathological conditions associated with these proteins.

The RNase A catalytic site with a bound RNA molecule is shown schematically in Fig. 1. According to the established nomenclature (2), the active site is partitioned into subsites $\{B_i\}$ and $\{P_i\}$, which interact, respectively, with the RNA bases and phosphate groups. Subsite B_1 has a strong specificity for pyrimidine bases, conferred by residue Thr^{45}, which is strictly conserved in all RNases. The protein cleaves the P-O5′ (scissile) bond on the 3′ side of pyrimidine bases bound at B_1. Residues His^{12}, His^{119}, and Lys^{41} (subsite P_1) are strictly conserved among RNase homologs and play key roles in the reaction (2,9). Subsite B_2 recognizes all bases, with a preference for adenine. It is highly conserved, but other subsites are more variable among homologs.

The high degree of B_1, P_1, and B_2 homology suggests that inhibitors of one protein may also act against other members of the same family. Based on this expectation, structure-assisted inhibitor design studies have mainly focused on RNase A. Structural and kinetic studies have examined the complexes between RNase A and several mono- or dinucleotide inhibitors containing adenine (10–17) or inosine (18) at the 3′ position of the scissile bond, and uracil (13–17) or cytosine (11,12) at the 5′ position. The high-resolution structures of several complexes have provided a detailed picture of the inhibitor binding modes.

The most potent inhibitors (13–17) contain a nonstandard pyrophosphate group, which increases the affinity of nucleotide ligands by two orders of magnitude (14). Compounds

Submitted July 18, 2006, and accepted for publication November 1, 2006.

Address reprint requests to G. Archontis, E-mail: archonti@ucy.ac.cy.

Abbreviations used: dUppA, 2′-deoxyuridine-3-pyrophosphate (P→5) adenosine; pdUppA-3′-p, 5′-phospho-2′-deoxyuridine-3-pyrophosphate (P→5)-adenosine-3-phosphate; and ppA-2′-p, 5′-diphosphoadenosine 2′-phosphate; and ppA-3′-p, 5′-diphosphoadenosine 3′-phosphate.

0006-3495/07/03/1659/14 $2.00 doi: 10.1529/biophysj.106.093419

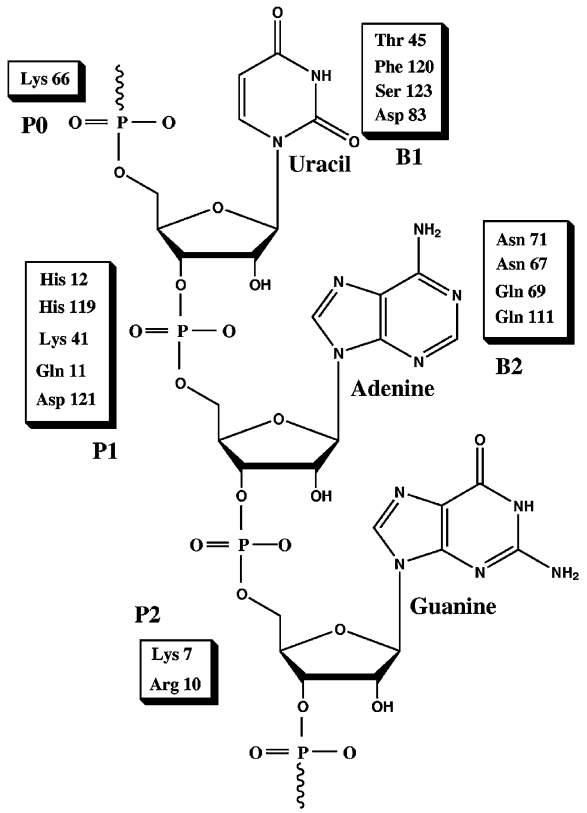


FIGURE 1 Schematic representation of the RNase A active site, with a bound RNA ligand. The subsites $\{B_i\}$ and $\{P_i\}$, interacting, respectively, with the RNA bases and phosphate groups are indicated. The figure is adapted from Raines (2).

ppA-3′-p ($K_i = 240$ nM) and ppA-2′-p ($K_i = 520$ nM) were the first inhibitors with this group (13,14). The compounds bind to RNase A with the β-phosphate at site P_1 and the adenosine χ angle in the syn range (13), structural features that are unique to pyrophosphate-containing ligands (15–17). Both inhibitors are strong, but do not make use of residues at subsites B_1 and P_0. To exploit potential interactions with these sites, a 2′-deoxy-5′-phosphoyuridine group was added to ppA-3′-p to create the new ligand pdUppA-3′-p. The uridine moiety of pdUppA-3′-p made numerous van der Waals and hydrogen-bonding interactions with the RNase A B_1 and the 5′-phosphate made medium-range Coulombic interactions with the Lys^{66} NZ group at the P_0 site; a hydrogen bond between the two groups could not be ruled out from the crystal structure (15). Compound pdUppA-3′-p is a ninefold stronger inhibitor than ppA-3′-p and the most potent RNase inhibitor to date, with a K_i of 27 nm. It is also effective against eosinophil-derived neurotoxin and RNase-4, with respective K_i values of 180 nM and 260 nM.

In this work we investigate for the first time by multi-nanosecond simulations in explicit water the dynamical behavior of RNase A complexes with two pyrophosphate compounds: 1), the most potent inhibitor pdUppA-3′-p; and 2), the related compound dUppA (16), which lacks the 3′, 5′ phosphate groups of pdUppA-3′-p and has a K_i of 11.3 μM for RNase A (16), corresponding to a weaker affinity by 3.6 kcal/mol. The chemical structures of the two molecules are shown in Fig. 2. The simulations reproduce essential features of the crystallographic structures, and provide insights on the flexibility of the bound ligands and the surrounding active-site residues.

The accurate computation of relative (19–26) or absolute (27–33) binding affinities has important applications in protein-ligand docking and design, and has attracted considerable computational effort in recent years. In this work, we evaluate the relative binding affinity of the two RNase A complexes by using the simulation conformations in conjunction with a continuum-electrostatics (Poisson-Boltzmann) approximation (34–37). This continuum model yields results in very good agreement with experiment, suggesting that the electrostatic interactions contribute mostly to the stability differences between the two complexes. We also evaluate the contribution to binding due to specific protein-ligand interactions and desolvation terms by a free-energy decomposition analysis of the Poisson-Boltzmann free energies (38,39).

In the next section, we describe the simulation methods and the continuum model. The results are given in the following section. The final section discusses the results and summarizes the conclusions.

THEORY AND METHODS

Molecular dynamics simulations

The crystallographic structures of the pdUppA-3′-p and dUppA complex have been determined at a resolution of 1.7 Å. The initial protein coordinates were taken from these structures (accession codes 1QHC (15) and 1JN4 (16)). The simulation system corresponded to a 29 Å sphere containing the entire protein, the ligand, and 2997 water molecules. The sphere was centered on the PB atom of the ligand. The water environment was created by retaining 135 crystallographic waters; additional water molecules were included by overlaying a preequilibrated 32 Å water sphere on the complex and deleting waters beyond 29 Å from the center of the complex, or overlapping with protein-heavy atoms or crystallographic-water oxygen atoms. This overlaying procedure was repeated 2–3 times during the equilibration of the two complexes.

Atomic charges, van der Waals, and force-field parameters corresponded to the CHARMM22 all-atom force field (40). The water was reproduced by a modified TIP3P water model (41). Electrostatic interactions were calculated by use of a multipole approximation (extended electrostatics) for groups >14 Å apart (42); van der Waals interactions were switched to zero at distances >12 Å. Water molecules were restrained to a spherical region of 29 Å radius by the stochastic boundary method (43). Water oxygen atoms in a buffer beyond 23 Å from the center were subjected to random and frictional forces mimicking a thermal bath at 293 Kelvin (44). Protein heavy atoms beyond 16 Å from the center were harmonically restrained, based on the crystallographic B-factors. Bond lengths with hydrogen atoms, and the geometry of water molecules were constrained by the SHAKE algorithm (45). The classical equations of motion were integrated by a Verlet algorithm modified for Langevin dynamics, using a time step of 1 fs.

To minimize structural perturbations in the molecular dynamics (MD) simulations due to the omission of the bulk solvent beyond the simulation

