University of Cyprus

DEPARTMENT OF BIOLOGICAL SCIENCES

PROBING THE ROLE OF RARE CODONS IN CODING SEQUENCES OF ESCHERICHIA COLI

DOCTOR OF PHILOSOPHY DISSERTATION

ATHINA THEODOSIOU

2015

University of Cyprus

DEPARTMENT OF BIOLOGICAL SCIENCES

PROBING THE ROLE OF RARE CODONS IN CODING SEQUENCES OF ESCHERICHIA COLI

ATHINA THEODOSIOU

A Dissertation Submitted to the University of Cyprus in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

May 2015

VALIDATION PAGE

Doctoral Candidate: Athina Theodosiou

Doctoral Thesis Title: Probing the role of rare codons in coding sequences of *Escherichia coli*

The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the ***Department of Biological Sciences*** *and was approved on the ……by the members of the* ***Examination Committee.***

Examination Committee

Research Supervisor:_____

(Name, position and signature)

Committee Member:_____

(Name, position and signature)

Committee Member:_____

(Name, position and signature)

Committee Member:_____

(Name, position and signature)

Committee Member:_____

(Name, position and signature)

DECLARATION OF DOCTORAL CANDIDATE

The present doctoral dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes or any other statements.

…….[Full Name of Doctoral Candidate]

…….[Signature]

Abstract (In Greek)

Στα διάφορα είδη οργανισμών που υπάρχουν παρουσιάζονται ειδικές προτιμήσεις στη χρήση κωδικονίων που κωδικοποιούν το ίδιο αμινοξύ και αυτό αντανακλάται στη συχνότητα εμφάνισης των συνώνυμων κωδικονίων στο γονιδιωματικό DNA. Στους μονοκύτταρους οργανισμούς, είναι γενικά αποδεκτό, ότι η προτίμηση των κωδικονίων αντικατοπτρίζει την ισορροπία μεταξύ μεταλλάξεων και φυσικής επιλογής που μεγιστοποιεί την μετάφραση (Sharp και Li 1986). Πιο ειδικά, στην *Escherichia coli* έχει αποδειχθεί ότι η μη τυχαία επιλογή κωδικονίων, ως επί το πλείστο, οφείλεται στην διαθεσιμότητα του μεταφορικού RNA (tRNA) εντός του κυττάρου (Ikemura 1981a, Ikemura 1981b). Πρόσφατα, έχουν γίνει πειραματικές και υπολογιστικές μελέτες για να μπορέσει να ποσοτικοποιηθεί η μεταφραστική παύση που προκαλείται από ομάδες σπάνιων κωδικονίων (που ονομάζουμε Rare Codon Clusters–RCCs), επιδεικνύοντας μια πιθανή σχέση με το συν-μεταφραστικό δίπλωμα των πρωτεϊνών (Zhang, Hubalewska, Ignatova 2009). Η αποτελεσματικότητα της μετάφρασης μπορεί να μεγιστοποιείται σε τοπικό επίπεδο (ελαχιστοποιείται, αντίστοιχα), όταν κωδικόνια με άφθονα (ή σπάνια αντίστοιχα) συγγενή tRNAs βρίσκονται συγκεντρωμένα κατά μήκος των αντίστοιχων μορίων πληροφοριακού RNA (mRNA).

Η διαδικασία ανίχνευσης των RCCs ορίζεται ως ο προσδιορισμός των ομάδων κωδικονίων που αντιστοιχεί σε σπάνια tRNA ή σπάνια κωδικόνια κατά μήκος του mRNA. Βασικός στόχος της έρευνάς μας ήταν να ανακαλύψουμε και να περιγράψουμε το βιολογικό ρόλο της ύπαρξης των RCCs χρησιμοποιώντας πλήρη στοιχεία του γονιδιώματος της *E. coli* σε συνδυασμό με τις λειτουργικές και δομικές πληροφορίες που είναι διαθέσιμες. Εν συντομία, αναπτύξαμε νέες μεθόδους και εργαλεία λογισμικού για την ανίχνευση των RCCs και στη συνέχεια διερευνήσαμε τα πιθανά πρότυπα συσχετισμού των RCCs (για παράδειγμα παρουσία, απουσία, θέση) με δομικά και λειτουργικά χαρακτηριστικά των γονιδίων και των αντίστοιχων πρωτεϊνών.

Αναπτύξαμε ένα νέο και ευέλικτο διαδικτυακό διακομιστή (LaTcOm; Theodosiou and Promponas, 2012) μαζί με μια αυτόνομη έκδοση του λογισμικού, με στόχο να αντιμετωπίσουμε ελλείψεις των υφιστάμενων μεθόδων και να παρέχουμε επίσης νέα εργαλεία και δυνατότητες για την ανίχνευση των RCCs. Η συγκριτική ανάλυση που εφαρμόστηκε στο σύνολο των κωδικών γονιδίων της *E. coli* αποκάλυψε ότι δεν υπάρχει καμία σαφής συμφωνία μεταξύ των διαφόρων προσεγγίσεων. Παρόλα αυτά, η καλύτερη

θετική συσχέτιση βρέθηκε μεταξύ των εργαλείων %MinMax και MSS. Για να αποφευχθούν προβλήματα που σχετίζονται με τη χρήση κυλιόμενου παραθύρου, προτείνουμε πως το MSS μπορεί να χρησιμοποιηθεί εναλλακτικά για την ανίχνευση των RCCs.

Στην ανάλυση σχετικά με την κατανομή των RCCs επιβεβαιώσαμε πως υπάρχει προτίμηση στις περιοχές αυτές να βρίσκονται στα 5’ και 3’ άκρα των γονιδίων και επιπρόσθετα υπάρχει στατιστικά σημαντική διαφορά της κατανομής της απόστασης των πρώτων RCCs από το 5’ άκρο με την κατανομή των τελευταίων RCCs από το 3’ άκρο. Αυτό υποδεικνύει πιθανότατα πως υπάρχει διαφορετικός λειτουργικός ρόλος της ύπαρξης των RCCs στα δύο άκρα.

Επιπρόσθετα, βρήκαμε πως η παρουσία των RCCs σχετίζεται κυρίως με εκκρινόμενες πρωτεΐνες, με πρωτεΐνες που αλληλεπιδρούν με την κυτταρική μεμβράνη ή με το εξωτερικό κυτταρικό τοίχωμα (διαμεμβρανικές εσωτερικής μεμβράνης, διαμεμβρανικές εξωτερικού τοιχώματος (β-βαρέλια)). Από την άλλη, η απουσία των RCCs σχετίζεται κυρίως με κυτταροπλασματικές πρωτεΐνες, με πρωτεΐνες που σχετίζονται με το ριβόσωμα και με τον μεταβολισμό. Ακόμη, δείξαμε πως τα RCCs σχετίζονται με τις πρωτεΐνες με πολλαπλά αυτοτελή δομικά στοιχεία (domains) και προτείνουμε πως μπορούν να χρησιμοποιηθούν σαν ένδειξη των ορίων τους. Πιθανόν η καθυστέρηση της μετάφρασης σε αυτές τις θέσεις να είναι αναγκαία για το σωστό δίπλωμα αυτών των πρωτεϊνών.

Ακόμη μια κύρια πρωτότυπη προσπάθεια της εργασίας αυτής ήταν η συσχέτιση των RCCs με τοπολογικά και δομικά χαρακτηριστικά των διαμεμβρανικών α-ελικοειδών πρωτεϊνών (αHTMP) με πειραματικά επιβεβαιωμένες τρισδιάστατες δομές. Βρήκαμε πως τα RCCs βρίσκονται κατά προτίμηση σε περιπλασματικές περιοχές των αHTMP δείχνοντας πως υπάρχει μια σύνδεση μεταξύ της επιβράδυνσης τους ριβοσώματος και τις βιογένεσης των αHTMP. Προτείνουμε πως το σήμα αυτό σχετίζεται με τον μηχανισμό εισαγωγής και τοπολογίας των πρωτεϊνών αυτών στην μεμβράνη.

Τα αποτελέσματά μας τονίζουν την σημαντικότητα των RCCs σε συγκεκριμένες περιοχές των κωδικών γονιδίων της *E. coli*. Αναμένουμε πως θα αποτελέσουν πηγή έμπνευσης για περαιτέρω αναλύσεις βασικής έρευνας για την κατανόηση αυτών των μηχανισμών και θα προτείνουν προβλέψεις που θα αξιοποιηθούν στο μέλλον σε εφαρμογές βιοτεχνολογίας, όπως για παράδειγμα στο σχεδιασμό γονιδίων για έκφραση σε ετερόλογους οργανισμούς.

Abstract (In English)

The various species that exist show specific preferences for codons encoding the same amino acid (codon bias), reflected in the frequency of occurrence of synonymous codons in genomic DNA. In unicellular organisms, it is generally accepted that the preference of codons reflects a balance between mutational biases and natural selection for translational optimization (Sharp and Li 1986). In *Escherichia coli* it has been shown that the non-random choice of codons is mostly attributable to the availability of transfer RNA within a cell (Ikemura 1981a; Ikemura 1981b). Recently, experimental and computational advances have been made in quantifying translational pausing caused by rare codon clusters (RCCs), demonstrating a possible relation to co-translational protein folding (Zhang, Hubalewska, Ignatova 2009). Translational efficiency can be locally maximized (minimized, respectively) when codons with highly abundant (rare) cognate tRNAs are clustered along the respective mRNA molecules.

The RCC detection process is the identification of codon clusters corresponding to rare tRNA species or rare codons along mRNAs. Within this project our main research goal was to unravel possible roles for RCCs in *E. coli* using complete genome data combined with functional and structural information available in disparate resources. Briefly, we developed novel methods and tools for RCC detection and consequently investigate patterns correlating RCCs (presence/absence, position etc.) in *E. coli* genes/proteins with their structural and functional features.

We have implemented a novel flexible web server (LaTcOm; Theodosiou and Promponas 2012) along with a standalone version, aiming to address shortcomings of existing methods and to also provide new tools and features for RCC detection. The benchmarking we applied on the *E. coli* set revealed that there is no clear evidence of concordance between the different approaches. Nevertheless, the best positive correlation was found between %MinMax and MSS. To avoid window bias issues, we propose that MSS can be alternatively used for detecting rare codon clusters.

We confirm previous findings that RCCs are preferentially located at the 5' and 3' terminal sites and additionally demonstrate a statistically significant difference between the distribution of distances of the first RCCs from 5' terminals and the last RCCs from the 3' terminal site. RCCs were found to lay closer to the 5' terminal, than to the 3' terminal site possibly indicating a different functional role for their existence at the two sites.

Moreover, we analysed the RCC detection results for the complement of genes encoded in the *E. coli* genome and identified that the existence of RCCs is related with secreted, inner and outer membrane proteins (inner transmembrane and outer membrane β-barrels). Interestingly, we reveal that most of the sequences with no detected RCCs are found in the cytoplasm, are involved with the ribosome or with metabolic processes. In addition, we demonstrated that RCCs and multidomain proteins are associated well and we propose that RCC coordinates can be used as indicators for domain boundaries. We suggest that translation slowdown at these sites is necessary for correct protein folding.

Another main innovation of our project was our effort to correlate RCCs to topological and structural features of *E. coli* α-helical transmembrane proteins (αHTMP) with experimentally derived atomic structures. We demonstrate the preferential position of RCCs in periplasmic loops of αHTMPs, indicating that a coupling exists between RCC-mediated ribosomal attenuation and biogenesis of αHTMPs. We propose that the signal at the periplasmic region may be related to the insertion mechanism and topology of the transmembrane protein in the membrane.

Our results highlight the importance of RCCs at specific locations of coding genes in *E. coli*. We anticipate that these results will inspire further basic research towards understanding the fine details of these mechanisms and provide predictions that may be exploited in future biotechnological applications, as for example for rationally designing heterologous gene expression studies.

Acknowledgments

First of all I would like to thank my supervisor Dr. Vasilis Promponas for giving me the opportunity to perform this research, for guiding and most of all inspiring me thought this journey. I am grateful and very lucky for being part of his team. Additionally, a big thank you goes for my examination committee members Dr. Paris Skourides, Dr. Giorgos Apidιanakis, Dr. Eitan Rubin and Dr. Chris Christodoulou for reading and commenting on this PhD thesis.

I wish to thank the anonymous referees for invaluable comments on the LaTcOm manuscript and the LaTcOm functionality. I also thank Professor Walter Ruzzo (University of Washington) and Shane Neph (University of Washington) for help with the MSS source code, professor Zoya Ignatova (University of Potsdam) and Dr. Gong Zhang (University of Potsdam) for providing their tRNA scale, professor Konstantinos Fokianos (University of Cyprus) for useful discussions on the RCC validation procedure and Ioanna Kalvari (University of Cyprus) for helping with interfacing MSS with the LaTcOm modules.

Moreover, I would like to thank the University of Cyprus and the A.G. Leventis Foundation for supporting my PhD with scholarships.

During these five years, I have enjoyed valuable discussions and enjoyable moments with colleagues and friends in our lab. Therefore, I would like to thank Ioannis Kirmitzoglou, Stella Tamana, Ioanna Kalvari, Maria Xenophontos, Eleni Gkotsi and Stalo Demetriou.

Last, none of this work would have been completed if I did not have the love, support and patience of my family.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

αHTMP	α-Helical Transmembrane Protein
CAI	Codon Adaptation Index
FN	False Negative
FP	False Positive
MCC	Matthews correlation coefficient
MSS	Maximal Scoring Subsequences
NPV	Negative Predictive Value
PPV	Positive Predictive Value
RCC	Rare Codon Cluster
RSCU	Relative Synonymous Codon Usage
SOV	Segment Overlap Measure
SP	Signal Peptide
tAI	tRNA adaptation index
TM	Transmembrane
TMβb	Transmembrane β-barrel
TN	True Negative
TP	True Positive

1 Introduction

One of the fundamental and most important discoveries in biological sciences, is the known central dogma of molecular biology (DNA↔RNA→protein), which describes the flow of information in a biological system (Crick, 1958; Crick, 1970). The ingredients for 'baking' DNA or RNA are the four nucleotides, which are combined in triplets to form codons, and then translated into the amino acids, the chemical building blocks of proteins. Does this simple letter code hide more information than we already know?

With an excitement for the potential and power of this 'simple' three letter code, we started this project seeking to explore the impact and specific role that this code has on the biology of genes and proteins, following a bioinformatics approach. In this introductory section, state of art knowledge will be explored, followed by our motivation and hypothesis, in an effort to include all the fundamental information regarding the impact of 'rare' codons in coding sequences.