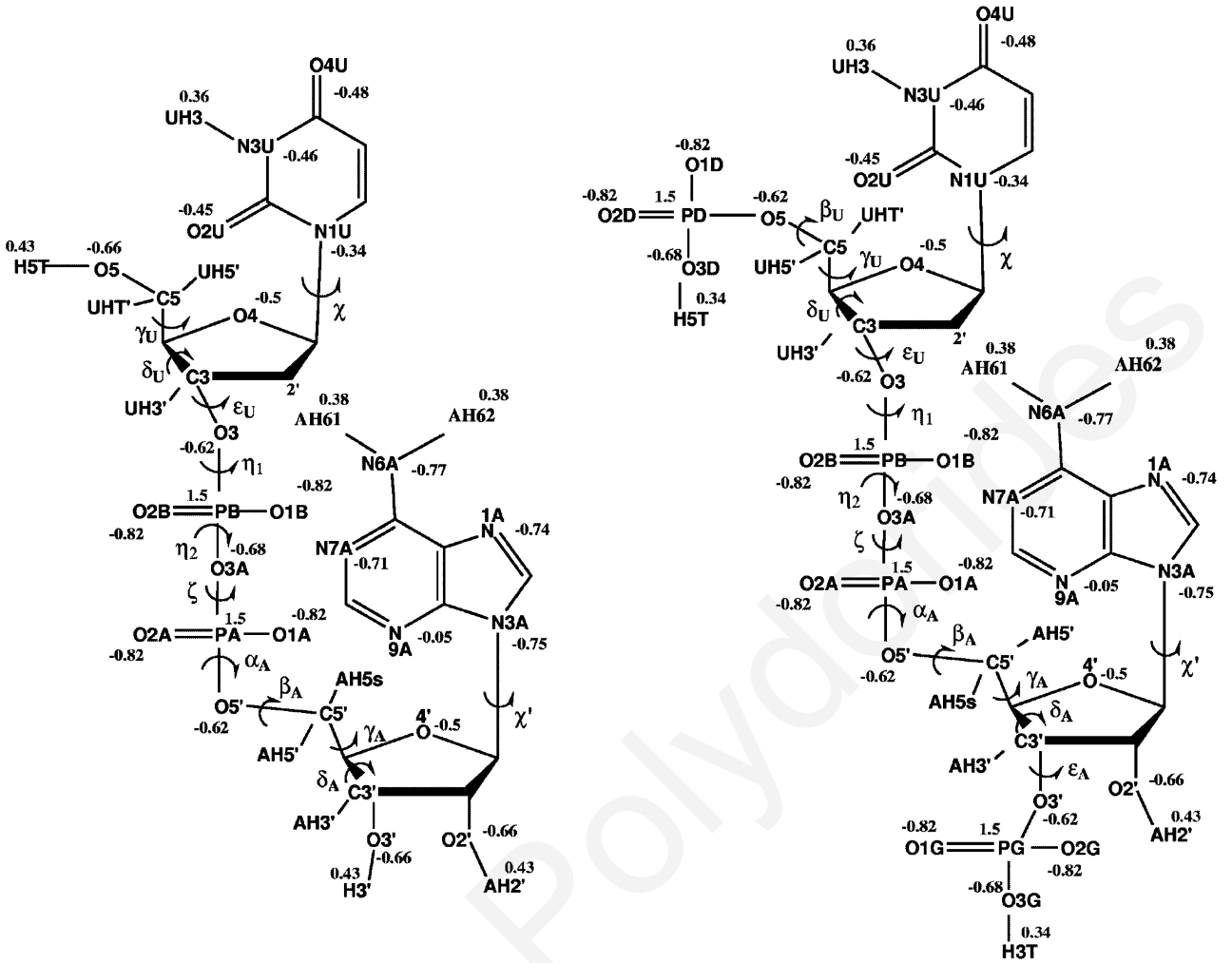


FIGURE 2 Chemical structures of the compounds dUppA (*a*) and pdUppA-3′-p (*b*). The atomic names and charges and the torsional angles discussed in the text are indicated.

sphere, the atomic charges of selected residues were scaled by a robust inhomogeneous reaction-field methodology, presented in Simonson et al. (46) and tested in detail in the aspartyl-tRNA synthetase system in the literature (46–48). All residues treated by this scheme did not participate in salt bridges and were located at least 20 Å away from the central phyrophosphate group of either ligand, and at least 14.5 Å from the nearest terminal (5′) phosphate group of pdUppA-3′-p. The scaled residues (scaling factors) were Lys^{31}(4.31), Lys^{61}(4.15), Arg^{85}(2.08), Lys^{91}(4.63), Lys^{98}(2.85), and His^{105}(2.39). The charges of side-chain atoms in these residues were divided by the corresponding scaling factor. The resulting simulation system had a total charge ≈0 for the dUppA complex.

In the crystallographic structures of the two complexes (15,16) there is no electronic density associated with bound ions. Thus, the structural features of the two complexes are not expected to depend on specific interactions with solution ions, located at well-defined positions. In the MD simulations, the solution screening of electrostatic interactions was taken into account by the scaling factors described above, without explicit counterions. In the Poisson-Boltzmann (PB) calculations (below), the average ionic effect is taken into account by setting the ionic concentration to the experimental ionic strength of 0.2 M.

The simulations had a first, 400-ps equilibration phase. In the subsequent 4-ns production phase, the protein and solvent coordinates were saved every 1 ps for analysis. All simulations were performed with the CHARMM biomolecular program (49), Ver. c28b1.

Poisson-Boltzmann calculations

To calculate the continuum-electrostatics free energies of binding for the various protein-ligand complexes, we evaluated the electrostatic potentials for the protein alone, the ligand alone, and the protein-ligand complex by solving the finite-difference Poisson equation on a three-dimensional grid. For the three states (protein, ligand, complex) the same three-dimensional grid and the same coordinates were used, so that the molecules were localized in the same positions on the grid. This allowed artificial contributions to the potential arising from the grid to be subtracted out exactly (50).

Finite-difference Poisson calculations were done for multiple (usually 200) structures taken from the 4 ns molecular dynamics trajectories of the two complexes. Averaging over multiple structures was expected to improve the precision and accuracy of the FDPB results (19,51). For each structure, the Poisson-Boltzmann equation was solved first on a large, coarse grid (0.8 Å spacing, 72–80 Å side), with screened Coulombic potentials on the grid boundary, and then on a finer grid (0.4 Å spacing, 56–64 Å side) with the coarse solution as boundary conditions (focusing method (50)). The sizes of

the coarser grid were chosen so that a high-dielectric layer of minimal width 20 Å would surround the complex. The dielectric constant of the solvent was set to 80. These of the protein and ligand were taken to be equal; values of 1–20 were compared. Atomic radii and charges were taken from the CHARMM22 force field used in the MD simulations, with the exception of the hydrogen radii, which were set to 1.0 Å. The protein-solvent boundary was defined by the molecular surface, constructed using a solvent probe radius of 2 Å. The ionic strength was set to the experimental value 0.2 M of monovalent counterions. Additional calculations were performed at 0 M, to check the dependence of the results on ionic concentration. All calculations were performed with the UHBD program (52).

Electrostatic free energy component analysis

The derivation of the electrostatic components has been presented in Archontis et al. (39) (see also (38)). We include it here, for the sake of convenience of the reader.

The electrostatic free energy of a protein-ligand complex (PL) is given by the expression (39,53)

$$G^{\mathrm{PL}} = \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{prot,lig}} q_{\mathrm{i}}V_{\mathrm{i}}^{\mathrm{PL}} = \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{prot}} q_{\mathrm{i}}V_{\mathrm{i}}^{\mathrm{PL}} + \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{lig}} q_{\mathrm{i}}V_{\mathrm{i}}^{\mathrm{PL}}, \qquad (1)$$

where q_{i} is the charge on atom i of the protein or ligand, and $V_{\mathrm{i}}^{\mathrm{PL}}$ is the electrostatic potential on atom i, in the solvated complex PL; analogous expressions yield the electrostatic free energies of the isolated protein (P) and ligand (L).

The total electrostatic potential on atom i inside the complex PL can be expressed as a sum over contributions from all ligand and protein atoms

$$V_{\mathrm{i}}^{\mathrm{PL}} = \sum_{\mathrm{j}\in\mathrm{lig}} V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} + \sum_{\mathrm{j}\in\mathrm{prot}} V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}}, \qquad (2)$$

with $V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}}$ the potential on atom i due to atom j in the complex PL. Using this decomposition and the reciprocity relation $q_{\mathrm{i}}V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} = q_{\mathrm{j}}V_{\mathrm{i}\to\mathrm{j}}^{\mathrm{PL}}$ (53), we arrive at the following expression for the electrostatic free energy of the complex:

$$G^{\mathrm{PL}} = \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{prot,j}\in\mathrm{prot}} q_{\mathrm{i}}V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} + \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{lig,j}\in\mathrm{lig}} q_{\mathrm{i}}V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} + \sum_{\mathrm{i}\in\mathrm{prot,j}\in\mathrm{lig}} q_{\mathrm{i}}V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}}. \qquad (3)$$

The electrostatic binding free energy of the complex PL, ΔG_{bind}, is equal to the difference between the free energies of the complex and the isolated protein and ligand in solution,

$$\Delta G_{\mathrm{bind}} = G^{\mathrm{PL}} - G^{\mathrm{P}} - G^{\mathrm{L}}. \qquad (4)$$

With the aid of Eq. 3, we obtain for ΔG_{bind} (38,39):

$$\begin{aligned}\Delta G_{\mathrm{bind}} &= \sum_{\mathrm{i}\in\mathrm{prot,j}\in\mathrm{lig}} q_{\mathrm{i}}V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} + \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{lig,j}\in\mathrm{lig}} q_{\mathrm{i}}[V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} - V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{L}}] \\ &\quad + \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{prot,j}\in\mathrm{prot}} q_{\mathrm{i}}[V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{PL}} - V_{\mathrm{j}\to\mathrm{i}}^{\mathrm{P}}] \\ &\equiv \Delta G_{\mathrm{int}}^{\mathrm{PL}} + \Delta G_{\mathrm{desolv}}^{\mathrm{L}} + \Delta G_{\mathrm{desolv}}^{\mathrm{P}}. \qquad (5)\end{aligned}$$

The first term on the right-hand side of the Eq. 5 is a free-energy component associated with the direct interaction between protein and ligand charges in the solvated complex. The second term corresponds to the ligand desolvation component, i.e., the change in the intraligand interactions upon binding, due to changes in the ligand geometry or charge distribution, as well as changes in the interaction of the ligand with polarization charge in the surrounding medium. If the ligand is assumed to have the same geometry and partial charges in the free and bound state (as here), this term arises entirely from ligand interactions with the polarization charge. The last term has an identical interpretation for the protein (38,39).

RESULTS

We first discuss the dynamical behavior of the two complexes and describe the important ligand-protein interactions observed in the simulations. We then evaluate the relative electrostatic binding free energy of the two complexes by a Poisson-Boltzmann approximation, and identify the protein residues, which mostly contribute to the stabilization of pdUppA-3′-p with respect to dUppA by a free-energy component analysis (38,39).

Simulation structures and interactions

Complex with dUppa

The Cartesian-coordinate root-mean-square (RMS) deviation from the initial conformation is plotted in Fig. 3 as a function of the simulation time. Note that $t = 0$ corresponds to the beginning of the production period, i.e., after 400 ps of equilibration. The left and right panels show, respectively, the dUppA and pdUppA-3′-p complex results. The total RMS deviation of the protein backbone heavy atoms (Fig. 3 *A*) is ≈0.65 Å at the end of the 4-ns production period (4.4 ns total time), showing that the protein conformation remains in the close vicinity of the crystal structure.

The ligand conformations can be described by a set of dihedral angles, defined in Fig. 2. The glycosyl dihedral angles χ and χ' describe, respectively, the orientation of the uracil and adenine rings, and the angles ϵ_{U}, η_1, η_2, ζ, and α_{A} characterize the conformation of the ligand pyrophosphate group. The time evolution of these dihedral angles is shown in Fig. 4 and the average values are listed in Table 1; for dihedrals, which fluctuate in more than one basin, we compute separately and report the average value in each basin.