1.1 Existing knowledge

1.1.1 Is the genetic code "degenerate"?

The beauty and at the same time mystery in all living organisms, prokaryotes and eukaryotes, is the fact that they share the same genetic code, despite some limited variations that exist. The 'standard' (or 'universal') genetic code (Figure 1) consists of 64 triplets of nucleotides referred to as codons and is known to be degenerate, meaning that each amino acid is represented by more than one synonymous codon. Even single silent substitutions that lead to synonymous codons can have a significant impact in the proteins' activity, expression, folding and function indicating that they are essential for the organism (reviewed in Chamary and Hurst, (2009)).

Each codon, with the exception of the three stop codons, encodes one of the 20 standard

amino acids that are used to synthesize proteins. However, there are two rare amino acids, the so called 21st and 22nd, selenocysteine (Sel, U) (Zinoni et al., 1986) and pyrrolysine (Pyl, O) (Srinivasan et al., 2002) respectively. These two amino acids are encoded by the triplets UGA and UAG which are stop codons in the standard genetic code.

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figure 1: The standard genetic code. Taken from (Alberts et al., 2007).

A number of alterations to the standard genetic code have been discovered for various prokaryotic and eukaryotic genomes, as well as for mitochondrial genomes. In mitochondrial systems, both 'sense' and 'non-sense' codons can change identity; on the other hand in prokaryotic and eukaryotic systems, only stop codons change identity with the exception of decoding of the leucine–CUG codon as serine in *Candida* species (Silva et al., 2004). How these codons are used (also referred as codon usage) in an organism has been an open issue for decades and is discussed further on.

1.1.2 Codon usage (Codon bias)

Different organisms show specific preferences for codons that encode the same amino acid. The choice among synonymous codons in both prokaryotic and eukaryotic genes is distinctly non-random (reviewed in (Ikemura, 1981a)). This is known as "codon bias", referring to the differences in frequency of occurrence of synonymous codons in genomic DNA. Patterns and degrees of codon usage bias differ between organisms (Grantham et al., 1980) but also between genes from the same genome (Sharp and Li, 1986). It is generally accepted that the preference of codons reflects a balance between mutational biases and natural selection for translational optimization (Sharp and Li, 1986). Some codons from each set of synonymous codons are more frequently used and tend to occur in highly expressed genes (Ikemura, 1981a; Bennetzen and Hall, 1982; Gouy and Gautier,

1982; Ikemura, 1985; Sharp and Li, 1986; Bulmer, 1991; Kanaya et al., 1999). It has been shown that codon choice influences heterologous expression in a host and Itakura et al., (1977) managed to express for the first time a human protein in a bacterium (somatostatin in *E. coli*). A strategy that is commonly used to increase heterologous expression is to alter the rare codons in the target gene, in order to reflect closely the respective usage in the respective host. Techniques to achieve this, range from site-directed mutagenesis steps to re-synthesizing of the entire gene (reviewed in Kink et al., (1991) and Gustafsson et al., (2004)). However, this can also lead to abnormal protein folding and decrease protein solubility (Cortazzo et al., 2002) and activity (Crombie et al., 1992; Komar et al., 1999). Assuming that rare codons correspond to under-expressed tRNA species, a strategy that has been adapted by biotechnological companies is the improvement of expression by expanding the tRNA pool of the host. This can be achieved by overexpressing rare tRNAs. There are commercially available *E. coli* strains overexpressing these tRNAs, from companies such as Stratagene (www.stratagene.com) and Novagen (www.emdbiosciences.com) (reviewed in (Gustafsson et al., 2004). Nevertheless, recent studies suggest that rare codons are essential for the proper folding of the protein (Makhoul and Trifonov, 2002; Zhang et al., 2009) and irrationally altering the translational kinetics may cause mis-folding.

1.1.3 Factors that contribute to synonymous preferences

Extensive studies have shown that there are many biological factors that shape codon bias including translational selection, tRNA abundance, gene expression level, gene length, GC composition, strand specific mutational bias, amino acid conservation, protein hydropathy, transcriptional selection and RNA stability (reviewed in Ermolaeva (2001)).

As indicated before, in unicellular organisms like *Saccharomyces cerevisiae* and *E. coli*, the synonymous codon preference is related to the relative tRNA abundance (the proportion of tRNA isoacceptors) and this correlation is stronger in highly expressed genes (Ikemura, 1981a; Bennetzen and Hall, 1982; Gouy and Gautier, 1982; Ikemura, 1985; Sharp and Li, 1986; Bulmer, 1991; Kanaya et al., 1999).

Multicellular organisms like *Drosophila melanogaster* and *Caenorhabditis elegans* demonstrate variation in codon usage and the balance between mutational biases and

translational selection has been described as the main driving factor (Akashi, 1994; Carulli et al., 1993; Moriyama and Powell, 1997; Sharp and Li, 1989; Shields et al., 1988). Codon usage in *Xenopus*, although determined mainly by compositional constraints, was also shown to be influenced by translational selection (Musto et al., 2001).

It has been generally accepted that GC composition is related to codon usage (Ermolaeva, 2001) and affects expression efficiency (Ikemura and Wada, 1991; Kanaya et al., 2001). Very low or very high GC content is associated with large codon bias. Indeed in mammals, the codon usage bias is found to be influenced by the variation in isochores (GC content) (Bernardi and Bernardi, 1985) whereas Knight et al., (2001) showed that it is the GC content that drives codon amino acid usage within and across genomes. Moreover, Lynn et al., (2002) have shown this correlation while studying 32 bacterial and 8 archaeal genomes. Nevertheless, more recently Dittmar et al., (2006) suggested that codon usage in mammals is influenced by the tissue specific tRNA pool of the cell. Therefore, what is ultimately shaping codon usage may be a combination of different factors each one contributing to the best fitness of an organism. Concluding, any information encoded in the sequences of information-carrying biological macromolecules (DNA, RNA, mRNA) could pose a constraint on the codon choice (Trifonov, 2011) and should be taken into consideration.

1.1.4 Methods for quantifying codon bias

Several statistical methods have been proposed to analyze the codon usage bias (reviewed in (Cannarozzi and Schneider, 2012)). The first measure proposed was F_{op} (Ikemura, 1981a, 1985), which made use of tRNA abundance data of *E. coli* and *S. cerevisiae*. Since then, several other measures have been proposed and revised (summarized in Table 1).

Table 1: Codon usage indices and methods for codon usage analysis.

Codon usage indices	**Reference**
Frequency of Optimal codons (Fop)	(Ikemura, 1981a, 1985)
Codon Bias Index (CBI)	(Bennetzen and Hall, 1982)
Codon usage preference bias measure (CPS), x^2 and scaled x^2	(McLachlan et al., 1984; Shields and Sharp, 1987)
Codon Adaptation indices (CAI)	(Sharp and Li, 1987)
The effective number of codons (N_C)	(Wright, 1990))
CODONs	(Lloyd and Sharp, 1992)
Codon bias (CB)	(Karlin et al., 1998)
Application of Shannon theory to compute synonymous coding bias in the Human and mouse genomes	(Zeeberg, 2002)

Nevertheless, one of the most popular measures remains the Codon Adaptation Index (CAI) (Sharp and Li, 1987). The authors of CAI examined codon usage in *E. coli* and proposed this measure of directional synonymous codon usage bias. They suggested that in unicellular organisms, there is a preference for high bias in highly expressed genes where the selective force is strong, and low bias in lowly expressed genes as was previously proposed elsewhere (Bennetzen and Hall, 1982; Gouy and Gautier, 1982). In order for the data to be comparable with data sets of different sizes, the codon usage numbers were converted into relative synonymous codon usage values (RSCU). The RSCU, as described by the authors, represents the ratio of codon frequency divided by the frequency expected under the assumption of equal usage of the synonymous codons of an amino acid:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

where the X_{ij} is the number of occurrences for the j codon for the i-th amino acid and n_i the number of alternative codons for the specific amino acid. The relative adaptiveness of a codon w_{ij}, is the frequency of use of that codon compared with the frequency of the optimal codon for that amino acid:

$$w_{ij} = \frac{\mathrm{RSCU}_{ij}}{RSCU_{imax}}$$

Therefore, the CAI for a gene is calculated as the geometric mean of the w_{ij} values corresponding to each codon used in that gene:

$$CAI = \left(\prod_{k=1}^{L} w_k\right)^{1/L}$$

In this definition, L is the number of codons and w_k is the relative adaptiveness value for the k-th codon in the gene. In conclusion, the CAI number is a very simple measure for the codon usage bias especially for bias seen in highly expressed genes. CAI was applied as a measure for predicting gene expression levels, since this figure reflects the level of expression and for comparing codon usage bias in different organisms (Sharp and Li, 1987).

1.1.5 Other measures correlated with codon usage

In *E. coli* it has been shown that the non-random choice of codons is mostly attributable to the availability of transfer RNA molecules within a cell (Ikemura, 1981a, 1981b). Synonymous codon choice has been shown in many studies to be affected by and correlated with tRNA abundance. Quantifications of cellular concentrations of tRNA species *in E. coli* were initially made by Ikemura (1981) who quantified 26 tRNA species. The amounts were measured and expressed as ratios to amounts of $tRNA^{leu}$ (CUG). (Emilsson and Kurland, 1990) as well as (Emilsson et al., 1993) calculated the relative abundance of a set of 18 tRNA species in *E. coli*. Moreover, Dong et al., (1996) made systematic measurements of the cellular concentrations of 46 tRNA species in *E. coli* and these were calculated for each individual tRNA isoacceptor at different growth rates. They showed that there is a biased distribution of tRNA abundance at all growth rates.

Positive correlations were demonstrated between codon usage and tRNA levels for *E. coli* (Ikemura, 1981a, 1981b; Dong et al., 1996), *S. cerevisiae* (Ikemura, 1982) and *Bacillus subtilis* (Kanaya et al., 1999). Figure 2 demonstrates these linear correlations between tRNA levels of *E. coli* and codon usage as calculated in (Theodosiou and Promponas, 2012).

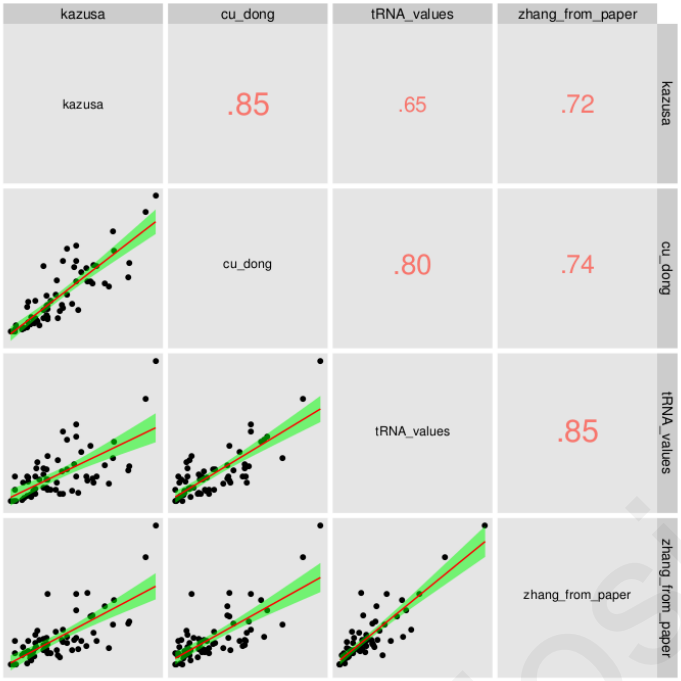

Figure 2: Correlations of RCC detection scales for *E. coli* available in LaTcOm.
Figure taken from Supplementary Data in (Theodosiou and Promponas, 2012). Displayed are pairwise scatterplots of *E. coli* codon usage from CUTG database (`kazusa'), from Table 4 of (Dong et al., 1996) (`cu dong') and tRNA-abundance based scales as calculated in-house (`tRNA values') or from (Zhang et al., 2009; Zhang and Ignatova, 2009) (`zhang from paper'). On the upper right part of the matrix the Pearson product moment correlation coefficients are displayed ($p < 0.05$). Plot generated using the ggcorplot function in the R statistical environment (R Development Core Team, 2008).
(http://groups.google.com/group/ggplot2/attach/6bf632a9718dddd6/ggcorplot.R?part=2)

Furthermore, the individual cellular levels of tRNAs were also shown to be approximately proportional to the copy number of the respective tRNA genes (Dong et al., 1996; Percudani et al., 1997; Kanaya et al., 1999; Tuller et al., 2010a). This strategy was successfully applied to determine optimal codons in species with no tRNA concentrations available (Kanaya et al., 2001). Moreover, the tRNA adaptation index (tAI) (dos Reis et al., 2003) was developed which is a measure that follows the mathematical model of CAI (Sharp and Li, 1987), but estimates the adaptiveness of a gene based in tRNA gene copy numbers.

1.1.6 Codon usage and rare codon clusters

The codon usage statistical measures described above were used, in most cases, to predict

gene expression levels or to estimate codon usage evenness. Nevertheless, none of them was able to detect the position of local prevalence of codons with rare cognate tRNAs. These regions were recently described (Makhoul and Trifonov, 2002; Clarke and Clark, 2008) and we refer to them, from this point on, as rare codon clusters (RCCs). These regions are thought to be important for translational speed control (Makhoul and Trifonov, 2002; Tuller et al., 2010a; Cannarozzi et al., 2010) and can slow down translational elongation (Pedersen, 1984). It has be known for some time that the ribosome slows down in areas that correspond to low abundance tRNAs and recently we have more experimental evidence on this fact (Zhang et al., 2009). Nevertheless, more factors seem to be implicated in the translational regulation.

1.1.7 Translation tuning factors

Protein synthesis by the ribosome is a well described process in prokaryotes and eukaryotes that varies in speed. The translation process is not uniform along an mRNA sequence and this is essential for the proper function of a protein product. The speed of translation can be controlled and several mechanisms have been proposed to cause the so-called translational pausing or ribosome stall. Figure 3 taken from (Gloge et al., 2014) shows some of the up to date proposed mechanisms of translational control.

Recent research efforts have often provided conflicting evidence regarding the factors that determine translation elongation rates. It is widely accepted that rare codons and their clusters are associated with translational pausing (Pedersen, 1984; Guisez et al., 1993; Goldman et al., 1995; Komar and Jaenicke, 1995; Thanaraj and Argos, 1996a; Zhang et al., 2009; Tuller et al., 2011). Sequence-dependent translational rates are reported in previous studies and tRNA abundance is documented as an important determinant of elongation rate, causing variation in the translational rate for each codon (Pedersen, 1984; Varenne et al., 1984; Zhang et al., 2009; Tuller et al., 2011).

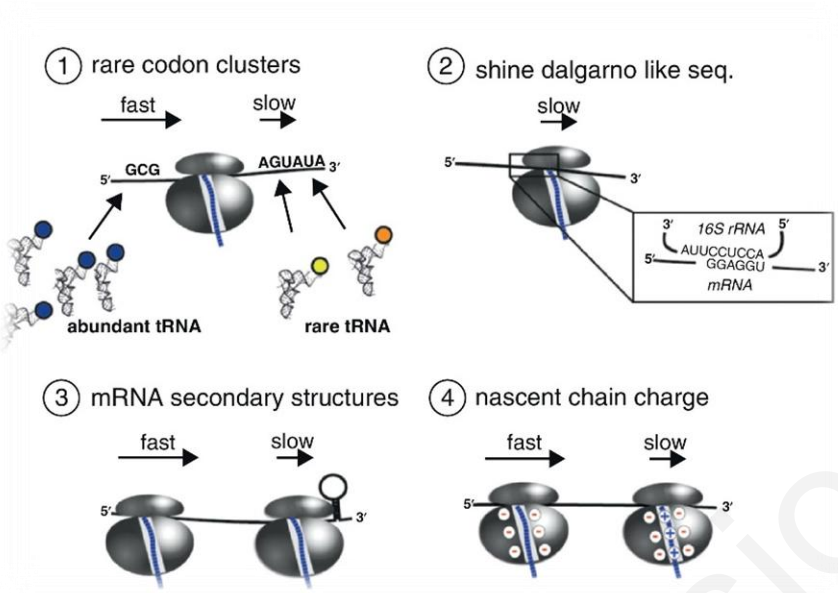

Figure 3: Proposed mechanisms of translational control.
Translational speed can be reduced by the existence of rare codon clusters, secondary structural elements, Shine-Dalgarno–like sequences in the mRNA and interaction of the assembled chain with the ribosomal exit tunnel. Figure from (Gloge et al., 2014).

On the other hand, early studies have indicated that base-pairing of mRNA at the initiation site was a major determinant for translational efficiency in prokaryotes (Schauder and McCarthy, 1989). Similarly, in eukaryotes, secondary structures located at close regions at initiation ATG sites were shown to reduce translation efficiency (Wang and Wessler, 2001). More recently, while investigating the determinants of gene expression of 154 synthetic GFP genes in *E. coli*, (Kudla et al., 2009) showed that mRNA folding stability could explain 10-fold more variation in expression/levels than codon bias or any other determinant. The authors of this work concluded that codon synonymous substitutions at the beginning of the sequence that reduced mRNA stability were correlated with GFP protein abundance and they suggested that this effect is caused by local nucleotide composition and not by codon usage (reviewed in (Angov, 2011)). Nevertheless, (Supek and Smuc, 2010) argued that Kudla et al., 2009 wrongly used a nonlinear regression analysis. Moreover, they argued that the effect of codon usage was masked by the inherent strong mRNA structure that exists in GFP. Additionally, (Tuller et al., 2010b) while investigating the determinants of translational efficiency for *E. coli* and *S. cerevisiae*, they identified a correlation between codon bias and protein abundance, showing that codon bias is an important determinant of translational efficiency.

Recent work, based on RNA-seq for revealing ribosome occupancy on complete transcriptomes, provided genome wide ribosomal profiling data both for eukaryotic (Ingolia, 2010; Ingolia et al., 2009) and bacterial (Oh et al., 2011) species. Ribosomal profiling is the sequencing of short ribosome protected fragments of mRNAs unravelling the positions where ribosomes stall more often, thus offering an (indirect) quantitative measure of translational speed. In a follow up work, the same group revealed that Shine-Dalgarno-like features within coding sequences are a primary drive for translational pausing possibly due to their hybridization to 3' region of the 16S ribosomal RNA in *E. coli* (Li et al., 2012). The initially described Shine-Dalgarno sequence is a purine rich region upstream the initiation site in prokaryotes (Shine and Dalgarno, 1974) that binds to the ribosome at the 3' end of 16S rRNA (Steitz and Jakes, 1975; Jacob et al., 1987).

Another recent analysis using ribosomal profiling data from (Ingolia et al., 2009), proposed that the mechanism of translational tuning is mainly determined by the positively charged residues on the nascent polypeptide (Charneski and Hurst, 2013). These residues have been proposed to stall ribosomes by interacting with the negatively charged ribosomal exit tunnel. However, the same authors on a follow up work (Charneski and Hurst, 2014) argue that positively charged amino acids are known to orient membrane proteins into the membrane with the positive-inside rule. Therefore, they suggested that the increased frequency of positive amino acids at the N-terminal is mostly due to membrane topology.

Nevertheless, a very recent study has challenged previous discussed determinants (Artieri and Fraser, 2014). This work proposed that although ribosomal profiling data present a great opportunity to study the determinants of elongation rate, the results of such efforts gave controversial results. Moreover, they suggest that ribosomal profiling data reported in (Ingolia et al., 2009) suffered from technical sequence bias not taken into account in that study. After incorporating these biases in their analysis Artieri and Fraser suggested that none of the aforementioned factors are implicated in ribosomal pausing but it is the proline amino acid, which has a unique side chain that stalls the ribosome (Artieri and Fraser, 2014). The addition of proline residues in the nascent polypeptide is a slow process due to the unusual cyclic nature of proline as previously shown (Pavlov et al., 2009). Screening for codon sequences that stall the ribosome Tanner et al., (2009) identified amino acid sequences with di-proline codons that cause ribosomal stalling in vivo and this effect has been also confirmed by (Chevance et al., 2014).

All the aforementioned orthogonal approaches try to shed light on the effect of different parameters on translation elongation rates. It is important to remember that each step of translation, initiation-elongation-termination, may be regulated differently. All these factors described above may contribute to ribosomal pausing, showing that the translational mechanism is more complex than previously believed. Nevertheless, the exact level of contribution of all these factors on how the ribosomal stalling is performed still remains unclear.

Having all this in mind, our study is only focused on the effect of rare codons clusters (RCCs) and what follows is evidence showing the correlation of RCCs not only as determinants of translational rate, but consequently as determinants of co-translational folding.

1.1.8 RCCs implicated in co-translational folding

How proteins are finally folded into their 3D structure is a process not yet fully understood. It was previously suggested that the protein sequence was enough to determine the 3D structure of a protein (Anfinsen, 1972). However, the exact code on how a protein finds its final conformation still remains unclear (Komar, 2009). More recent evidence exists showing that additional information lays on the mRNA sequence. For example two identical protein sequences, but with differences in their mRNA synonymous sites, may produce different tertiary structures due to the alterations in translational kinetics at the synonymous sites (Komar, 2009).

It is believed that rare codons within a coding sequence encode the instructions for regulation of protein synthesis and the formation of some secondary and tertiary structures (Purvis et al., 1987; Marin, 2008). In the late 80s, two groups have shown that sequential co-translational folding events can be separated by translational pauses located at domain boundaries (Purvis et al., 1987; Krasheninnikov et al., 1988). Later on, several groups have demonstrated evidence on this issue (Krasheninnikov et al., 1991; Guisez et al., 1993; Komar and Jaenicke, 1995; Thanaraj and Argos, 1996a; Komar et al., 1999; Makhoul and Trifonov, 2002). A computational analysis on the *E. coli* genome, has demonstrated that highly abundant codons are preferentially associated with α-helical

secondary structures, whereas RCCs are more often related with β-strands, random coils and domain boundaries (Thanaraj and Argos, 1996b). Nevertheless, recent computational work on large datasets found no evidence of enrichment in slow codons around domain boundaries in *E. coli*, human and yeast (Saunders and Deane, 2010). However, Saunders and Deane (2010) identified a signal of decreased translation in the transition into helix or strand.

Recently, it has been demonstrated by in vitro and in vivo experiments, that the folding efficiency of the *E. coli* protein SufI is altered, when the translational rate is affected with synonymous codon substitutions or alteration in tRNA concentrations (Zhang et al., 2009). Zhang and co-workers used a bioinformatics approach in which they built a simple algorithm to identify putative regions of slow translation described in (Zhang and Ignatova, 2009). They studied the expression patterns of several proteins, demonstrating that mRNAs with several slow translated regions gave rise to respective translation product intermediates, whereas proteins with no detected slow translated regions gave rise to full length proteins. To study this in more detail, they identified four slow translated patterns in SufI and three of them indeed matched with translation intermediates. They showed experimentally that synonymous changes of rare codons at these sites or increase of tRNAs for these codons lead the protein to degradation and also affected the translocation of the protein, demonstrating that the function of the protein is also affected. Taking these into consideration the authors suggested that slow-translating clusters control the folding process. Several other experimental evidence exist and are nicely reviewed in (Angov, 2011).

1.1.9 Folding of α-helical transmembrane proteins (αHTMPs) - are RCCs implicated in their biogenesis?

Almost all αHTMPs are co-translationally integrated into the membrane lipid bilayer through the translocon channels (SecYEG in prokaryotes and Sec61 in eukaryotes) (Rehling et al., 2003; Osborne et al., 2005). However, how the transmembrane (TM) helices are inserted and finally folded in the membrane still remains unclear. Codon usage in membrane proteins differs from that of soluble proteins (Nørholm et al., 2012) mainly reflecting the hydrophobic nature of many amino acids in membrane sequences (Hessa et

al., 2005). Surprisingly, there is U-bias (Uracil bias) in codons of membrane mRNAs (Prilusky and Bibi, 2009) but their role remains unclear. Interestingly, RCCs have been detected in mRNAs encoding membrane proteins for many species, e.g. in yeast (Képès, 1996), *Emericella nidulans* (Dessen and Képès, 2000), *E. coli* and *B. subtililis* (Zhang et al., 2009). Chartier et al., (2012), in a large scale analysis of conserved rare codons identified in Pfam domains, noted that the longest rare codon clusters are found in membrane sequences. In the work of Képès (1996) the author identified that rare codon clusters often occur at 45 or 70 codons downstream of the TM helix in *S. cerevisiae* and made the hypothesis that a translational pause may occur as the helix is leaving the ribosomal exit tunnel or the translocon. Many hypotheses may rise from this issue and one would be that translational pauses might occur as a part of the insertion mechanisms to facilitate the proper interaction of TM helices within the membrane for the protein to get its final confirmation. This issue is raised in the current work and detailed analysis is discussed further on. While our work was in its final stages before completion, two related works suggested that pause of translation at specific sites in mRNAs can cause local pause of translation elongation thus facilitating the co-translational targeting of membrane proteins to the translocon (Fluman et al., 2014; Pechmann et al., 2014). We discuss these works in more detail in the next section.

1.1.10 Rare codon distribution along mRNA sequences and functional implications

In the absence of any selection, RCCs would appear randomly in the coding genome. However, rare codons have been shown to be enriched at the terminal regions of sequences (Clarke and Clark, 2010). An early study (Ikemura, 1981a) showed that rare codons exist near the start sites of some *E. coli* genes. Clarke and Clark (Clarke and Clark, 2010) studied the distribution of rare codon clusters and showed that rare codons are enriched at both 5' and 3' termini in genes not only in *E. coli* but also in other prokaryotic coding genomes. Nørholm et al., (2012) have also shown the preference or rare codons at the 5' site. A ramp at the first 30-50 codons of 5' termini was also demonstrated by ribosomal profiling (Tuller et al., 2011). Several other studies have also reported the enrichment at 5' termini in different organisms (Allert et al., 2010; Goodman et al., 2013;

Pechmann and Frydman, 2013). A possible explanation of the enrichment of RCCs at the gene start is for keeping the ribosome binding site free from stable mRNA structures (Bentele et al., 2013), whereas another suggestion is a functional role in secretion of secretory sequences (Burns and Beacham, 1985; Power et al., 2004) and/or to allow correct folding of pre-secretory proteins (Zalucki and Jennings, 2007). In eukaryotes, the signal recognition particle (SRP) can pause translation of secreted proteins though its binding to SRP Alu domains to facilitate their translocation to the endoplasmic reticulum (Regalia et al., 2002; Lakkaraju et al., 2008). Nevertheless, an analogous mechanism is unclear to exist in *E. coli* due to the absence of Alu domain (Raine et al., 2003). Clarke and Clark (2010) suggested that rare codons at the 5' terminal in prokaryotes are present to serve the same purpose. Two recently published studies, provided experimental evidence regarding the local slowdown of translation at specific positions of mRNA elements downstream the SRP binding elements in yeast (Pechmann et al., 2014) and immediately before or after the first TM helices in many *E. coli* proteins (Fluman et al., 2014). Pechmann and colleagues analyzed a previously assembled experimental dataset of co-translational interactions of the SRP with nascent polypeptides and identified elements downstream the SRP binding site that might slow translational elongation. Ribosomal profiling data demonstrated increased ribosomal occupancy at these sites and the experimental removal of such a site resulted in inefficient translocation of the proteins through the translocon. Fluman and colleagues analyzed ribosomal profiling data in *E. coli* (Li et al., 2012), identifying Shine-Dalgarno-like elements that slow the elongation before or after the first TM helix. They showed experimentally that insertion of a Shine-Dalgarno element within the segment coding for a TM helix reduces the protein's aggregation. Both studies provide interesting insights for the role of non-optimal codons in co-translational targeting and consequently their involvement in the biogenesis of membrane proteins.

The importance of rare codons at the 5' site was additionally highlighted in a recent study, in which Mahlab and Linial (2014) suggested that only secreted and membrane proteins with a signal peptide (SP) have a rare codon region at the 5' terminal in the human coding genome.

As far as the enrichment at the 3' terminal is concerned, little information exists; Clarke and Clark, (2010) suggested that the mechanism of translation is very different in prokaryotes compared to eukaryotes, therefore the signal in 3' sites that they detected in

their research may be specific to prokaryotes. However, 3' end rare codons were also identified in human (McKown et al., 2013). Moreover, they discussed that ribosome stalling at the 3' termini may allow chaperones or other co-factors to interact with the newly synthesized polypeptide. The computational methods used in all the aforementioned studies to detect RCCs differ in their definition for scales and detection algorithms. These alternative approaches are described further on.