The time series of the glycosyl dihedral angles χ and χ' for the dUppA ligand correspond to the solid curves in Fig. 4 *A*. The fluctuations in the χ' angle are very small, signifying that the χ' syn conformation is very stable throughout the simulation; this conformation is observed in all RNase complexes with the pyrophosphate-containing ligands dUppA (16), pdUppA-3′-p (15), and ppA-3′-p (13). The χ fluctuations are somewhat larger; however, the conformations of both adenine and uracil rings stay close to the initial (x ray) structure, with an RMS deviation of 0.7–0.8 Å at the end of the 4-ns production period (Fig. 3 *B*) and an average RMS positional fluctuation of ≈0.6 Å.

The pyrophosphate backbone is somewhat more flexible; the phosphate dihedrals η_1, η_2, ζ, and α_{A} fluctuate in the vicinity of the x-ray values and around different rotamers (Fig. 4 *B* and Table 1). The overall RMS positional fluctuation of the pyrophosphate atoms ranges between 0.45 Å and 0.85 Å. Atom PB has the smallest RMS fluctuation (0.45 Å), and a 0.6 Å RMS deviation from its initial position (Fig. 3 *C*). Thus, the PB position is very stable, in accord with the observation that all pyrophosphate-containing ligands bind with this phosphate at the P_1 site (13,15–17). The RMS

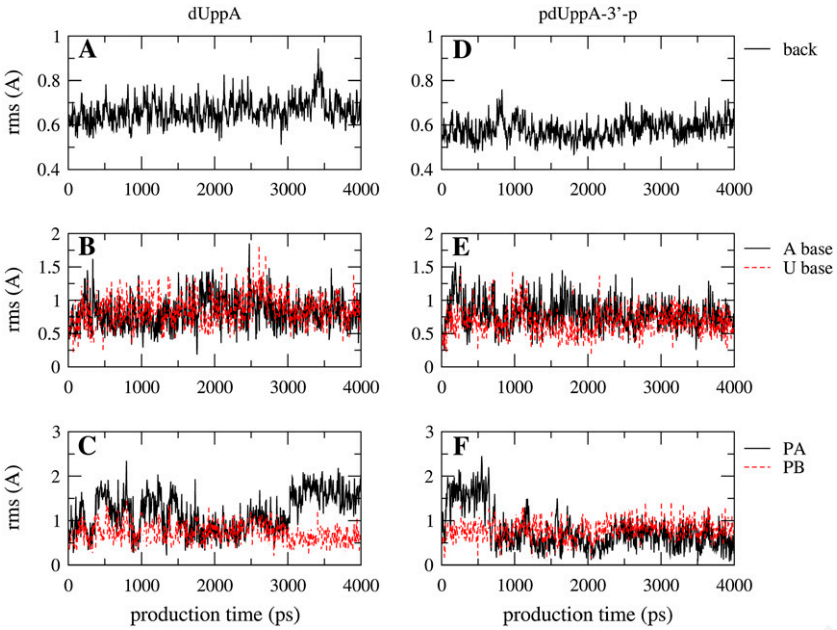


FIGURE 3 Root-mean-square deviation of selected groups of atoms from the starting conformation, plotted as a function of the simulation time. The $t = 0$ value corresponds to the end of the equilibration phase (400 ps). The results for complex dUppA are shown in plots *A–C*; those for pdUppA-3′-p are in plots *D–F*. Plots *A* and *D* are protein main chain heavy atoms; plots *B* and *E* are adenine and uracil ring atoms. Plots *C* and *F* are phosphate PA and PB atoms. The net rotation and translation has been removed, by orienting all trajectory frames with respect to the initial atomic coordinates of the protein backbone heavy atoms.

positional fluctuation of atom α-phosphate (PA) is larger (0.78 Å) and its RMS deviation from the initial position is ~1.6 Å at the end of the simulation. The observed abrupt changes in the PA RMS curve (Fig. 3 *C*) are due to transitions in the pyrophosphate dihedrals (Fig. 4 *B*). Thus, the PA position seems to be less stable. However, the PA atom forms interactions with important catalytic-site residues throughout the simulations, as we show below.

We next discuss the ligand interactions with protein residues and water molecules in the active site. The average distances between atoms participating in protein-ligand hydrogen bonds, along with the average hydrogen-bond lengths and occupancies are listed in Table 2. The time evolution of selected distances between pyrophosphate and protein atoms is plotted in Fig. 5.

In the crystal structures of RNase complexes with pyrophosphate-containing ligands (13,15,16), the ligand β-phosphate (PB) group interacts with the two catalytic histidines His^{12} and His^{119}, the proximal residues Gln^{11}, Phe^{120}, and a water molecule. In our simulations this group makes strong interactions with His^{12} and one or two water molecules, and somewhat weaker interactions with His^{119}, Phe^{120}, and Gln^{11}. In the first 2 ns, atom O1B interacts with His^{119} and Phe^{120} (see Fig. 5). Atom O2B forms a strong hydrogen bond with His^{12} and a weaker, direct or water-mediated interaction with Q11. At ≈2 ns, the pyrophosphate

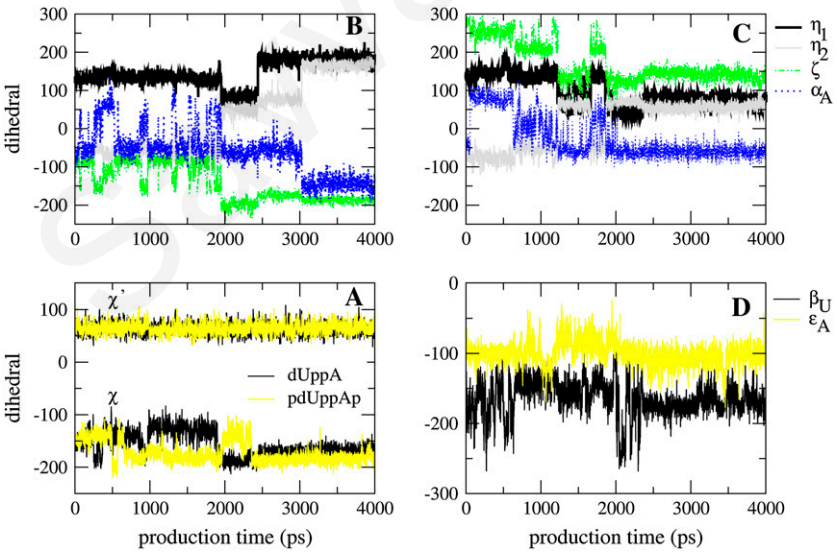


FIGURE 4 Time evolution of selected ligand dihedral angles in the simulations of the two complexes. The dihedral angles are defined in Fig. 2 and Table 1. (*A*) Dihedral angles χ' and χ. The black and yellow colors correspond, respectively, to the dUppA and pdUppA-3′-p ligand. (*B*) Ligand UppA pyrophosphate dihedral angles η_1, η_2, ζ, and α_A; (*C*) same as panel *B*, for ligand pdUppA-3′-p; and (*D*) terminal phosphate dihedrals β_U and ϵ_A of the ligand pdUppA-3′-p.

TABLE 1 Average values of selected ligand dihedral angles, observed in the simulations of the two complexes

Dihedral angle*	dUppA	X ray†	pdUppA-3′-p	X ray‡
Glycosyl dihedrals				
O4′U-C1′U-N1U-C2U(χ)	−135/−175§	−136	−182/−141	−130
O4′A-C1′A-N9A-4A(χ')	64	76	65	85
Phosphate dihedrals				
C4-C3-O3-PB(ϵ_U)	−173/91	176	−170	−82
C3-O3-PB-O3A(η_1)	80/133/180	121	69/142	106
O3-PB-O3A-PA(η_2)	−75/63/166	−82	−65/60	−71
PB-O3A-PA-O5′A(ζ)	−86/−179	−100	−102/−152/138	−95
O3A-PA-O5′-C5′(α_A)	−55/−145/49	−72	−57/79	−58
PA-O5′-C5′-C4′(β_A)	−173	−133	175	−136
PG-O3′-C3′-C4′(ϵ_A)			−100	−103
PD-O5′-C5′-C4′(β_U)			−165	−142
Backbone dihedrals				
O5′-C5′-C4′-C3′(γ_U)	−72/56	73	36/68	78
O5′-C5′-C4′-C3′(γ_A)	−48/41	51	−65/48	52
C5′-C4′-C3′-O3′(δ_U)	78/133	140	84/136	148
C5′-C4′-C3′-O3′(δ_A)	96	81	126	140

*The employed atom names and the corresponding dihedrals are shown in Fig. 2.
†From Leonidas et al. (15).
‡From Jardine et al. (16).
§For dihedrals which fluctuate in more than one basin (see Fig. 4), we compute separately and report the average value in each basin.

ζ dihedral angle undergoes a conformational transition (Fig. 4 *B*). Subsequently, O1B interacts with His^{12} and Phe^{120}; the Q11 interaction with both oxygens O1B and O2B is improved. The O1B and O2B atoms are also hydrogen-bonded to waters throughout the simulation (see Table 2). Atom O3 interacts with His^{119} in the last 1.5 ns.

The α-phosphate (PA) group interacts with residues Lys^7 and His^{119}. Atom O1A forms a strong hydrogen bond with Lys^7 in the first 3 ns. Atoms O2A and O3A interact with water molecules throughout the simulation and with His^{119} in the last 2 ns. For a brief period in the middle of the simulation, atom O2A makes an intramolecular hydrogen bond with O5 (not shown). A simulation structure of the active site region that is typical of the last 2-ns trajectory segment is shown in Fig. 6.

The His^{119} side chain can adopt two conformations A and B in the free RNase A, which differ by a rotation around the χ_1 angle (54). In the crystallographic structure of the complex, it is observed in conformation A (16). In the simulations it is retained in this conformation, with a mean χ_1 value of 152°. The His^{119} and adenine rings form continuous π–π stacking interactions, which presumably contribute to the stabilization of the His^{119} A orientation and the adenine ring syn orientation; the distance between the ring centers varies between ≈3.0 and 5.0 Å.

Residue Lys^{41} is located at a distance of 3.3 Å from atom O3 in the crystal structure. In the simulation, it forms water-mediated interactions with atoms O3 and the phosphate groups of the ligand, and a (noncontinuous) direct hydrogen bond for ~40% of the time with Gln^{11}. The positional fluctuation of its terminal NZ atom is 1.5 Å.