1.1.11 RCC detection methods

In principle, the RCC detection process is the identification of codon clusters corresponding to rare tRNA species along mRNAs, as quantified using scales of experimental tRNA levels. A complete dataset of this type is available for *E. coli* (Dong et al., 1996) and since there is an approximate linear correlation with codon usage (Figure 2) –even though some authors argue against this view (Saunders and Deane, 2010)– approaches based on codon usage scales may and have been alternatively used.

A number of different definitions and algorithms have been proposed for identifying RCCs in coding sequences. These definitions are based on identifying clusters of codons corresponding to rare tRNA species along mRNAs, as quantified using (i) experimental cellular tRNA level data, or (ii) inferred from codon usage based on data from complete genomes or highly expressed genes, and iii) more recently from tRNA gene copy numbers. In the following, we briefly present three recently described methods for RCC identification that make use tRNA concentration or codon usage data.

The %MinMax algorithm (Clarke and Clark, 2008)

The %MinMax algorithm (Clarke and Clark, 2008), a freely available web server at http://www.codons.org, relies on codon usage scales and scans a query sequence with overlapping sliding windows of fixed size of 18 codons. For each such window, the following codon usage frequency-related mean values are computed: (i) actual: the actual codon usage, (ii) Max/Min: the maximum/minimum possible codon usage values for a nucleotide sequence encoding the same peptide, and (iii) average: the average codon usage

value for synonymous codons at each codon in the window. Then, two new quantities (%Max and %Min) are computed, reflecting the deviation of actual from average codon usage, as compared to the maximum (or minimum respectively) possible deviation.

Depending on whether the difference between actual and average is positive or negative, the algorithm reports either the %Max or the %Min value, respectively. An important feature of the %MinMax web server, is the option to compute a measure for validating RCC significance. This is based on an empirical estimation of the expected score, i.e. the average %MinMax score for randomly generated coding sequences based on the relative codon usage at synonymous sites to the sequence analyzed. Even though this approach provides invaluable information for the detected RCCs, it poses significant computational limitations: for validating a single sequence a large number of simulated randomly reverse translated sequences (200 in the current implementation) should be also analyzed. Moreover, neither the text nor the graphical output list the exact locations of RCCs which should be deduced by the user.

The RiboTempo method (Zhang et al., 2009; Zhang and Ignatova, 2009)

In an effort to quantify translational elongation rate, Zhang and colleagues (Zhang et al., 2009; Zhang and Ignatova, 2009) described a tRNA abundance-based scale. The translational rate is calculated as a reciprocal value of the cognate tRNA concentration. For the tRNAs with overlapping codon specificity, the parameters for the tRNA fraction that pairs to each codon were set according to the experimentally determined specificities of the ternary complexes (Bonekamp et al., 1989; Curran and Yarus, 1989; Krüger et al., 1998; Sørensen and Pedersen, 1998) or to the codon-usage index. They went on to develop a window-based approach (RiboTempo), where a moving average for the translational rate is calculated along a query sequence using a window with a fixed size of 19 codons (available at http://hxapp.hexun.com/RiboTempo/Default.aspx). In the current implementation of the RiboTempo web server only graphical output is provided (without explicitly defining RCC locations), and no option is available for statistical RCC validation.

Sliding window issue

Both web servers (%MinMax and RiboTempo) offer different features, e.g. scale options and output formats and they use a simple sliding window approach, with fixed window size w equal to 18 and 19, respectively. These w values correspond to the optimal window size 18 proposed for the problem of identifying translational pause sites by (Makhoul and Trifonov, 2002). An obvious limitation of window-based methods is the detection of RCCs with length at least equal to the applied window size, which probably generates artifacts. In our analysis we demonstrate that with both methods the length of the detected RCCs is dictated by the length of the window applied in the detection process.

Spatial scan statistics approach (Ponnala, 2010)

Recently, another algorithm was proposed (Ponnala, 2010) that applies the spatial statistic to detect rare codons clusters. Ponnala used the tRNA abundance set of *E. coli* as in (Zhang and Ignatova, 2009) to estimate a "waiting time" for each codon. For the corresponding tRNA available for each codon (r_i), the waiting time is the inverse: $t_i = \frac{1}{r_i}$. Moreover, he used the spatial scan statistic approach as proposed in (Huang et al., 2007) for detecting rare codon clusters. The software code is written in MATLAB* (http://sites.google.com/site/jbrpaper/). In this implementation of spatial statistics, the likelihood of zones is estimated on codons that have estimated time more than 0.1. Significant clusters are found by taking the top 100 most-likely highest λ zones and filtering the smallest and most dense clusters of slow translating codons. However, no statistical significance is given for the clusters.

1.2 Motivation, biological hypothesis and specific aims of this work

Taking all this knowledge into consideration, it is here appropriate to introduce the reasons

* Even though the code is freely distributed, the MATLAB platform is available under paid license, so this method was no further validated in this work.

that inspired and finally made me implement this PhD thesis. When I started this research, I was motivated by the work of Zhang and co-workers (Zhang et al., 2009), who as previously discussed, presented experimental evidence regarding the fundamental role of rare codon clusters (RCCs) in regulating the kinetics of proteins synthesis (translation). Previous studies suggested that translational pauses are necessary, to give the appropriate time for the right interactions of the newly synthesized domain to occur, while being correctly folded into the three dimensional conformation (Purvis et al., 1987; Krasheninnikov et al., 1988).

Our working biological hypothesis is that the location of RCCs in coding genes correlates well with other topological and structural characteristics/properties, while their existence may reveal higher level functional features. Proteins with disordered regions, mutlidomain proteins, outer membrane β-barrels, secreted and transmembrane proteins may have a distinct pattern of RCC preference and is very interesting to unravel potential correlations. Moreover, it was intriguing to clarify whether a coupling exists between rare codon-mediated ribosomal attenuation and the biogenesis of αHTMPs, since most of them are integrated into the bilayer co-translationally (Rehling et al., 2003; Osborne et al., 2005). We would expect a different pattern in comparison with transmembrane β-barrel (TMβb) proteins, since folding and insertion mechanism differs from the two-stage insertion model proposed for αHTMPs. TMβb proteins are proposed to fold post-translationally (Kleinschmidt and Tamm, 2002; Tamm et al., 2004).

From what is already known, there is scarce evidence regarding the ideal method or scale to use in order to identify computationally the exact position of translational pauses. There is a need to address shortcomings of the existing RCC detection methods, benchmark existing ones and to also provide new tools and features for RCC detection. Moreover, the evidence regarding any possible roles of RCCs in the function and structure (including the membrane topology of TM proteins) remain unclear. We hypothesize that there are functional and structural implications of the existence of RCCs and in this work we follow a computational approach in order to address these issues. We introduced and explored the concepts and state of the art knowledge regarding codon usage, rare codons, rare codon clusters (RCCs) and what has been demonstrated to date regarding their fundamental role in translational regulation and folding, along with other translation regulation determinants that were described so far. In the next sections of this work, the

implementation of the LaTcOm web server is described in detail (Theodosiou and Promponas, 2012); to the best of our knowledge LaTcOm is the first flexible web application offering alternative methods for detecting RCCs. Furthermore, we present *in silico* experiments on genes encoded in the *E. coli* genome, initially to compare the different methods available in the LaTcOm package and then unravel the general characteristics of RCCs in the coding genome. Moreover, we explore possible correlations of biologically significant positions of RCCs along the sequences. In the last section we present and discuss results related with the co-occurrence of RCCs with several structural and functional characteristics of *E. coli* sequences i.e. potential correlation of RCCs with the distance of genes that belong to operons, relation of RCCs with disordered, transmembrane, secreted, peripheral inner membrane, multidomain, and outer membrane TMβb proteins. Finally, we focused on a correlation analysis regarding the position of RCCs in αHTMP sequences along with topological and structural features. We chose *E. coli* for our study, since for this species there is a wealth of data available: complete genome (for a number of strains), tRNA abundance concentration measurements, highly reliable functional annotation for a significant fraction of genes/proteins, protein localization data, protein 3D structures and topology of transmembrane proteins.

2 Detection of rare codon clusters (RCCs)

2.1 Background

While searching for the methods available and published to detect RCC regions, we came across the limitations of currently existing methodologies. Two of the existing methods, namely RiboTempo (Zhang and Ignatova, 2009) and MinMax (Clarke and Clark, 2010) are sliding window based approaches that are based on an optimal predefined window size (Makhoul and Trifonov, 2002). To overcome this limitation we developed another window-less RCC detection approach, based on the linear time Maximal Scoring Subsequences algorithm (Ruzzo and Tompa, 1999). As described by its authors, this approach could be applied to detect transmembrane regions, DNA binding domains or highly charged residues.

Along with the MSS method, we re-implemented %MinMax and RiboTempo, and provided all these tools freely to the research community under the LaTcOm web server (Theodosiou and Promponas, 2012). In addition, for executing batch analyses we have also developed a standalone version of the suite of tools (unpublished work). Several parameters are supported for tuning the analysis, such as tRNA-abundance and codon usage scales, including the option for users to enter their own scales, and a selection of novel transformations that may prove useful for RCC analyses and research on the field. Moreover, the ability to choose different values of w for window-based RCC detection schemes, the explicit report of RCC coordinates and simulation-based p-values as a measure for statistical RCC validation, enables more sophisticated analyses of RCCs.

2.2 Data and Methods

We developed in Perl a generic sliding window algorithm, for implementing the two recently published RCC identification methods: the %MinMax algorithm (Clarke and Clark, 2008) and the RiboTempo proposed by (Zhang and Ignatova, 2009). These were integrated in a novel modular software package the lowAbundanceMethods.pm. Information regarding the individual algorithms was described in the introduction and

more details can be found in the original publications. Contrary to the initial implementations, the sliding window in LaTcOm is purposely of decreasing size when approaching sequence extremities for enabling computing biologically relevant values near the 5' and 3' termini. In addition to implementations of the aforementioned published algorithms, LaTcOm offers access to a novel RCC identification scheme, based on tailoring the linear time Maximal Scoring Subsequences algorithm (Ruzzo and Tompa, 1999). Briefly, this window-less scheme is introduced to overcome the inherent limitation of window-based algorithms, i.e. detected RCCs are technically restricted to have at least the size of the window. The MSS algorithm was made available by the original authors as the source code of a C++ library and was used in this form with a specialized Perl wrapper module.

2.2.1 The choice of w in sliding window approaches

The window size choices by both the aforementioned algorithms (w: 18, 19 respectively) are based on the optimal window size (18 codons) proposed in (Makhoul and Trifonov, 2002). By varying the window size between 2-32 codons, these authors identified a broad maximum within a range of 16 to 25 codons with 18 being the local maximum (Makhoul and Trifonov, 2002). However, their estimation was based on a rather small collection of sequences (491 mRNAs) from different bacterial species. Additionally, they mention that for a more accurate estimate of optimal w values, larger datasets would be required and more work would be necessary in order to conclude whether there exists "a fine structure in the clusters of rare triplets". Moreover, it is not clear that this optimum is valid in other settings (e.g. eukaryotic coding sequences). Thus, providing in LaTcOm the option for changing w enables addressing those important questions.

2.2.2 Tailoring the Maximal Scoring Subsequences (MSS) algorithm for RCC identification (Ruzzo and Tompa, 1999)

Assuming a sequence $(x_1,x_2,...,x_n)$ of real numbers ("scores"), the score $S_{i,j}$ of subsequence $(x_i, x_{i+1},..., x_j)$ is defined as $S_{i,j} = \sum_{i \le k \le j} x_k$. A candidate maximal scoring subsequence is

simply the subsequence ($x_i, x_{i+1}, ..., x_j$) that locally maximizes $S_{i,j}$, i.e. a subsequence which cannot be extended or shortened from either end without reducing its score. MSS was introduced as a practical algorithm, capable of finding in linear time all mutually disjoint (non-overlapping), contiguous subsequences with greatest total scores given a sequence of numeric scores (Ruzzo and Tompa, 1999). The authors describe some practicalities of implementing the algorithm in software, and also provide source code for a C++ library at http://bio.cs.washington.edu/software. The MSS algorithm is generic, and may be used for solving other interesting biological problems (in addition to RCC identification), however it is not trivial for an average molecular biology researcher to utilize it in the form provided by the authors. We interfaced this library to the LaTcOm web server with modifying the C++ source (with the help of Ioanna Kalvari, University of Cyprus) through a specialized Perl module, which generates numerical arrays based on the scale values (see next section) that are the input to the MSS algorithm. RCCs can then be detected in a sequence as the set of Maximal Scoring Subsequences detected by MSS based on the selected scale and transformation. An inherent feature of MSS is that the exact extents of RCCs are reported. In our implementation, we report those RCCs accompanied with the respective average scale value.

2.2.3 Available scales for RCC detection

Naturally, when interested for detecting slowly translating regions in coding sequences, users are expected to rely on tRNA abundance scales as demonstrated in (Zhang et al., 2009). However, since a complete dataset of this type is currently available only for *E. coli*, codon usage tables may also be used based on the observation that codon usage correlates well with tRNA abundance (see for example Figure 2) Thus, in the current LaTcOm implementation we provide options for using tRNA abundance-based scales, pre-compiled codon usage tables, or user-supplied codon usage tables in GCG format.

Scales currently provided through the LaTcOm web interface are:

i. *E. coli* tRNA abundance values, based on data from (Dong et al., 1996), as calculated in (Zhang et al., 2009), and our own in-house calculated variant (see following section).

ii. (Weighted) codon usage from a subset of highly expressed *E. coli* genes, taken from Table 4 of (Dong et al., 1996) for a growth rate of 0.4 doublings/hour.

iii. Codon usage tables available from the CUTG online database (available at http://www.kazusa.or.jp/codon/; (Nakamura et al., 2000). Currently, tables for *Homo sapiens*, *Mus musculus musculus*, *Saccharomyces cerevisiae*, *Bacillus subtilis*, and *E. coli* are programmatically accessed from our package with a modified version of the Bio::DB::CUTG.pm BioPerl module which we named CUTG2.pm[†].

iv. User-defined tRNA abundance or codon usage scales can be copied and pasted or uploaded on the server. These scales should follow the GCG tabular format, with four columns of data, namely:

– AmAcid: The amino acid code in three letter format.

– Codon: The specific codon for the respective amino acid.

– Number: The actual frequency of the codon in the dataset from which the respective codon usage has been calculated.

– /1000: Frequencies as a fraction of 1000.