Thr^{45} confers to subsite B_1 of RNases the specificity for pyrimidine bases (2) by forming two hydrogen bonds with the uridine atoms N3U and O2U. In the simulations, both hydrogen bonds are almost continuously present (see Table 2). The uridine atom O4U forms indirect interactions with the Ser^{123} and the Asp^{83} side chains, mediated by one or two water molecules. Asp^{83} forms a second, continuous hydrogen bond with Thr^{45}. The uridine ring forms displaced stacking interactions with Phe^{120}. The adenine base forms two strong hydrogen bonds with Asn^{71} and weaker interactions with Gln^{69}. All these interactions are also observed in the crystal structure of the complex (16).

The ligand makes extensive interactions with the solvent (see Table 2). Each hydrogen-bonding atom of the ligand interacts with several hundred different water molecules in the course of the 4-ns simulation. The average (over all atoms) lifetime of these bonds is ≈2.3 ps; the largest average duration (5.9 ps) corresponds to water interactions with atom O2B.

Residues Arg^{10} (site P_2) and Lys^{66} (P_0) are located at distances >9 Å from the ligand phosphate groups and interact with water. Lys^{66} forms a salt bridge with Asp^{121} for ~0.5 ns. Its side chain undergoes frequent conformational transitions and the terminal NZ atom has a positional fluctuation of 2.9 Å. As we show below, this residue contributes to the stronger affinity of pdUppA-3′-p for RNase A.

Complex with pdUppA-3′-p

The time evolution of the Cartesian coordinates RMS deviation is shown in the right panel of Fig. 3. The simulation conformations remain close to the crystallographic structure. The total RMS deviation of the protein backbone heavy atoms (plot 3 *D*) and the ligand adenine and uracil rings (plot 3 *E*) converge to ≈0.6–0.7 Å at the end of the 4-ns production period, in close similarity with the behavior of dUppA complex. The glycosyl dihedral angles χ and χ' are maintained near the x-ray values (plot 4 *A*, *yellow curves*), indicating that the orientations of the two nucleotide rings are very stable. The β-phosphate (PB) group of the pyrophosphate linker occupies a stable position at the P_1 site, whereas the α-phosphate group (PA) is more mobile; the average positional fluctuations of the two phosphate atoms are, respectively, 0.41 Å and 0.81 Å. The pyrophosphate dihedrals η_1, η_2, ζ, and α_A undergo conformational transitions, as in the dUppA complex (Fig. 4 *C*). The two terminal 5′- and 3′-phosphate dihedrals β_U and ϵ_A (see Fig. 2) fluctuate mostly in a single basin (plot 4 *D*).

The average distances of atoms participating in ligand-protein hydrogen bonds, and the corresponding hydrogen-bond statistics are included in Table 2. The time evolution of selected distances is presented in Fig. 7. As in the dUppA complex, the β-phosphate group forms two strong hydrogen bonds with His^{12} and Phe^{120}, and three somewhat weaker

TABLE 2 Statistics of ligand-protein and ligand-water hydrogen bonds, observed in the simulations of the RNase A:dUppA and RNase A:pdUppA-3′-p complexes

	dUppA				pdUppA-3′-p			
	Distance		hb		Distance		hb	
Atom pair*	MD†	X ray	Length‡	Occupancy	MD†	X ray	Length‡	Occupancy
O1B-F120 (N)	3.4 (0.4)	3.2	2.9 (0.0)	52 (%)	3.0 (0.4)	2.9	3.0 (0.2)	82 (%)
O1B-H12 (NE2)	3.6 (0.1)		3.2 (0.1)	51	3.3 (0.9)		3.1 (0.2)	63
O1B-H119 (ND1)	3.5 (0.8)	3.1	2.7 (0.0)	44	3.5 (0.7)	2.8	2.6 (0.1)	28
O1B-water				25				
O2B-H12 (NE2)	3.5 (0.9)	2.8	2.7 (0.1)	49	3.4 (0.6)	2.4	2.7 (0.1)	37
O2B-K7 (NZ)					4.2 (1.7)		2.9 (0.2)	34
O2B-Q11 (NE2)	3.6 (0.6)	2.8	3.1 (0.1)	25	3.3 (0.5)	2.9	3.0 (0.2)	17
O2B-F120 (N)	4.6 (0.9)		3.2 (0.1)	8	4.8 (0.6)		3.2 (0.1)	2
O2B-water				138				100
				394§				**365**
O3-H119 (ND1)	3.9 (0.8)		3.0 (0.1)	29				
O3-Q11 (NE2)					3.8 (0.6)		3.3 (0.0)	15
O3-water				40				62
				70				**88**
O1A-K7 (NZ)	3.3 (0.9)		2.7 (0.1)	61	3.0 (0.7)	2.8	2.6 (0.0)	68
O1A-water				154				160
O2A-H119 (ND1)	4.2 (1.1)	3.0	3.0 (0.3)	28	3.3 (0.7)	3.4	3.2 (0.1)	58
O2A-water				198				148
O3A-H119 (ND1)	4.0 (0.8)		3.3 (0.1)	27	3.6 (0.8)		3.0 (0.2)	45
O3A-Q11 (NE2)	4.6 (0.8)		3.2 (0.1)	3	4.7 (0.8)		3.1 (0.2)	8
				484				**458**
O2D-K66 (NZ)					5.7 (2.4)		2.8 (0.1)	17
								17
O2G-K7 (NZ)					5.4 (2.1)		2.7 (0.1)	25
O3G-K7 (NZ)						2.7		
								25
N6A-N71 (OD1)	3.1 (0.4)	2.9	3.0 (0.3)	84	3.0 (0.2)	3.1	2.9 (0.2)	90
N6A-N67 (OD1)	3.6 (0.5)		3.0 (0.1)	21	3.8 (0.5)	3.3	3.2 (0.3)	17
				106				**109**
N7A-N71 (ND2)	3.1 (0.2)	2.9	3.1 (0.1)	91	3.1 (0.2)	3.2	3.1 (0.2)	92
N7A-water				14				25
				109				**120**
N1A-N67 (ND2)	3.3 (0.3)	3.1	3.0 (0.2)	26	3.3 (0.5)	3.4	3.2 (0.2)	33
N1A-water				37				44
				63				**77**
N3U-T45 (OG1)	2.9 (0.1)	2.9	2.8 (0.1)	100	2.8 (0.1)	2.8	2.8 (0.1)	100
				100				**100**
O2U-T45 (N)	3.0 (0.1)	2.8	3.0 (0.1)	83	3.0 (0.1)	2.8	3.0 (0.1)	89
				83				**89**
O4U-water				107				101

All distances in Å. The standard deviations are included in parentheses.

*The atomic names of the two ligands are indicated in Fig. 2.

†Average distance (with standard deviations in parentheses) between the corresponding heavy atoms in the entire 4-ns trajectory.

‡Average distance (with standard deviations in parentheses) in the portion of the 4-ns trajectory, where a hydrogen bond is formed. The criteria for the existence of a hydrogen bond employed here are 1), a maximum donor (D)-acceptor (A) distance of 3.4 Å; and 2), a minimum ∠ DHA of 120°.

§Total hydrogen-bond occupancy for each ligand group.

bonds with His^{119}, Gln^{11}, and Lys^{7}; the first four interactions are also present in the crystal structure (15). In the first 2 ns, atom O1B interacts with His^{119} and Phe^{120}, whereas O2B interacts with His^{12}. Both oxygens form hydrogen bonds with one or two water molecules, which sometimes bridge an interaction with Gln^{11}. The α-phosphate atoms O1A and O2A interact, respectively, with Lys^{7} and His^{119}; O3A interacts with Q11. A typical MD structure of the first 2-ns portion of the simulation is shown in Fig. 8. At $\approx$2 ns, the pyrophosphate ζ dihedral angle undergoes a conformational transition; subsequently, O1B interacts with His^{12}, Phe^{120}, and one or two waters, and O2B interacts with Lys^{7}, water, and Gln^{11}. The interaction between the α-phosphate (atoms O2A, O3A) and His^{119} is also improved.

Residue Lys^{41} forms a hydrogen bond with the ligand O3 atom in complex I of the crystallographic unit cell, with a

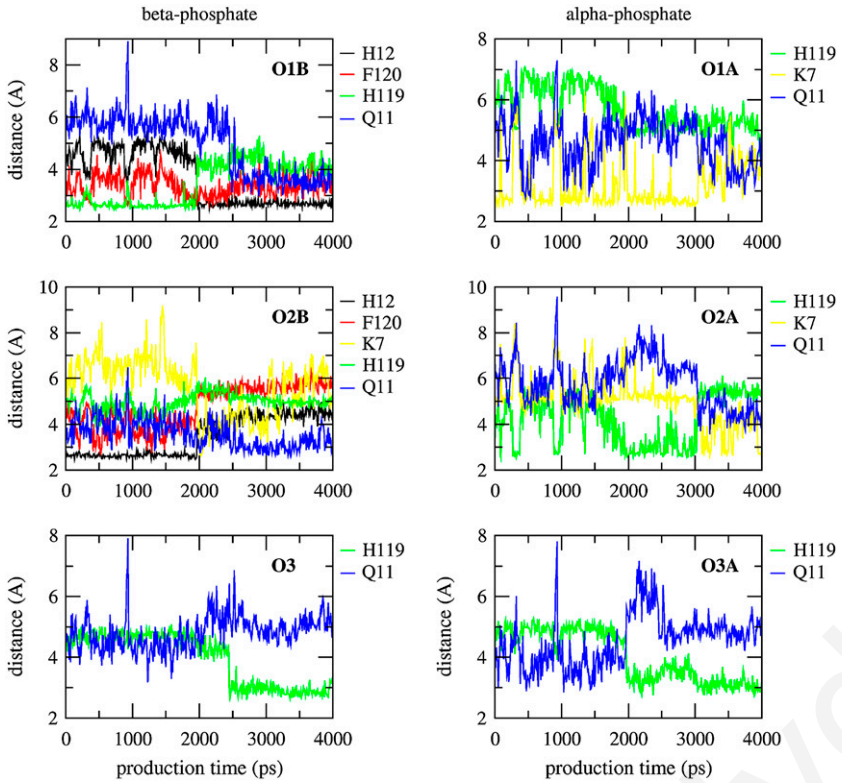


FIGURE 5 Time evolution of selected dUppA-protein distances in the 4-ns production-phase simulation.

length of 3.1 Å. In the simulations this hydrogen bond is observed in the first ≈500 ps (Fig. 7). Subsequently, Lys[41] forms water-mediated interactions with the two pyrophosphate groups as in the dUppA complex; the positional fluctuation of its terminal NZ atom is 1.5 Å.