– Fraction: The relative synonymous codon usage.

An example of this format may be found at the CUTG web server at http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=37011&aa=1&style=GCG. The multiple scales available to LaTcOm users for *E. coli* are positively correlated; however, subtle differences do exist (see Figure 2 in introduction for details).

2.2.4 A custom *E. coli* tRNA abundance-based scale

In addition to using tRNA concentration measurements for *E. coli* calculated in (Zhang et al., 2009; Zhang and Ignatova, 2009) (kindly provided by Dr. Zhang and Professor Ignatova) we also compiled a similar scale with a simplified procedure. More specifically,

[†] In the initial version of the module CUTG.pm if multiple codon usage table exists in the html search i.e. when searching for 'Mus musculus musculus' the first table is the mitochondrion Mus musculus musculus table and this is selected from CUTG by default. Therefore we modified the subroutine *get _request* in order to get the correct table.

we used as a starting point the experimentally determined intracellular concentrations of tRNA species in slowly growing *E. coli* (see (Dong et al., 1996); Table 5, growth rate of 0.4 doublings/hour). To calculate a unique value for each codon, we take into account the different isoacceptor tRNA species present in *E. coli* along with the (weighted) codon usage from gene subsets, (see (Dong et al., 1996); Table 4, growth rate of 0.4 doublings/hour). In more detail, we calculate the scale value for codon i as:

$$\text{Scale}_i = \sum_{j \in K} T_j$$

where

$$T_j = t_j \frac{c_i}{\sum_{i \in K} c_i}$$

is the fraction of the abundance of tRNA species j that contributes to Scale_i.

- t_j is the tRNA abundance of tRNA species j (taken from Table 5 of (Dong et al., 1996)),
- K represents the set of indices for codons recognized by tRNA species j ,
- c_i represents the weighted codon usage of codon i (taken from Table 4 of (Dong et al., 1996)).

This scale was implemented in a new Perl module named CodonUsageScale.pm.

2.2.5 Available scale transformations

A number of scale transformations are available to LaTcOm users (see below and Figure 4 for details):

- 'Linear shift' ($x \to \hat{x} - x$), where the scale-average $\hat{x}$ is subtracted from each given value, followed by reversing the sign. The motivation for introducing this transformation is for obtaining scales with both positive and negative values in order to identify Maximal Scoring Sub-sequences with the MSS algorithm. Therefore, with the linear transformation applied to scales typically used for quantifying ribosomal attenuation or translational rates, positive numbers in the graphs correspond to "slowly" translating clusters irrespective of the identification method used. With the MSS identification algorithm a linear transformation is

necessary for such scales in order to be able to estimate clusters based on the definition of the algorithm.

- 'Multiplicative inverse' ($x \rightarrow \frac{1}{x}$), where each scale value x is substituted by $\frac{1}{x}$. This transformation may be used when users want to interpret their results as translational rates, starting from a codon/tRNA relative quantity scale. Using this transformation enables the use of the %MinMax algorithm in a translational-rate setting.

- 'Sigmoid' applied to linearly shifted scale ($x \rightarrow \frac{2}{1+e^{-(\hat{x}-x)}} - 1$), which facilitates smoothing the contribution of extremely rare or frequent codons/tRNAs, and

- A combination of the multiplicative inverse and linear transformations ($x \rightarrow \frac{\hat{1}}{x} - \frac{1}{x}$). Using this transformation on a tRNA-abundance based scale with the 'sliding window' algorithm, practically implements the RiboTempo method as presented in (Zhang et al., 2009; Zhang and Ignatova, 2009).

In addition, when a codon usage scale (with no transformation) is combined with the '%MinMax' option of the LaTcOm web server the resulting output is in principle equivalent to the '%MinMax' method[‡] as presented in (Clarke and Clark, 2008). It is worth mentioning that after 'linear' shift the transformed scales are approximately zero-centered, thus we naturally choose zero as the threshold value for RCC detection. By definition, zero is the threshold using the %MinMax algorithm. The transformations are implemented in a new Perl module named TrCodonUsageScales.pm.

[‡] In LaTcOm, sliding windows are implemented with gradually decreasing width when approaching the sequence termini. Thus, at least at the sequence extremities, our implementation may differ to the original RiboTempo and %MinMax algorithms, even when all other parameters have been appropriately set.

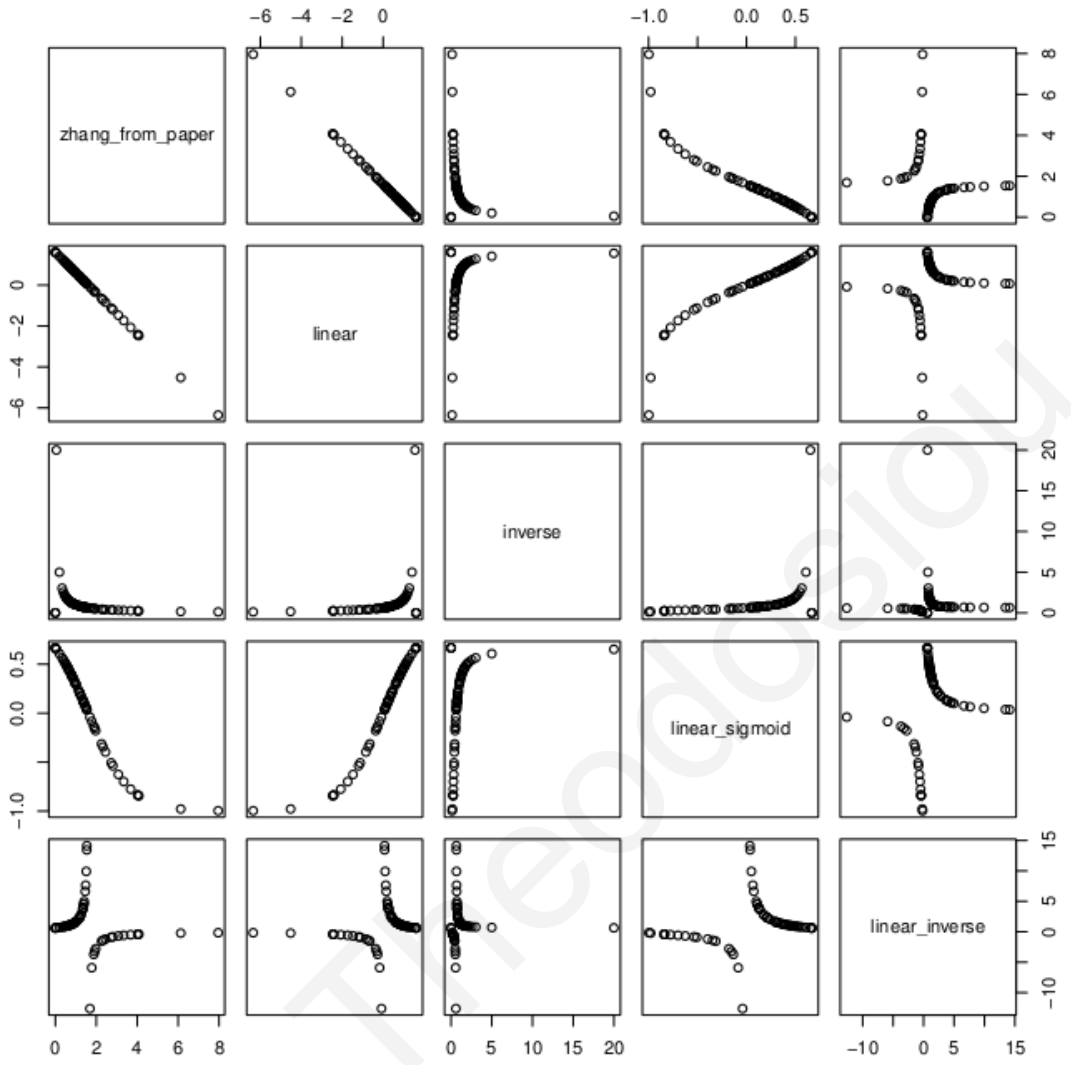

Figure 4: Available scale transformations.
The original values of the *E. coli* tRNA abundance-based scale introduced in (Zhang et al., 2009; Zhang and Ignatova, 2009) (`zhang from paper') are compared pairwise to its all possible transformations from the LaTcOm web server. Plots were generated in the R statistical environment (R Development Core Team, 2008). Figure taken from (Theodosiou and Promponas, 2012).

2.2.6 Reporting RCC ranges

The extents of RCCs are explicitly detected and reported only using the MSS algorithm; the original implementations of %MinMax and RiboTempo do not offer this option. However, it is important for users to know the exact cluster locations. Moreover, for performing RCC validation (see next section) knowing the exact range of a cluster is

necessary. In LaTcOm, we follow a simple procedure to compute the start and end positions of RCCs when using '%MinMax' and 'sliding window', based on the window size and those window centers (i.e. codon positions along the sequence) which correspond to values indicating a RCC (termed `cluster centers'). Starting from each cluster center, we initiate a RCC and extend it up- and down-stream according to the window size. Following this approach, RCCs corresponding to cluster centers located less than $\left[\frac{w}{2}\right]$ [§] codons apart in the sequence are merged to a larger RCC.

2.2.7 Statistical validation

In order to provide a measure for validating RCC significance, (Clarke and Clark, 2008) proposed a simulation based approach. More specifically, an option is provided to %MinMax users to compute as control the %MinMax scores for n = 200 randomly reverse translated sequences, i.e. artificial coding sequences translating to the same amino acid sequence to the complete sequence analyzed. These sequences are generated based on the selected relative codon usage. Then, the average %MinMax score obtained per window on the random sequences is reported along the original sequence for validating cluster scores against the random dataset. Clearly, this approach provides a rough estimate of the expected score of analyzed windows, and is time consuming, since 200 (unnecessary) %MinMax computations are performed for the complete length random sequences.

Calculating the exact expected score value for each RCC

Let assume that for a set of synonymous codons, the scale (score) value is a discrete random variable observed with probability which can be estimated by the frequency of occurrence for this specific codon. This frequency can be obtained by the specific dataset (e.g. genome, gene set) to which the respective scale refers. Then, we can simply compute for each amino acid type aa the exact expected score for the scale of interest, based on the analytical formula:

[§] [x] denotes the floor function, i.e. the largest integer value not exceeding x.

$$E_{\text{aa}} = \sum_{j \in \text{codons(aa)}} \text{freq}_j \times \text{Scale}_j$$

where

- codons(aa) is the set of synonymous codons encoding aa,
- freq_j is the relative frequency of codon j, and
- Scale_j is the (potentially transformed) scale value for codon j.

Apparently, for any set of synonymous codons we can assign the E_{aa}values computed for the respective amino acid. This computation is enabled by using the codon usage scale (pre-defined or provided by the user) for obtaining freq_j values and the selected RCC detection scale for obtaining Scale_jvalues. Then, we populate a hash table keyed by the respective codons and cluster validation can be performed very fast, by simply averaging E_{aa}values for codons within a RCC. Actually, this is equivalent to computing the expected score of the sum of scores, normalized by cluster length. This is a valid estimate for the expected value for the sum even in the case that the variables are dependent. The result we obtain with our approach is theoretically equivalent to the result which would be obtained by the simulation procedure adopted for validating %MinMax results for $n \to \infty$ randomly reverse translated sequences for the cluster in question.

Computing simulation-based p-values

Unfortunately, even though, RCC detection is a problem where we try to locally maximize the segment score, the well-known Karlin-Altschul limit distribution for the maximal segment scores (also known as Karlin-Altschul statistics; equations 1 and 2 in (Karlin and Altschul, 1990) cannot be applied directly to this problem. Therefore, an analytical formula for computing p-values is not available at the moment and further work would be necessary for achieving a solution (even approximate) to analytically tackle this task. It is worth mentioning, that by generating a few sets of simulation data (data not shown; see below) we observe that scores of maximal RCCs may not follow an extreme value distribution (not even approximately); however, more experimentation would be necessary to study this specific empirical distribution in more detail, e.g. to determine the score distributions of RCCs detected using different algorithms or parameter settings.

In order to statistically validate candidate RCCs detected by LaTcOm, we generate 500

artificial sequences by randomly reverse translating the corresponding amino acid sequence based on synonymous codon frequencies. A 'simulation' based p-value (termed: simulated p-value, psim) is reported, based on the fraction of times the observed score was detected to be more extreme compared to the simulation scores. Importantly, constraining the simulation only within the candidate RCCs reduces computational resources, enabling a larger number of simulations for more accurate psim calculations. The analytically computed expected value is still reported in the text output, along with psim. Two levels of significance are included for assisting the users decide of the suggested RCCs: psim < 0.01 (denoted as '**'), and $0.01 \leq$ psim < 0.05 (denoted as '*'); RCCs are also marked on the graphical output accordingly.

2.2.8 Web server architecture

Input to the LaTcOm server is enabled through an HTML page (latcom.html), with enhanced functionalities provided by JavaScript code. More specifically, custom JavaScript code was developed for initially screening and validating input form data. Moreover, the overlibmws library (http://www.macridesweb.com/oltest/) provides pop-up menu display and control for providing useful messages to the users.

Any submitted sequence is validated for

- FASTA format
- upload file size and type
- unknown/ambiguous nucleotides (only standard nucleotides A, C, G, T, U permitted),
- in-frame stop codons, and
- compliance to a coding sequence, i.e. sequence length is a multiple of three, reporting meaningful error messages to users. Nucleotide sequences may be presented to LaTcOm in upper or lower-case, as all sequence characters are converted to upper-case. Additional validation is performed for user defined scales (e.g. valid GCG format) and file uploads (e.g. only ASCII files permitted).

On the LaTcOm back-end, the core CGI program written in Perl (latcom.cgi) utilizes the CPAN CGI.pm module, BioPerl modules for handling sequence data, and the Chart::Graph::Gnuplot Perl package for interfacing the Gnuplot[**] plotting utility for generating graphical output. A number of custom-made Perl subroutines and modules have also been developed for handling codon usage and tRNA abundance-based scales (uploading, parsing, transforming) and for implementing the different RCC detection algorithms. The MSS algorithm was made available by the original authors as the source code of a C++ library and was used with modifications for acquiring input in form compatible to the problem of RCC detection, interfaced with the main LaTcOm application with a specialized Perl wrapper module. The binary file implementing the MSS algorithm was compiled using the GNU g++ compiler, version 4.3.3. LaTcOm has been developed on a Linux workstation, running Ubuntu server 9.0, with Perl version 5.10.0 and the Apache 2.0 HTTP server and was tested on most of the common web browsers (Firefox, Chrome, Safari and Opera). Web server architecture is illustrated in Figure 5.

[**] Available at http://www.gnuplot.info/

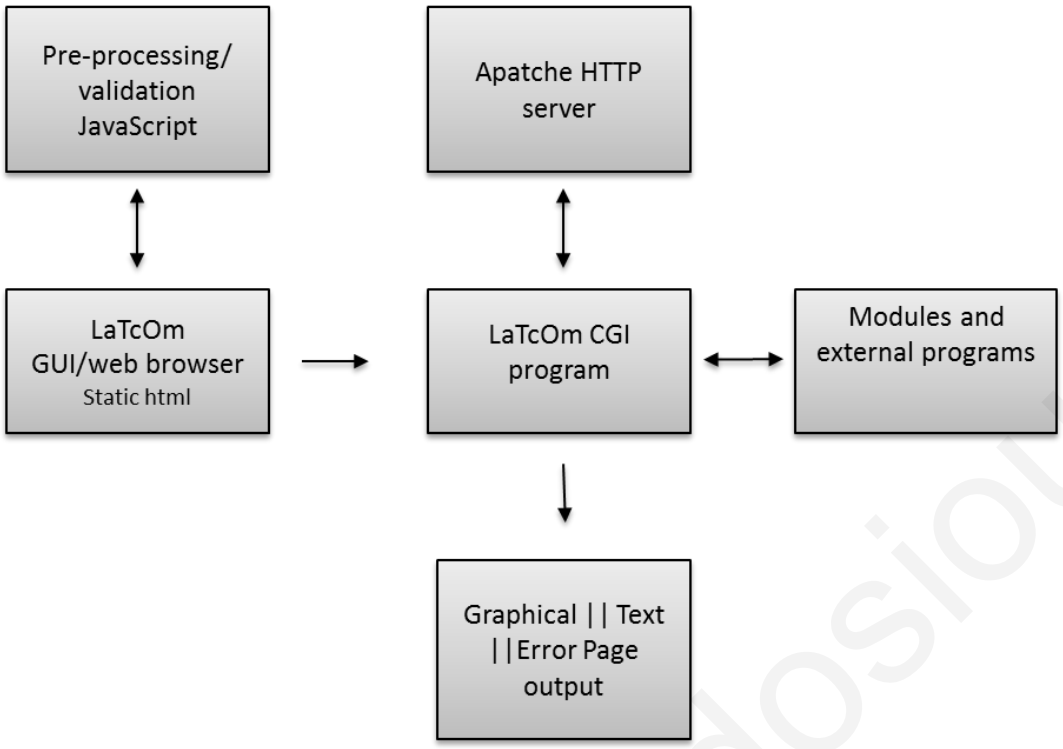

Figure 5: LaTcOm web server architecture as described in the text.

2.3 Results and Discussion

We developed LaTcOm (Theodosiou and Promponas, 2012), which is to the best of our knowledge the first flexible web application offering alternative methods for detecting RCCs, and we introduce a new window-less RCC identification algorithm. RCC detection can be performed from a single and simple graphical user interface. In the current version, three RCC detection schemes are implemented: the recently described %MinMax algorithm (Clarke and Clark, 2008) and a simplified sliding window approach (Zhang and Ignatova, 2009), along with a novel modification of a linear-time algorithm MSS (Ruzzo and Tompa, 1999) for the detection of maximally scoring subsequences tailored to the RCC detection problem. Among a number of user tunable parameters, several codon-based scales relevant for RCC detection are available, including tRNA abundance values from *E. coli* and several codon usage tables from a selection of genomes. Furthermore, useful scale transformations may be performed upon user request (e.g. linear, sigmoid). Users may choose to visualize RCC positions within the submitted sequences either with

graphical representations or in textual form for further processing. Moreover, the ability to choose different values of *w* for window-based RCC detection schemes, the explicit report of RCC coordinates and simulation-based *P*-values as a measure for statistical RCC validation enable more sophisticated analyses of RCCs.

It is worth mentioning that when the LaTcOm web server was being developed, another window-less RCC- detection approach, based on the spatial scan method [introduced by (Huang et al., 2007) was published (Ponnala, 2010) and this it is scheduled to be implemented in a the next version of LaTcOm (LaTcOm++).

A detailed comparison of the features offered by different RCC-detection algorithms is available at Table 2.

Table 2: Comparison of features available by different RCC detection methods.
Taken from (Theodosiou and Promponas 2012).

	Method			
Feature	**%MinMax**	**RiboTempo**	Ponnala 2010	**LaTcOm****
	(Clarke and Clark, 2008)	(Zhang and Ignatova, 2009)	(Ponnala, 2010)*	(Theodosiou and Promponas, 2012)
Availability	Online	Online	Matlab code	Online
Window-less	No	No	Yes	Optional***
User defined scales	Yes	No	N/A	Yes
Scale transformations	No	No	N/A	Yes
Graphical output	Yes	Yes	N/A	Yes
Text output	Yes	No	N/A	Yes
Experimental RCC validation	No	Yes†	No	No
Statistical RCC validation	Yes(optional)	No	Yes	Yes
Explicit RCC location‡	No	No	N/A	Yes
Multi-FASTA input	Yes	No	N/A	Yes
Maximum Input Size	25Kbytes	N/A	N/A	500Kbytes
File upload	No	No	N/A	Yes
Maximum Upload Size	N/A	N/A	N/A	1Mbyte

* Features reported on this table are based on information available in the published work only.

** [This work] With window-based methods, LaTcOm leaves the window size selection to the user, rather than relying on previously optimal values (as for example the 18 window length reported in(Makhoul and Trifonov, 2002). This feature enables users to experiment with different values of this parameter.

*** LaTcOm enables window-less RCC detection by the MSS algorithm.

† RiboTempo was extensively validated by experimental methods for the correlation of RCCs with ribosomal attenuation for several *E. coli* protein coding genes (Zhang et al., 2009) and screened against *E. coli* proteins of known three dimensional structure (Zhang and Ignatova, 2009).

‡ The original implementations of %MinMax and RiboTempo do not explicitly report RCC locations. In our implementation, RCC coordinates are deduced for "RCC centers", as described in the Supplementary Methods section. For MSS, RCC coordinate deduction is inherent in the detection scheme.

2.3.1 LaTcOm Input/Output

Input form

A simple input form enables users to load their sequence data in a text area or upload a file from a local drive. Users may copy/paste one or more cDNA or mRNA sequences in the text area provided in FASTA format (maximum limit 50 sequences with up to 50000 characters). Alternatively, a FASTA formatted file may be uploaded to the server for analysis (maximum file upload size 1Mbytes, unlimited number of sequences). Furthermore, the form provides control elements for selecting the RCC detection method to be applied and all related tunable parameters. The 'Example' button loads the form with the sequence encoding the *E. coli* SufI (FtsP) protein and pre-selects suitable parameters for an example LaTcOm run. Moreover, the 'Pre-run queries' link opens a new page with results computed on the same example sequence using different algorithms and parameters. The LaTcOm input page is displayed in Figure 6.

LaTcOm Output

When LaTcOm is invoked the submitted sequence data are validated and then passed to the core module. Based on the parameters selected by the user the relevant scales and transformations are applied and are passed to the chosen RCC method for identifying putative clusters and their significance. An intermediate results page is displayed linking to the results page and to a compressed archive with all relevant files (graphics, text or both) In those cases where multiple sequences are submitted, this archives contains separate files for each sequence. In any case, results are retained on the LaTcOm server for one week. Examples of the LaTcOm output with the different detection methods are illustrated in Figure 7,8 and 9.

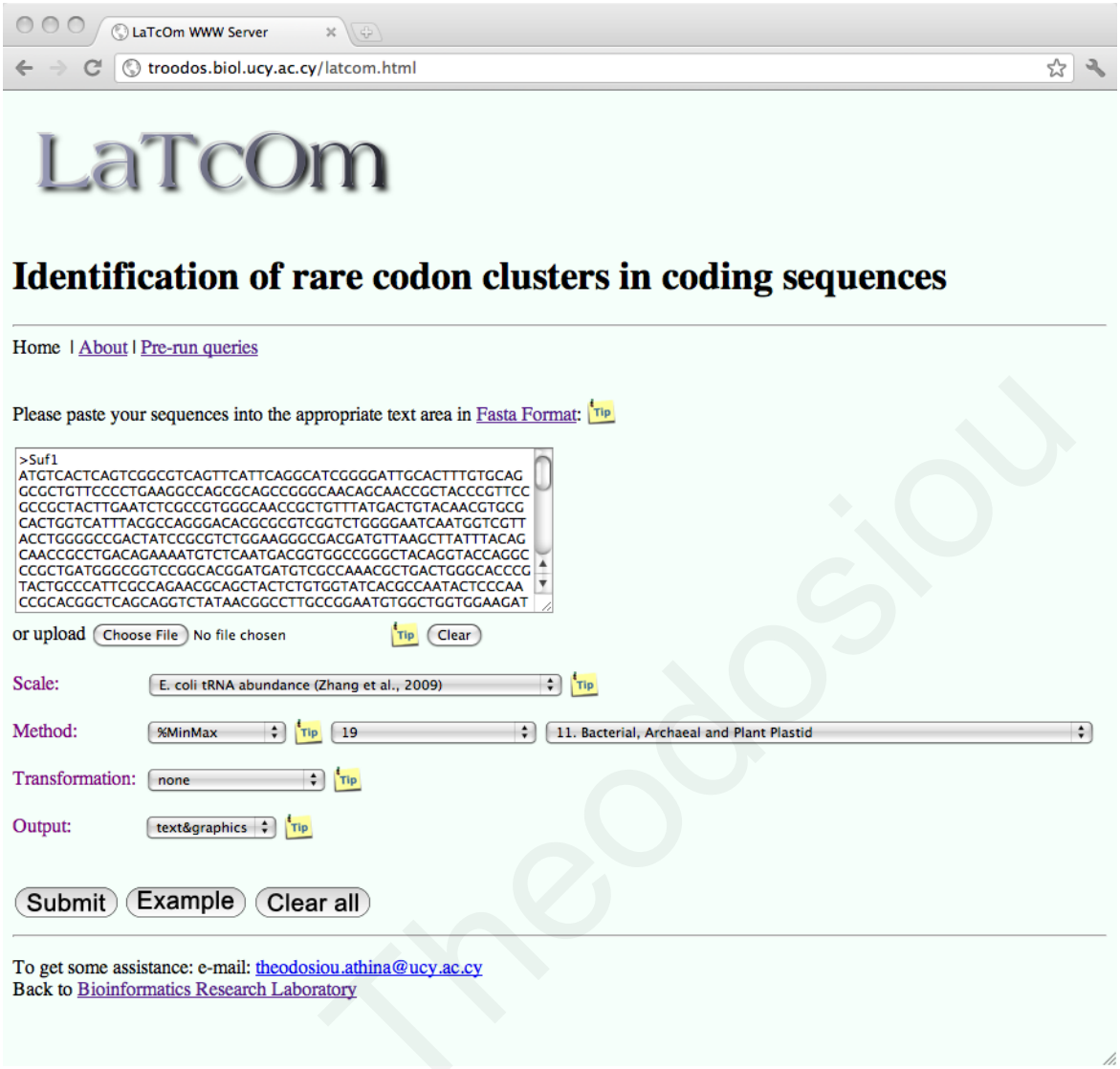

Figure 6: LaTcOm web interface input form.

Graphical output

Graphical output is displayed in png formatted image files, and the data plotted may differ based on the RCC detection algorithm and the selected parameters. For the '%MinMax' (Figure 7) and 'sliding window' (Figure 8) methods, RCCs are plotted using consistently red color in the graphical output and a legend provides a key for users to interpret the different components of the plots, while the parameters used for each run are summarized in a HTML text table below the graphical output. For the MSS algorithm (Figure 9), the raw scale values are plotted for each codon in the background, and alternating clusters and

'non-clusters' are represented on the graph by the average scale values for the respective range. For all methods, putative RCCs are depicted as green horizontal lines in the bottom of the plot (for easily deducing positional coordinates without referring to the detailed text output), and clusters found to be significant are denoted by the '**' and '*' symbols (see Methods on statistical validation).

Text output

Text output in LaTcOm (Figure 10) is designed to be simple and machine readable, and is intended to aid downstream analyses by third party software. The following data/results are included:

- initial values: the scale values per codon position, possibly after transformation.
- processed values: the values computed by the RCC detection algorithm, or (in the case of MSS) the average scale value for a cluster/non-cluster.
- RCC information: a tab-delimited table containing information for all detected RCCs. This table lists:
  - the RCC coordinates (start and end position, designated as "Position of clusters"),
  - the average value for the RCC computed on the initial values normalized according to RCC length (designated as "Score (per position)"),
  - the expected value for the specific RCC, computed as described in the Methods section, and
  - a statistical significance indicator ('**', '*') and the psim value.

Initial and processed values are available as space delimited lists for easy parsing and processing with other third party software.