Residue Lys[7] forms two hydrogen bonds with the PA and PG groups, with respective occupancies 70% and 25% (Table 2). This behavior is in accord with the crystallographic structure, where Lys[7] hydrogen-bonds to O1A in one of the two complexes in the crystallographic unit cell and to O3G in the second complex (15). The His[119] ring is maintained in conformation A with an average χ_1 angle of 153°, and forms continuous π–π stacking interactions with the adenine ring, as in the dUppA complex. Both residues contribute to the higher relative affinity of pdUppA-3′-p (see below).

The uridine and adenosine moieties of pdUppA-3′-p interact, respectively, with Thr[45] and Asn[71] via two strong hydrogen

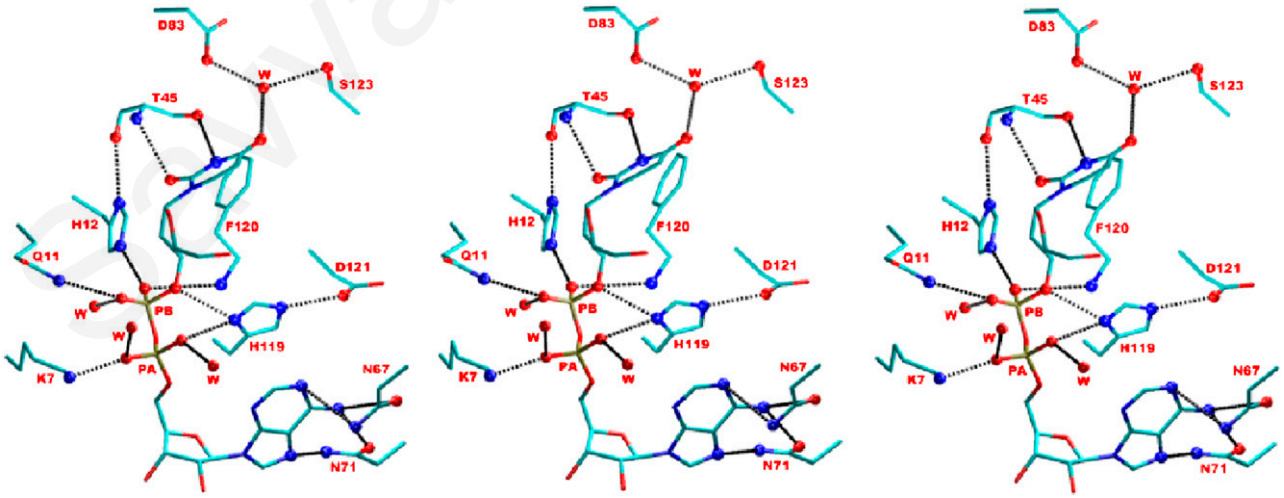


FIGURE 6 Wall (*left*) and cross (*right*) stereo representations of a typical structure of the dUppA-complex active site, observed in the second 2-ns simulation segment. The important interactions of dUppA with surrounding protein residues and waters are indicated.

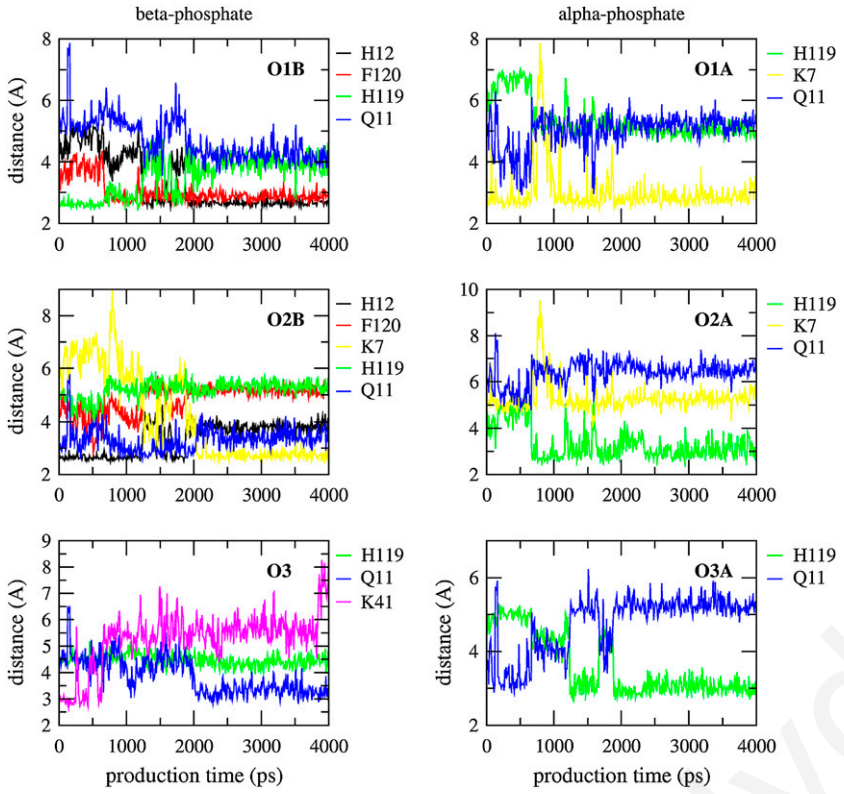


FIGURE 7 Time evolution of selected pdUppA-3′-p-protein distances in the 4-ns production-phase simulation.

bonds. The uridine ring makes off-centered stacking interactions with the Phe[120] ring. The adenine moiety interacts also with Asn[67] and Gln[69]. Ser[123] often makes water-mediated interactions with O4U and Asp[83]. Arg[10] is more remote (site P_2) and interacts mostly with water. All these interactions are in close agreement with the crystal structure (15) and similar to what was observed in the simulations of the dUppA complex.

The 5′ (PD) terminal phosphate interacts with water molecules throughout the run. In ≈75% of the simulation length it makes an intramolecular hydrogen bond with the oxygen O2A of the α-phosphate (not shown). In the last ns it also interacts on and off with Lys[66] of subsite P_0. This lysine interacts also with solvent molecules throughout the simulation. The positional fluctuation of the Lys[66] NZ atom is 3.2 Å, slightly larger than in the dUppA complex (2.9 Å). In the crystal structure, the position of the same atom is well defined in molecule I and its distance from the nearest oxygen of the PD phosphate is ≈4.7 Å. In molecule II of the

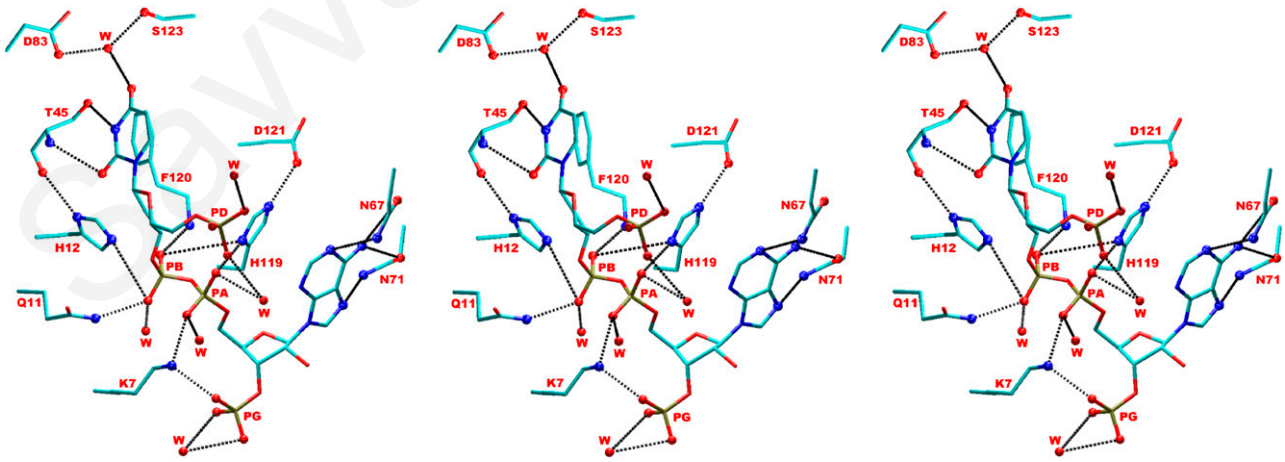


FIGURE 8 Wall (*left*) and cross (*right*) stereo representations of a typical structure of the pdUppA-3′-p -complex active site, observed in the first 2-ns simulation segment. The important interactions of dUppA with surrounding protein residues and waters are indicated.

unit cell there is no density beyond the C_β atom, suggesting that Lys^{66} is flexible. Even though Lys^{66} does not make strong interactions with the ligand, its contribution in the higher stability of the dUppA-3′-p complex is significant, as we show in the next section.

As in the dUppA complex, the pdUppa-3′-p ligand makes numerous hydrogen-bonding interactions with the solvent (see Table 2). Atom O2B hydrogen-bonds with 11 different waters and forms the longest-living interactions (with average lifetime of 16.1 ps). Other ligand atoms interact typically with several hundred different water molecules; when averaged over all ligand atoms, the mean water-ligand hydrogen bond lifetime is 3.1 ps.

Poisson-Boltzmann electrostatic association free energies

Based on the experimental K_i values of the pdUppA-3′-p and dUppA ligands (27 nM and 11.3 μM), the pdUppA-3′-p complex is estimated to be more stable by −3.6 kcal/mol. The net charge of the two ligands is −4 (pdUppA-3′-p) and −2 (dUppA); that of RNase is +8 at the experimental pH 5.5 of the solution (15,16). Thus, electrostatic interactions are expected to be the most important factor affecting the stability of the two complexes.

The electrostatic association free energies of the two complexes can be calculated by a Poisson-Boltzmann model (34–37). Furthermore, a component analysis (38,39) can identify which protein or ligand groups contribute mostly to the total free energy, and provide a better understanding of the differences between the two complexes.

The total Poisson-Boltzmann binding free energies for the two complexes are included in Table 3. To increase accuracy (39,51,55), the results were averaged over 200 snapshots, spanning the 4 ns simulations. The ionic strength of the solution corresponded to the experimental value of 0.2 M monovalent ion concentration. Additional calculations were performed at 0 M, to investigate the dependence of the results on the ionic concentration.

TABLE 3 Total electrostatic binding energies for the two complexes (kcal/mol)

	dUppA		pdUppA-3′-p		$\Delta\Delta G_{bind}$	
ϵ_p	0 M	0.2 M	0 M	0.2 M	0 M	0.2 M
2	−17.5 (4.0)	−14.1 (4.0)	−28.1 (3.5)	−22.0 (3.3)	−10.6	−7.9
4	−14.3 (2.0)	−11.0 (2.0)	−22.1 (2.1)	−16.1 (2.0)	−7.8	−5.1
8	−12.3 (1.0)	−9.1 (1.0)	−18.7 (1.4)	−12.8 (1.3)	−6.4	−3.7
16	−10.9 (0.5)	−7.8 (0.5)	−16.6 (1.0)	−10.7 (0.9)	−5.7	−2.9
20	−10.4 (0.4)	−7.4 (0.4)	−16.0 (0.9)	−10.2 (0.8)	−5.6	−2.8

Averages over 200 conformations spanning the 4 ns simulations. Values are calculated with a water dielectric constant of 80 and a protein/ligand constant of ϵ_p. The uncertainty (in parentheses) is computed as the standard deviation of averages, which are obtained by partitioning the 4 ns trajectory into four groups of 1 ns. The ionic strength of the experimental binding measurements is 0.2 M.

The binding free energies of both complexes are negative, implying that the electrostatic interactions promote association. The pdUppA-3′-p complex is more stable by 7.9 kcal/mol ($\epsilon_p = 2$) to 2.8 ($\epsilon_p = 20$) kcal/mol at the experimental (0.2 M) ionic strength. Thus, the PB results are in good agreement with the experimental value of 3.6 kcal/mol for $\epsilon_p > 2$; notably, the experimental value is exactly reproduced with a dielectric constant $\epsilon_p = 8$.

At 0 M, the association free energies are somewhat more negative. This increase in the affinity of oligonucleotide ligands for RNase A at low salt concentrations has been established experimentally (56,57).

Poisson-Boltzmann free energy component analysis

The relative electrostatic binding free energy obtained by PB is in good agreement with experiment. To gain insight on the interactions that increase the affinity of pdUppA-3′-p, we separate the total electrostatic free energy of both complexes into interaction and desolvation components. The interpretation of these components was presented in Theory and Methods. For a detailed discussion, see Archontis et al. (39).

The results are included in Table 4. The interaction term ΔG^{PL}_{int} is related to the direct electrostatic interaction energy between protein and ligand charges in the solvated complex. For both complexes it is negative, reflecting the fact that the direct electrostatic interactions between protein and ligand charges favor association. Its (absolute) value is larger in the pdUppa-3′-p complex (−70.0 kcal/mol compared to −54.1 kcal/mol) due to the higher charge of this inhibitor.

The protein and ligand desolvation terms are related to the change in the interaction of the protein (or ligand) with the polarization charge of the surrounding medium, upon binding. Both terms are positive, reflecting the fact that the protein and ligand interact more strongly with the induced charge in their unbound state, where they are surrounded by the high-dielectric solvent. The protein desolvation component is somewhat more positive for the large inhibitor, in accord

TABLE 4 Decomposition of the total electrostatic free-energies of Table 3 into interaction and desolvation components (kcal/mol)

	dUppA	pdUppA-3′-p	$\Delta\Delta G$
ΔG^{PL}_{inter}	−54.0 (2.1)	−70.0 (3.7)	−15.9
ΔG^{P}_{desolv}	20.0 (0.6)	22.7 (0.8)	2.7
ΔG^{L}_{desolv}	23.0 (0.6)	31.2 (1.1)	8.2
Total*	−11.0 (4.5)	−16.1 (5.4)	−5.1

The various components were explained in Theory and Methods. Values correspond to a protein/ligand dielectric constant of 4 and an ionic strength of 0.2 M. The uncertainty (in parentheses) is computed as the standard deviation of averages, which are obtained by partitioning the 4 ns trajectory into four groups of 1 ns.
*The total values are taken from Table 3.

with the larger reduction in the solvent-accessible surface area of RNase A in the simulations of the pdUppA-3′-p complex (418 Å^2, compared to 372 Å^2 in the dUppA complex, evaluated by the Lee-Richards algorithm with a probe radius of 1.4 Å). The ligand desolvation component is significantly more positive for pdUppA-3′-p (31.2 kcal/mol, compared to 23.0 kcal/mol). This can be attributed to the additional two phosphate groups, which increase the total interaction between pdUppA-3′-p and the induced charge at the ligand-solvent interface. The reduction in the solvent-accessible surface area of pdUppA-3′-p in the simulations is 664 Å^2, compared to 624 Å^2 for dUppA.

From the above analysis it follows that the increased relative stability of the pdUppA-3′-p complex originates mainly from enhanced electrostatic interactions between the pdUppA-3′-p ligand and the protein, reflected by the more negative component $\Delta G_{\text{int}}^{\text{PL}}$. Insight on the protein residues, which mostly contribute to the increased affinity of pdUppA-3′-p, can be obtained by further decomposing the term $\Delta G_{\text{int}}^{\text{PL}}$ into residue contributions. The largest values are included in Table 5. The meaning of these components is further analyzed in the next section and in Archontis et al. (39).

The most important contributions are due to residues Lys^7, His^{119}, and Lys^{66}. In Table 6 we decompose further these residue terms into contributions from interactions with various ligand moieties. Lys^7 forms stronger interactions with the pyrophosphate group and the 3′-end moiety of pdUppa-3′-p, which contains the γ-phosphate. This is in accord with the hydrogen-bond analysis presented above. Residue Lys^7 makes 1.27 hydrogen bonds (per frame) with the atoms O1A, O2B, and O2G of pdUppA-3′-p, compared to 0.61 hydrogen bonds with atom O1A in dUppa. The next most important residues His^{119} and Lys^{66} interact more strongly with the 5′-end (δ-phosphate) and to a lesser extent with the intermediate pyrophosphate group of pdUppa-3′-p. His^{119} forms, respectively, 1.31 and 1.28 hydrogen bonds (per frame) with the pyrophosphate atoms O1B, O3, O2A, O2B of ligands pdUppA-3′-p and dUppA; the average H119(Nδ1)–O3A distance is somewhat smaller (3.0 Å, compared to 3.3 Å) in the pdUppA-3′-p complex, possibly contributing to a stronger interaction.

TABLE 5 Contributions from selected residues to the protein-ligand interaction component $\Delta G_{\text{inter}}^{\text{PL}}$ (kcal/mol)

Residue	dUppA	pdUppA-3′-p	$\Delta\Delta G_{\text{inter}}^{\text{R}}$
Lys^7	−9.3 (0.6)	−17.9 (1.1)	−8.6
His^{119}	−18.4 (0.3)	−25.4 (0.6)	−7.0
Lys^{66}	−0.8 (0.4)	−4.3 (0.4)	−3.5
His^{12}	−19.6 (0.1)	−21.0 (0.0)	−1.3
Lys^{41}	−5.2 (0.6)	−5.5 (1.9)	−0.2
Asp^{121}	2.8 (0.1)	5.9 (0.3)	3.1
Total*	−54.1 (2.1)	−70.0 (3.7)	−16.1

Values are calculated with a protein/ligand dielectric constant of 4 and ionic strength of 0.2 M. The uncertainty (in parentheses) is computed as the standard deviation of averages, which are obtained by partitioning the 4 ns trajectory into 4 groups of 1 ns.
*Total value, listed in the first row of Table 4.

TABLE 6 Decomposition of the residue-ligand relative interaction free energies of Table 5 (term $\Delta\Delta G_{\text{inter}}^{\text{R}}$) into contributions from ligand moieties (kcal/mol)

Ligand moiety	Lys^7	His^{119}	Lys^{66}	His^{12}	Lys^{41}
5′-end*	−0.5	−4.1	−2.7	−0.9	−0.7
Pyrophosphate†	−5.3	−1.7	−0.7	−0.1	0.8
3′-end‡	−2.9	−0.9	−0.1	−0.6	−0.3
Total§	−8.6	−7.0	−3.5	−1.4	−0.3

All values are relative to the dUppA complex and correspond to a protein/ligand dielectric constant of 4, a solvent dielectric constant of 80 and an ionic strength of 0.2 M.
*The 5′-end includes atoms H5T, O5, C5, HC51, HC52 (ligand dUppA) and H5T, O3D, PD, O1D, O2D, O5, C5, HC51, HC52 (ligand pdUppA-3′-p) (see Fig. 2).
†The pyrophosphate moiety includes atoms C3, HC3, O3, PB, O1B, O2B, O3A, PA, O1A, O2A, O5′, C5′, HC5′1, HC5′2 for both ligands.
‡The 3′-end moiety includes atoms C3′, HC3′, O3′, H3′ (ligand dUppA) and C3′, HC3′, O3′, PG, O1G, O2G, O3G, H3T (ligand pdUppA-3′-p).
§Total value, listed in the last column of Table 5.

DISCUSSION AND CONCLUSIONS

The Coulombic forces exerted by various RNase A subsites on RNA analog substrates have been investigated by mutagenesis experiments. According to these studies, Lys^{66} (site P_0) and the pair Lys^7/Arg^{10} (site P_2) contribute, respectively, 0.9 kcal/mol and 1.2 kcal/mol to the binding of fluorescein-dAUAA (56) at a pH of 6.0; His^{12} and His^{119} (site P_1) contribute 1.4 kcal/mol and 1.1 kcal/mol to the binding of 3′-UMP at the same pH (58). These values are much smaller than the corresponding residue free-energy components $\Delta G_{\text{int}}^{\text{PL}}$ (Table 5). However, it should be noted that a PB interaction free-energy component associated with residue *R* (as the ones reported in Table 5) does not correspond quantitatively to the total free energy change of the complex due to neutralization of *R* (even though it does provide a qualitative measure of the R contribution to the total binding affinity). Two other factors are likely to partly compensate for this interaction free-energy component, yielding a smaller total free-energy change. First, the neutralization of *R* changes the polarization charge that is induced at the entire protein/solvent interface (dissociated state) or complex/solvent interface (bound complex state), modifying thereby the total protein desolvation free-energy component. Furthermore, the charge perturbation on *R* could cause structural relaxation in the complex, which might perturb the interaction components of other residues and change the total protein or ligand desolvation components. A detailed discussion of the connection between the PB free-energy components and mutagenesis, and specific numerical examples for complexes of the protein aspartyl-tRNA synthetase, are presented in Archontis et al. (39).