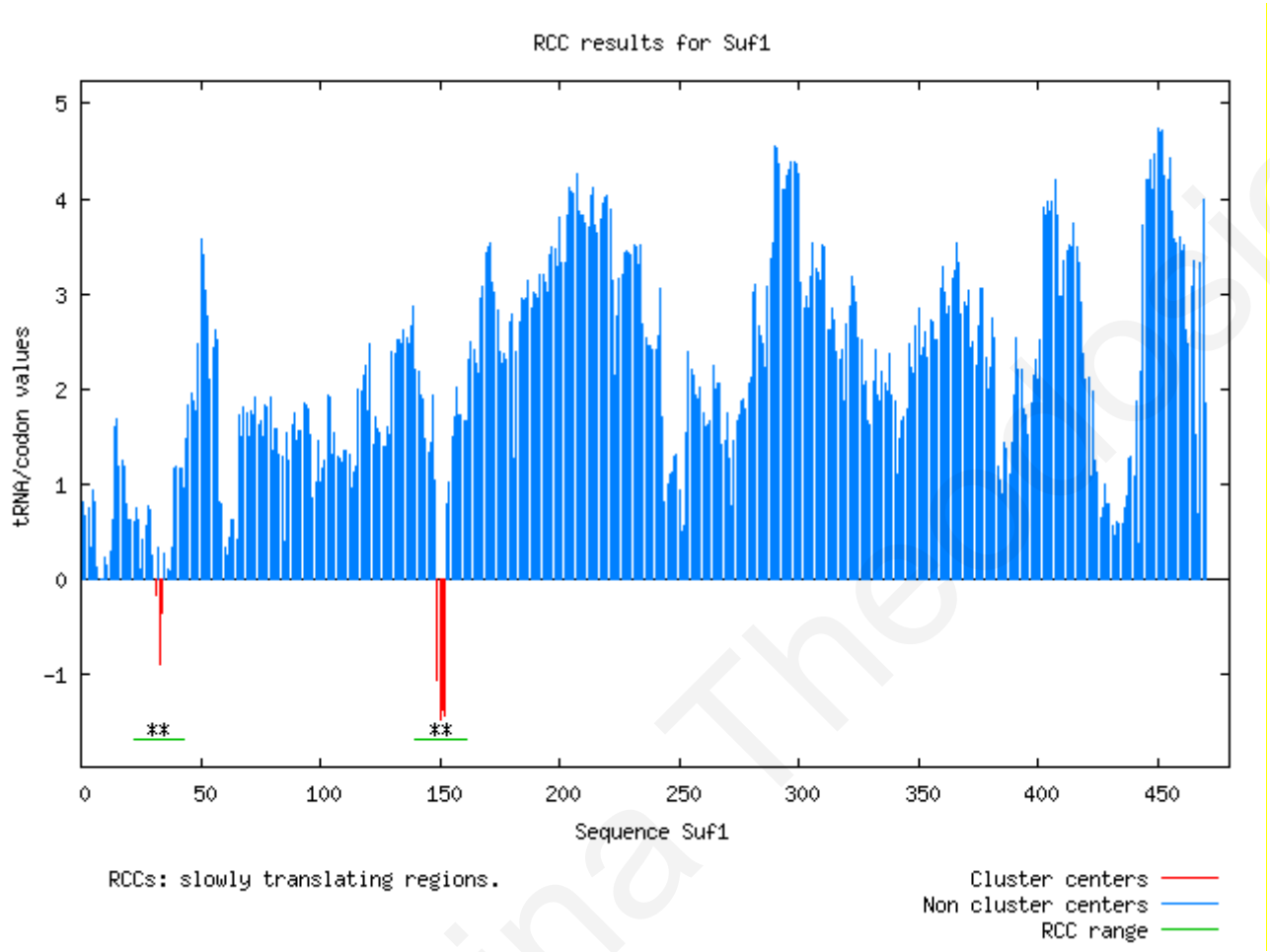

Figure 7: LaTcOm output example of %MinMax.
Graphical output for the coding sequence of *E. coli* SufI by the %MinMax algorithm. *E. coli* tRNA abundance based scale (Zhang et al., 2009) with transformation set to `None' and a window size w = 18.

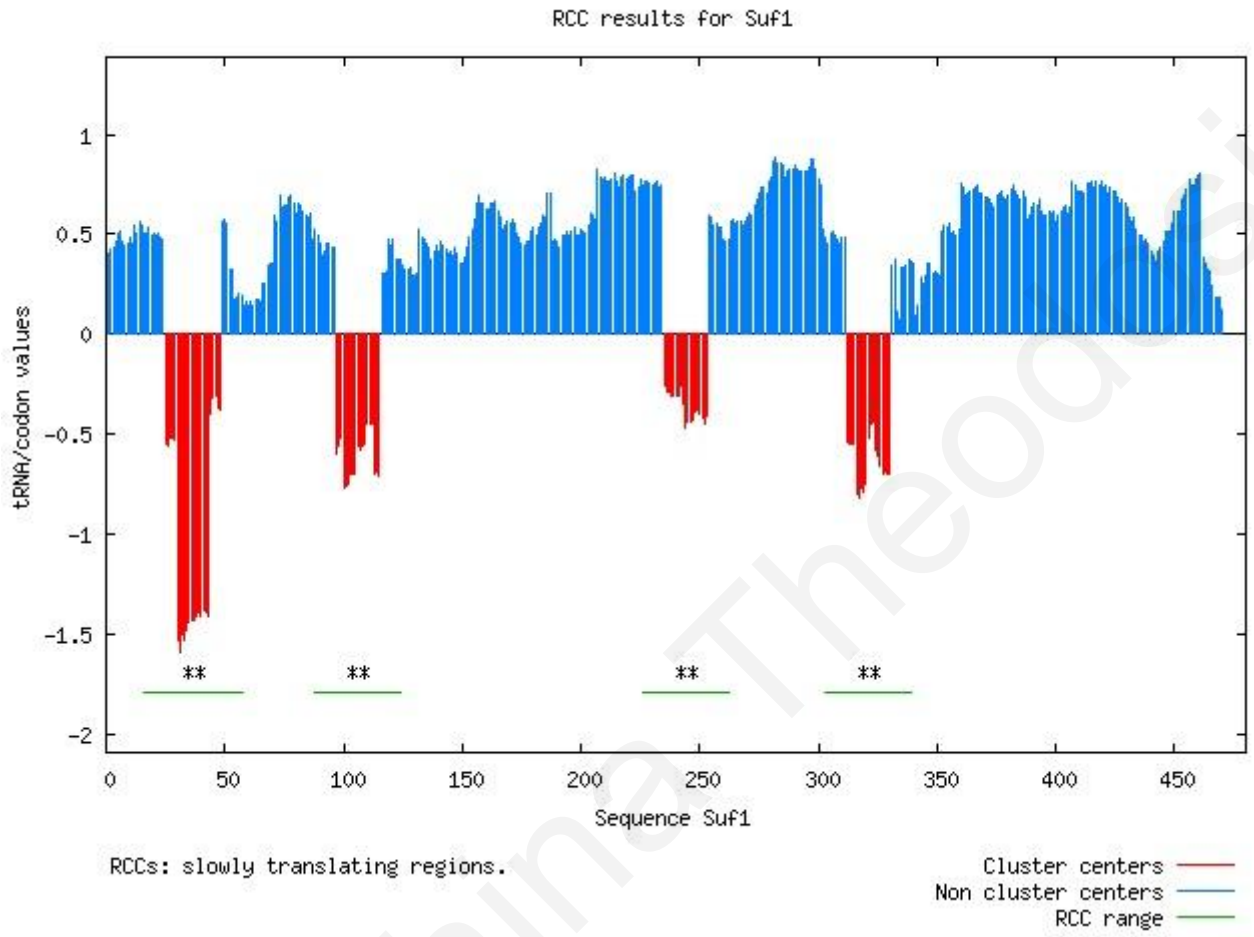

Figure 8: LaTcOm output example of RiboTempo.
Graphical output for the coding sequence of *E. coli* SufI by the RiboTempo algorithm. *E. coli* tRNA abundance based scale (Zhang et al., 2009) with transformation set to `Inverse+Linear and a window size w = 19.

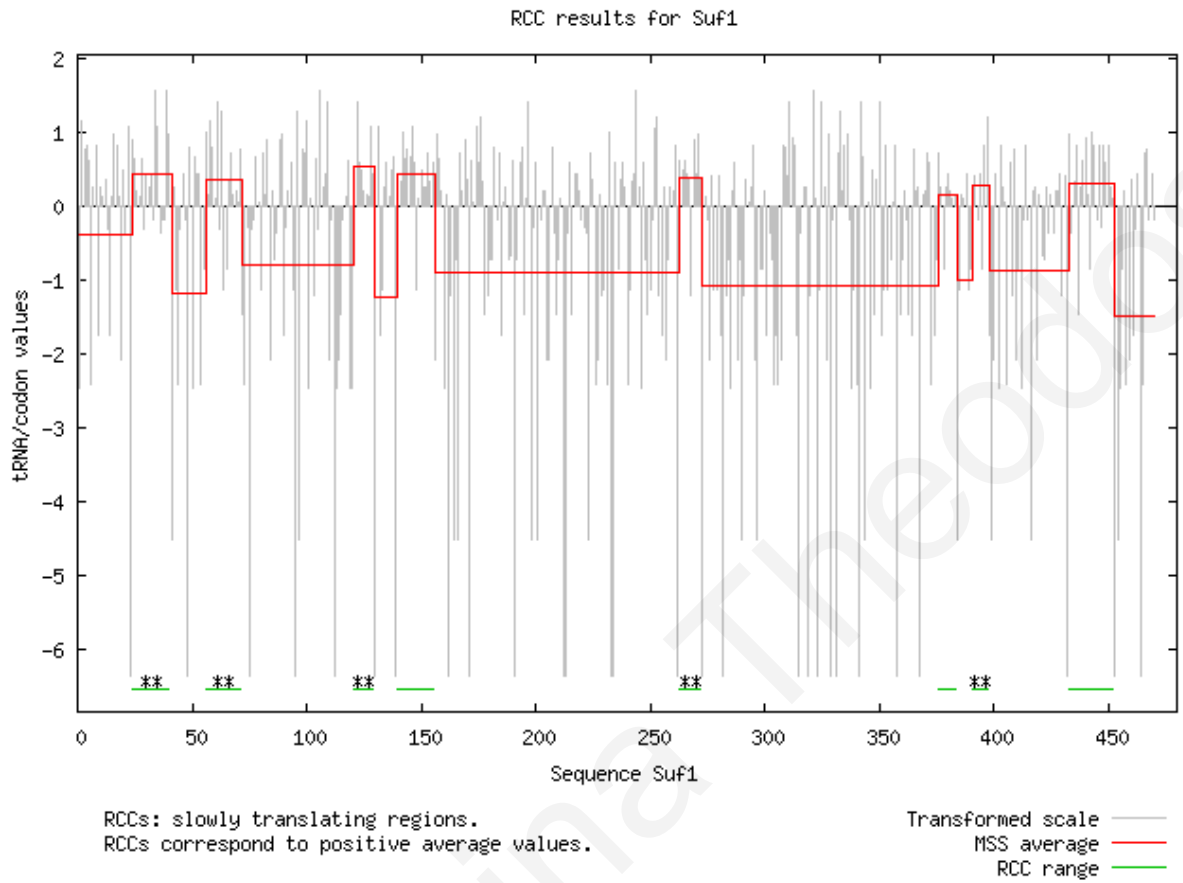

Figure 9: LaTcOm output example of MSS.
Graphical output for the coding sequence of *E. coli* SufI by the MSS algorithm. *E. coli* tRNA abundance based scale (Zhang et al., 2009) with transformation set to `Linear' and a clusterlength = 7.

```
>Sufl |RCC method:mss|Scale:E.coli_tRNA_values_Zhang_et al 2009|Cluster length:11|Transformation:linear+sigmoid
initial_values
-0.844 0.523 0.372 0.398 0.301 -0.836 0.125 0.389 -0.698 0.125 0.066 0.184 -0.153 -0.698 0.066 0.455 0.398 0.06
0.101 0.066 0.319 -0.153 0.213 0.125 0.213 -0.084 0.653 0.494 -0.177 -0.084 -0.084 0.653 0.455 -0.978 0.125 -0.
0.844 0.324 0.213 0.213 -0.836 -0.396 0.467 0.076 0.523 0.381 0.051 0.608 -0.153 0.566 -0.507 -0.396 0.184 0.34
-0.997 -0.153 -0.084 0.324 0.036 -0.507 0.346 0.076 0.423 -0.776 0.111 -0.344 -0.177 0.423 0.455 -0.698 -0.133 
0.346 0.523 -0.844 0.051 -0.536 0.319 -0.153 0.653 0.125 0.213 0.608 -0.776 -0.084 -0.997 -0.844 -0.776 -0.626 -
0.292 0.241 0.111 0.076 0.066 0.494 0.213 -0.997 0.494 -0.698 -0.507 0.125 0.292 0.066 0.241 0.324 -0.997 0.076
0.507 0.051 0.241 0.125 0.125 0.346 0.174 0.292 -0.776 0.455 0.319 0.111 -0.844 0.076 -0.997 -0.536 -0.978 -0.3
0.036 0.494 0.292 0.545 0.174 -0.626 -0.536 -0.344 -0.344 0.381 -0.084 0.346 -0.698 0.036 0.125 -0.344 -0.313 0.
0.608 -0.978 -0.133 0.292 -0.978 -0.084 0.111 0.101 -0.776 -0.776 0.381 -0.177 -0.626 -0.344 0.051 -0.997 -0.99
-0.133 -0.177 -0.978 0.346 0.184 -0.836 -0.776 0.076 -0.536 -0.507 -0.836 0.463 -0.997 -0.997 0.292 -0.396 0.18
0.125 -0.844 0.292 -0.344 -0.626 -0.507 -0.084 0.490 0.545 -0.536 -0.698 0.125 -0.776 -0.344 0.125 0.111 0.389 -
0.423 0.213 0.455 0.184 -0.997 0.066 -0.084 -0.776 0.208 -0.507 -0.507 0.208 -0.698 -0.997 -0.536 -0.344 -0.844
0.036 0.135 0.398 -0.776 -0.978 -0.396 -0.396 0.101 -0.698 -0.177 -0.344 -0.836 -0.698 -0.836 -0.776 0.389 0.38
0.125 0.135 -0.997 -0.536 0.653 0.051 -0.997 -0.507 0.608 0.135 -0.776 0.455 -0.997 -0.084 -0.997 0.346 0.566 0.
0.997 0.608 0.324 -0.978 0.036 -0.844 0.241 -0.626 0.184 0.608 -0.698 -0.507 0.398 -0.507 -0.344 0.036 0.398 -0.
0.368 0.111 0.125 -0.997 0.076 0.111 0.346 0.292 -0.836 -0.698 -0.344 0.346 0.135 -0.396 0.125 0.213 0.111 0.05
-0.396 0.111 0.208 -0.084 0.213 -0.396 0.389 0.545 -0.698 -0.978 -0.776 0.213 -0.844 0.389 0.125 0.036 -0.836 0.
0.389 -0.084 -0.978 0.111 0.135 -0.776 0.076 -0.313 -0.344 0.135 -0.177 0.076 -0.177 0.111 0.111 0.213 -0.536 0.
0.125 0.301 0.439 0.076 -0.396 0.467 0.389 -0.084 0.389 -0.133 0.389 0.292 0.398 0.125 0.051 -0.997 -0.978 -0.8
0.153 0.213 -0.997 -0.836 0.346 0.368 -0.084 0.213 -0.084
processed values
-0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042 -0.042
0.042 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 0.203 -0.
-0.277 -0.277 -0.277 -0.277 -0.277 -0.277 -0.277 -0.277 0.171 0.171 0.171 0.171 0.171 0.171 0.171 0.171 0.171 0.
0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -
0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -0.176 -
0.176 -0.176 -0.176 -0.176 -0.176 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105
0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.1
0.657 0.173 0.173 0.173 0.173 0.173 0.173 0.173 0.173 0.173 0.173 0.173 -0.194 -0.194 -0.194 -0.194 -0.194 -0.1
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -0.194 -
0.194 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.152 0.1
0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347 -0.347
Position of clusters|   Score (per position)|   Expected value| Significance (simulated p-value)
24-40    0.203    -0.035     (1.000)
56-71    0.171    0.024    **(0.006)
121-159 0.105     0.008      (1.000)
167-177 0.173     0.064      (1.000)
433-452 0.152     0.139      (1.000)
```

Figure 10: An example of text output in LaTcOm.
This figure is illustrating the results of MSS on SufI (RCCs longer than 11 reported, transformation: `Linear+Sigmoid'). Notice that the text output has been wrapped and truncated for illustration purposes.