In addition to the electrostatic free energies computed here, the total absolute binding free energies contain contributions from other factors, such as the nonpolar interactions, the translational, conformational and vibrational entropy loss of the ligand and protein, and the entropy gain due to the release of water molecules (23,28–33,59–61).

An estimate of the nonpolar contribution to the relative binding affinity can be obtained by a standard surface model (62). Given that the average reduction in the total solvent-accessible surface area of the pdUppA-3′-p and dUppA complexes in the simulations is 1082 Å^2 and 996 Å^2, respectively, the model yields an additional stabilization of the pdUppA-3′-p complex by ≈0.5 kcal/mol.

Contributions from ligand/protein entropic terms could also be estimated by various approximate methods (23,28–33,59–61). For example, the two additional dihedral angles of pdUppA-3-′p, PD-O5′-C5′-C4′(β_U), and PG-O3′-C3′-C4′(ϵ_A), fluctuate in a single minimum throughout the 4-ns simulations (see Table 1 and Fig. 4 *D*). Assuming an average loss of entropy per rotatable bond of 0.4–0.5 kcal/mol (31,63), the restriction of these angles upon association destabilizes the pdUppA-3-′p complex by ≈0.8–1.0 kcal/mol, with respect to dUppA. Additional contributions due to the ligand translational/rotational and the protein side-chain entropy are likely to cancel approximately in the relative free energy difference, because the two ligands are similar, bind at the same position/orientation in the complex and have similar positional fluctuations in the simulations. We thus do not compute them here.

Other factors that could introduce inaccuracies in the continuum model are the use of charges/radii optimized for molecular mechanics calculations and the use of uniform dielectric constants for the protein/ligand and solvent, despite the contrary evidence from simulations (37,64–68). Furthermore, the protein and ligand structures are assumed identical in the complex and the dissociated states in our calculations. The structural fluctuations of the complex are taken into account by averaging the results over the 4-ns trajectories, but any protein/ligand relaxation upon dissociation is only implicitly taken into account via the dielectric constant (36,55,69–71). Given these approximations and the appreciable ligand sizes (68 and 73 atoms), it is noteworthy that our continuum model yields relative binding free energies in excellent agreement with experiment for $\epsilon_p = 8$, and fair agreement in the entire range $\epsilon_p = 2$–20.

Ligand-design methods based on PB calculations of single (e.g., energy-minimized) conformations may be successful in ranking a family of complexes, provided their docking/energy minimization protocols produce representative conformations. An example of such a high-throughput docking/continuum electrostatics method, which discovered a set of β-secretase inhibitors, is presented in Huang et al. (25). In the case of the two RNase A complexes considered here, the x-ray conformations are indeed sufficient for the accurate evaluation of the relative binding affinity; PB calculations with the x-ray structures yield a result of −5.3 kcal/mol in favor of the pdUppa-3′-p ligand (with a protein/ligand dielectric constant $\epsilon_p = 4$ and an ionic strength of 0.2 M), in perfect agreement with the corresponding simulation average (−5.1 kcal/mol; see Table 3). On the other hand, MD simulations are likely to provide more representative structures in some systems, e.g., when a ligand or protein mutation is associated with structural relaxation (26,39). The computational requirements of such MD-based ligand design methods may be reduced by using a more approximate free-energy scoring function, such as in the linear interaction energy approach (20,26).

In conclusion, in this work we have studied by molecular dynamics simulations and Poisson-Boltzmann calculations the complexes between RNase A and the two dinucleotide inhibitors dUppA and pdUppa-3′-p. The simulations reproduce structural features which are characteristic of complexes between RNases and pyrophosphate-containing nucleotidic inhibitors (13–17); in particular, the β-phosphate group of each ligand interacts with the P_1 site residues His^{12} and His^{119}, Gln^{11}, and Phe^{120} (Table 2), in agreement with the crystal structures (15,16). The β-phosphate position is very stable, with an RMS deviation from the crystal structure of 0.6 Å at the end of the 4.4 ns simulation (Fig. 3) and an overall positional (RMS) fluctuation of 0.41–0.45 Å. The interaction with His^{12}, which is persistent throughout the simulations, presumably contributes to the stability. The α-phosphates are more mobile, with a positional fluctuation of 0.78–0.81 Å, and interact mainly with Lys^7 and His^{119}. The syn orientation of the adenosine ring, which characterizes RNase complexes with pyrophosphate-containing ligands, is maintained throughout the simulations; it is stabilized in part by stacking interactions with His^{119}. The direct and water-mediated hydrogen-bonding interactions of the two bases with the RNase catalytic-site residues are well reproduced.

Continuum electrostatics calculations predict that the pdUppa-3′-p ligand binds to RNase A with a relative affinity that ranges between −7.9 kcal/mol (protein/ligand dielectric $\epsilon_p = 2$) and −2.8 kcal/mol ($\epsilon_p = 20$). The experimental relative value, −3.7 kcal/mol, is reproduced with an $\epsilon_p = 8$. The good agreement of the continuum model with the experimental relative affinities suggests that the electrostatic interactions are mainly responsible for the different stability of the two complexes.

The present computational study provides further insight into the remarkable differences in potency of the two inhibitors, which complements efficiently the previous crystallographic results (15,16). A free-energy decomposition shows that the increased affinity of RNase for the pdUppa-3′-p ligand is mainly due to interactions with Lys^7, His^{119}, and Lys^{66} (Table 5). The interactions of Lys^7 and Lys^{66}, respectively, with the 3′ and 5′ phosphate groups and the stabilizing role of these two groups in the overall conformation of pdUppA-3′-p, both inferred from the present study and from the crystal structures (15,16), seem to be the main

reason for the differences in the potency of the two inhibitors.

This work has been partly supported by a ''Joined Research and Technology Grant'' between Cyprus and Greece (to G.A., D.L., and N.G.O.), and PENEK grants for the ''Research Integration of Young Cypriot Researchers'' and the ''Research Reinforcement of Young Cypriot Researchers'' (to G.A. and S.P.).

All calculations were performed on a Linux cluster of the Biophysics group at the University of Cyprus.

REFERENCES

1. Cho, S., J. J. Bentema, and J. Zhang. 2005. The ribonuclease A superfamily of mammals and birds: identifying new members and tracing evolutionary histories. *Genomics*. 85:208–220.
2. Raines, R. T. 1998. Ribonuclease A. *Chem. Rev*. 98:1045–1065.
3. Riordan, J. F. 2001. Angiogenin. *Methods Enzymol*. 341:263–273.
4. Fett, J. W., D. J. Strydom, R. R. Lobb, E. Alderman, J. L. Bethune, J. F. Riordan, and B. L. Vallee. 1985. Isolation and characterization of angiogenin, an angiogenic protein from human carcinoma cells. *Biochemistry*. 24:5480–5486.
5. Gleich, G., and C. R. Adolphson. 1986. The eosinophilic leukocyte: structure and function. *Adv. Immunol*. 39:177–253.
6. Vescia, S., D. Tramontano, G. Augusti-Tocco, and G. D. Alessio. 1980. In vitro studies on selective inhibition of tumor cell growth by seminal ribonuclease. *Cancer Res*. 40:3740–3744.
7. Rosenberg, H. F., and J. B. Domachowske. 2001. Eosinophils, eosinophil ribonucleases and their role in host defense against respiratory virus pathogens. *J. Leukoc. Biol*. 70:691–698.
8. Boix, E. 2001. Eosinophil cationic protein. *Methods Enzymol*. 341: 287–305.
9. Lopez, X., D. M. York, A. Dejaegere, and M. Karplus. 2002. Theoretical studies on the hydrolysis of phosphate diesters in the gas phase, solution and RNase A. *Int. J. Quantum Chem*. 86:10–26.
10. McPherson, A., G. Brayer, and R. D. Morrison. 1986. Crystal structure of RNase A complexed with $d(pA)_4$. *J. Mol. Biol*. 189:305–327.
11. Zegers, I., D. Maes, M.-H. Dao-Thi, F. Poortmans, R. Palmer, and L. Wyns. 1994. The structures of RNase A complexed with 3′-CMP and dCpA: active site conformation and conserved water molecules. *Protein Sci*. 3:2322–2339.
12. Toiron, C., C. Gonzalez, M. Bruix, and M. Rico. 1996. Three-dimensional structure of the complexes of ribonuclease A with 2′,5′-CpA and 3′,5′-CpA in aqueous solution, as obtained by NMR and restrained molecular dynamics. *Protein Sci*. 5:1633–1647.
13. Leonidas, D. D., R. Shapiro, L. I. Irons, N. Russo, and K. R. Acharya. 1997. Crystal structures of ribonuclease A complexes with 5′-diphosphoadenosine 3′-phosphate and 5′-diphosphoadenosine 2′-phosphate at 1.7 Å resolution. *Biochemistry*. 36:5578–5588.
14. Russo, N., and R. Shapiro. 1999. Potent inhibition of ribonuclease A by 3′,5′- pyrophosphate-linked nucleotides. *J. Biochem*. (*Tokyo*). 274: 14902–14908.
15. Leonidas, D. D., R. Shapiro, L. I. Irons, N. Russo, and K. R. Acharya. 1999. Toward rational design of Ribonuclease inhibitors: high resolution structure of a ribonuclease A complex with a potent 3′,5′-pyrophosphate-linked dinucleotide inhibitor. *Biochemistry*. 38:10287–10297.
16. Jardine, A. M., D. D. Leonidas, J. L. Jenkins, C. Park, R. T. Raines, K. R. Acharya, and R. Shapiro. 2001. Cleavage of 3′,5′-pyrophosphate-linked dinucleotides by ribonuclease A and angiogenin. *Biochemistry*. 40:10262–10272.
17. Leonidas, D. D., G. B. Chavali, N. G. Oikonomakos, E. D. Chrysina, M. N. Kosmopoulou, M. Vlassi, C. Franking, and K. R. Acharya. 2003. High resolution crystal structures of ribonuclease A with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors. *Protein Sci*. 12:2559–2574.
18. Hatzopoulos, G. N., D. D. Leonidas, R. Kardakaris, J. Kobe, and N. G. Oikonomakos. 2005. The binding of IMP to Ribonuclease A. *FEBS J*. 272:3988–4001.
19. Massova, I., and P. Kollman. 1999. Computational alanine-scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc*. 121:8133–8143.
20. Aqvist, J., V. B. Luzhkov, and B. O. Brandsal. 2002. Ligand binding affinities from MD simulations. *Acc. Chem. Res*. 35:358–365.
21. Simonson, T., G. Archontis, and M. Karplus. 2002. Free-energy simulations come of age: protein-ligand recognition. *Acc. Chem. Res*. 35: 430–437.
22. Gouda, H., I. D. Kuntz, D. A. Case, and P. A. Kollman. 2003. Free energy calculations for theophylline binding to an RNA aptamer: comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers*. 68:16–34.
23. Swanson, J. M. J., R. H. Henchman, and J. A. McCammon. 2004. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J*. 86:67–74.
24. Archontis, G., K. A. Watson, Q. Xie, G. Andreou, E. Chrysina, S. E. Zographos, N. G. Oikonomakos, and M. Karplus. 2005. Glycogen phosphorylase inhibitors: a free energy perturbation analysis of glucopyranose spirohydantoin analogues. *Proteins Struct. Funct. Bioinf*. 61: 984–998.
25. Huang, D., U. Luthi, P. Kolb, K. Elder, M. Cecchini, S. Audetat, A. Barberis, and A. Caflisch. 2005. Discovery of cell-permeable non-peptide inhibitors of β-secretase by high-throughput docking and continuum electrostatics calculations. *J. Medicin. Chem*. 48:5108–5111.
26. Ersmark, K., M. Nervall, E. Hamelink, L. K. Janka, J. C. Clemente, B. M. Dunn, M. J. Blackman, B. Samuelsson, J. Aqvist, and A. Hallberg. 2005. Synthesis of malarial plasmepsin inhibitors and prediction of binding modes by molecular dynamics simulations. *J. Medicin. Chem*. 48:6090–6106.
27. Straatsma, T., and J. A. McCammon. 1991. Theoretical calculations of relative affinities of binding. *Methods Enzymol*. 202:497–511.
28. Hermans, J., and L. Wang. 1997. Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding: application to a complex of benzene and mutant T4 lysozyme. *J. Am. Chem. Soc*. 119:2707–2714.
29. Luo, R., and M. K. Gilson. 2000. Synthetic adenine receptors: direct calculation of binding affinity and entropy. *J. Am. Chem. Soc*. 122: 2934–2937.
30. Luo, H., and K. Sharp. 2002. On the calculation of absolute macromolecular binding free energies. *Proc. Natl. Acad. Sci. USA*. 99:10399–10404.
31. Lazaridis, T., A. Masunov, and F. Gandolfo. 2002. Contributions to the binding free energy of ligands to avidin and streptavidin. *Proteins Struct. Funct. Bioinf*. 47:194–208.
32. Boresch, S., F. Tettinger, M. Leitgeb, and M. Karplus. 2003. Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B*. 107:9535–9551.
33. Woo, H.-J., and B. Roux. 2005. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA*. 102:6825–6830.
34. Gilson, M. K., and B. H. Honig. 1987. Calculation of electrostatic potentials in an enzyme active site. *Nature*. 330:84–86.
35. Honig, B., and A. Nicholls. 1995. Classical electrostatics in biology and chemistry. *Science*. 268:1144–1149.
36. Warshel, A., and A. Papazyan. 1998. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol*. 8:211–217.
37. Simonson, T. 2003. Electrostatics and dynamics of proteins. *Rep. Prog. Phys*. 66:737–787.

38. Hendsch, Z., and B. Tidor. 1999. Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.* 8:1381–1392.

39. Archontis, G., T. Simonson, and M. Karplus. 2001. Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations; application to amino acid recognition by aspartyl-tRNA synthetase. *J. Mol. Biol.* 306:307–327.

40. Mackerell, D. A., D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkewicz-Kuczera, D. Yin, and M. Karplus. 1998. An all-atom empirical potential for molecular modeling and dynamics study of proteins. *J. Phys. Chem. B.* 102:3586–3616.

41. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

42. Stote, R., D. States, and M. Karplus. 1991. On the treatment of electrostatic interactions in biomolecular simulation. *J. Chem. Phys.* 88:2419–2433.

43. Brooks, III, C. L., and M. Karplus. 1983. Deformable stochastic boundaries in molecular dynamics. *J. Chem. Phys.* 79:6312–6325.

44. Brooks, III, C. L., A. T. Brünger, and M. Karplus. 1985. Active site dynamics in protein molecules: a stochastic boundary molecular-dynamics approach. *Biopolymers.* 24:843–865.

45. Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* 23:327–341.

46. Simonson, T., G. Archontis, and M. Karplus. 1997. Continuum treatment of long-range interactions in free energy calculations. application to protein-ligand binding. *J. Phys. Chem. B.* 101:8349–8362.

47. Archontis, G., T. Simonson, D. Moras, and M. Karplus. 1998. Specific amino acid recognition by aspartyl-tRNA synthetase studied by free energy simulations. *J. Mol. Biol.* 275:823–846.

48. Archontis, G., and T. Simonson. 2001. Dielectric relaxation in an enzyme active site: molecular dynamics simulations interpreted with a macroscopic continuum model. *J. Am. Chem. Soc.* 123:11047–11056.

49. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

50. Gilson, M. K., and B. H. Honig. 1988. Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.* 9:327–335.

51. van Vlijmen, H. W., M. Schaefer, and M. Karplus. 1998. Improving the accuracy of protein pKa calculations: conformational averaging versus the average structure. *Proteins Struct. Funct. Bioinf.* 33:145–158.

52. Madura, J., J. Briggs, R. Wade, M. Davis, B. Luty, A. Ilin, J. Antosiewicz, M. Gilson, B. Baheri, L. Scott, and J. McCammon. 1995. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics Program. *Comput. Phys. Comm.* 91:57–95.

53. Landau, L., and E. Lifschitz. 1980. Electrodynamics of Continuous Media. Pergamon Press, New York, NY.

54. Leonidas, D. D., T. K. Maiti, A. Samanta, S. Dasgupta, T. Pathak, S. E. Zographos, and N. G. Oikonomakos. 2006. The binding of 3-′-*n*-piperidine-4-carboxyl-3′-deoxy-ara-uridine to ribonuclease A in the crystal. *Bioorg. Med. Chem.* In press.

55. Archontis, G., and T. Simonson. 2005. Proton binding to proteins: a free energy component analysis using a dielectric continuum model. *Biophys. J.* 88:3888–3904.

56. Fisher, B. M., J.-H. Ma, and R. T. Raines. 1998. Coulombic forces in protein-RNA interactions: binding and cleavage by ribonuclease A and variants at Lys^{7}, Arg^{10} and Lys^{66}. *Biochemistry.* 37:12121–12132.

57. Fisher, B. M., L. W. Schultz, and R. T. Raines. 1998. Coulombic effects of remote subsites on the active site of ribonuclease A. *Biochemistry.* 37:17386–17401.

58. Park, C., L. W. Schultz, and R. T. Raines. 2001. Contribution of the active site histidine residues of ribonuclease A to nucleic acid binding. *Biochemistry.* 40:4949–4956.

59. Finkelstein, A. V., and J. Janin. 1989. The price of lost freedom: entropy of bimolecular complex formation. *Protein Eng.* 3:1–3.

60. Tidor, B., and M. Karplus. 1994. The contribution of vibrational entropy to molecular association—the dimerization of insulin. *J. Mol. Biol.* 238:405–414.

61. Gilson, M. K., J. A. Given, B. L. Bush, and J. A. McCammon. 1997. The statistical-thermodynamical basis for computation of binding affinities: a critical review. *Biophys. J.* 72:1047–1069.

62. Sitkoff, D., K. A. Sharp, and B. Honig. 1994. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* 98:1978–1988.

63. Doig, A. J., and M. J. E. Sternberg. 1995. Side-chain conformational entropy in protein folding. *Protein Sci.* 4:2247–2251.

64. Simonson, T., and D. Perahia. 1995. Microscopic dielectric properties of cytochrome-c from molecular dynamics simulations in aqueous solution. *J. Am. Chem. Soc.* 117:7987–8000.

65. Simonson, T. 1998. The dielectric constant of cytochrome-c from simulations in a water droplet including all electrostatic interactions. *J. Am. Chem. Soc.* 120:4875–4876.

66. Pitera, J. W., M. Falta, and W. F. van Gunsteren. 2001. Dielectric properties of proteins from simulations: the effects of solvent, ligands, pH, and temperature. *Biophys. J.* 80:2546–2555.

67. Stern, H. A., and S. E. Feller. 2003. Calculation of the dielectric permittivity profile of a nonuniform system: application to a lipid bilayer simulation. *J. Chem. Phys.* 118:3401–3412.

68. Ballenegger, V., and J. P. Hansen. 2005. Dielectric permittivity profiles of confined polar fluids. *J. Chem. Phys.* 122:114711.

69. Warshel, A., S. T. Russell, and A. K. Churg. 1984. Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability. *Proc. Natl. Acad. Sci. USA.* 81:4785–4789.

70. Schutz, C. N., and A. Warshel. 2001. What are the dielectric constants of proteins and how to validate electrostatic models. *Proteins Struct. Funct. Bioinf.* 44:400–417.

71. Eberini, I., A. Baptista, E. Gianazza, F. Fraternali, and T. Beringhelli. 2004. Reorganization in apo- and holo-β-lactoglobulin upon protonation of Glu^{89}: molecular dynamics and pK_a calculations. *Proteins Struct. Funct. Bioinf.* 54:744–758.