2.3.2 Performance of the LaTcOm web server

The current version of the server provides users with the option to perform RCC analysis by copying-pasting one or more input coding sequences or uploading a local file. The LaTcOm server supports "multi-FASTA" submissions through a text-area for up to 50 sequences, with a total maximum of 50000 nucleotides. For larger submissions, the "upload" feature enables "multi-FASTA" ASCII text files up to 1Mbyte without any limitation on the number of sequences. In all cases, users have the option to view the results online through their web browser, or download them locally for inspection (for further analysis in the case of textual output) in compressed .tar.gz format. For a coding sequence of moderate size (between 1000 and 2000 codons), it takes only a few moments for the computation to complete using any combination of RCC detection algorithm and chosen parameters. This compares well with the time needed to get results from the %MinMax and the RiboTempo web server. When turning on the random reverse translation option on the %MinMax server, LaTcOm execution (with cluster validation executed by default) for the same sequences is clearly faster. In an extreme case, when the coding sequence of the human titin gene was used (>80000bps; EMBL- Bank: X90568.1) computing RCCs took something more than five minutes. The above mentioned figures refer to submitting works to a non-dedicated web server over a network connection and include web server and communication overhead.

As far as the usability of the tool is concerned, LaTcOm has been extensively used from the time of its publication until today (30/03/2015). More specifically, there were 2432 unique submissions (some of them possibly with more than one query sequence) from 78 unique IP addresses in total.

Last, a standalone version of LaTcOm, (batch_latcom.pl) is available and enables batch runs, as a main reasons of implementation this software suite was to facilitate large scale analyses of RCCs in the *E. coli* genome.

2.3.3 Translational profile of SufI

We compared the translational profile of SufI studied in (Zhang et al., 2009) with RiboTempo, with the translational profile of SufI with MSS along with the transient ribosomal arrest at the autoradiograms they provide. By visualizing the results from MSS we demonstrate that even if some of the reporting results are not statistically significant (those without an asterisk symbol) the multiple bands or even single bands can be better described by the MSS method. An arrest at: "full length (FL)>band>46kDa" is shown in the autoradiogram (Figure 11 b); this possibly corresponds to an RCC not making the 77 threshold set by RiboTempo (Figure 11 a) and therefore missed, but identified by MSS (Figure 11 c) at position 433-452, although not statistically significant. The three arrest fragments at 33-40kDa are not that clear with the RiboTempo-derived profile but can be seen in Figure 12 with MSS, when we lowered the threshold, in order to include smaller stretches of RCCs in the analysis. Arrest at 25-28kDa can be seen in both analyses.

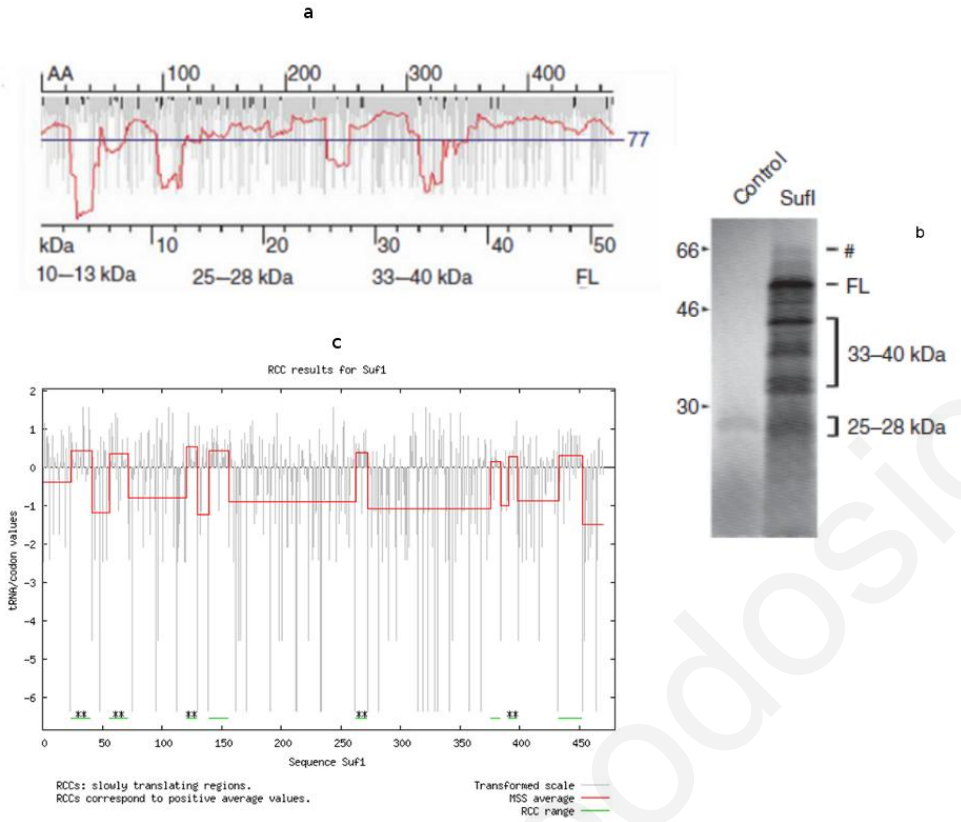

Figure 11: Translational profiles for SufI.
Figure (a) is the translational profile computed with RiboTempo in Zhang et al., 2009 and (b) is an autoradiogram also taken from (Zhang et al., 2009). Figure (c) is the graphical output for the coding sequence by the MSS algorithm in LaTcOm. *E. coli* tRNA abundance based scale (Zhang et al., 2009) was used with transformation set to `Linear' and a clusterlength of c =15.

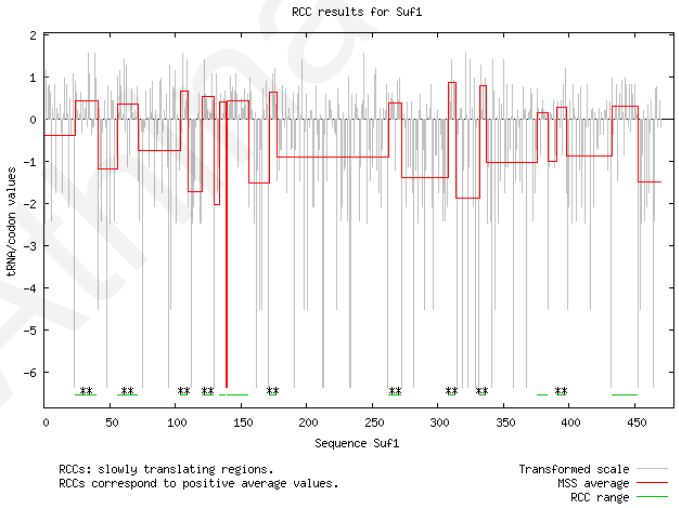

Figure 12: Graphical output for the coding sequence of *E. coli* SufI by the MSS algorithm.
E. coli tRNA abundance based scale (Zhang et al., 2009) with transformation set to `Linear' and a clusterlength c =7.

2.3.4 Use case: RCCs and protein domain structure

One potential use of RCC identification methods is using RCCs as indicators of protein domain organization, due to the proposed impact of slowly translating regions in protein folding. To illustrate such a scenario, we have applied the LaTcOm server on the coding sequences of two *E. coli* proteins with experimentally determined three dimensional structure, previously analyzed with the RiboTempo algorithm in (Zhang and Ignatova, 2009). More specifically, we demonstrate results in Figure 13 on:

• endonucelase III (PDB ID: 2ABK, chain A; GI:16129591), and

• blue copper oxidase CueO (PDB ID: 1KV7, chain A; GI:16128116).

For these proteins, we report the results of three different RCC detection schemes – %MinMax, MSS and the RiboTempo), and cross-reference the detected clusters with structural domain information, as available in the CATH database (Cuff et al., 2011).

It is obvious that, even though the different methods do not agree in the detected RCCs, in several cases the RCCs detected are in proximity to domain boundaries. From these two examples it is evident that MSS detects the RCCs in proximity to domain boundaries. However, more work is necessary, in order to evaluate in detail the performance of different RCC detection schemes with regards to protein folding. This is addressed in the following chapters, in which we benchmark the performance of the different detection methods and we study in more detail the correlation of RCCs and domain boundaries.

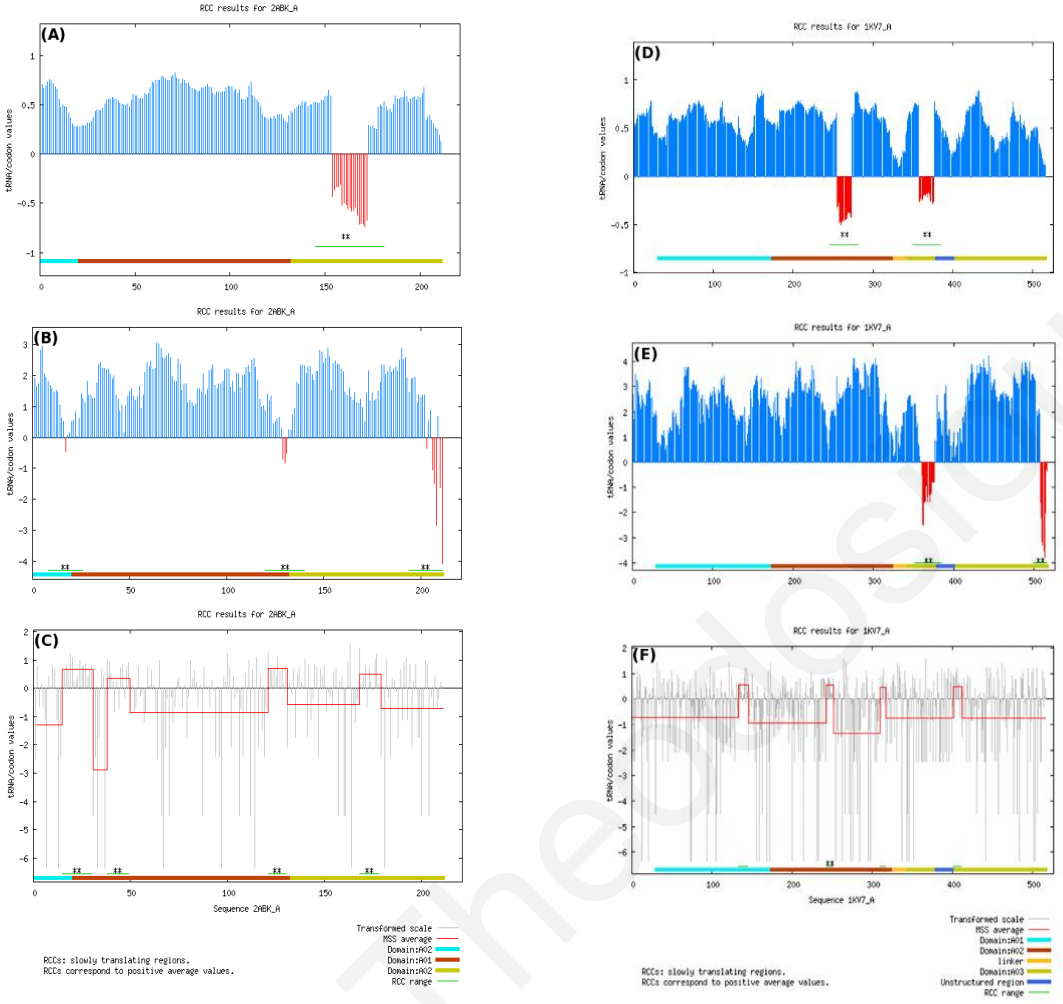

Figure 13: LaTcOm use case: RCCs and protein domain organization.
Results of the (a, d) Sliding window, (b,e) %MinMax, and (c, f) MSS methods for the *E. coli* endonuclease III (left panel; PDB ID:2ABK A) and blue copper oxidase CueO (right panel; PDB ID: 1KV7 A), where we have overlaid structural domain information from the CATH database (Cuff et al., 2011). The different methods were invoked with the following parameters: Sliding window: `Inverse+Linear' transform on the *E. coli* tRNA abundance based scale (Zhang et al., 2009), w = 19, %MinMax*: E. coli* codon usage from the CUTG database (transformation `None'), w = 19, MSS: `Linear' transform on the *E. coli* tRNA abundance based scale (Zhang et al., 2009), with least cluster length selected to 7.

We anticipate that the availability of a versatile online tool for RCC identification will enable a number of analyses to be performed. For example, when optimizing coding sequences for heterologous gene expression experiments, LaTcOm results could be indicative of codon choices that may interfere with proper folding of the polypeptide chain. More specifically, RCCs according to the host organism's tRNA abundance/codon

usage may have to be preserved for expressing functional proteins. In addition, LaTcOm may be used to study patterns of translational rate within diverged protein families, or the correlation of translational rate with protein structural and functional features, such as protein disorder, aggregation and co-translational folding. Such applications may trigger extensions of the current methods, as for example for the analysis of multiple sequence alignments (Widmann et al., 2008) and the study of the mechanics of translation (Tuller et al., 2011).

In this analysis we demonstrated that the flexibility of changing window thresholds or cluster length thresholds may provide more biologically significant results since we don't really know what the optimal window size to use is. The choice of window can be searched further and other thresholds may be applied.

While performing an 'omics' analysis, it is logical and significant to keep as much of the false positives out and one option would be to keep only the long stretches of RCCs. But do we really know if the bigger the stretch the longer the pause or the more significant for proper folding? A single nucleotide change to a synonymous codon can change the translational kinetics and provide a non-functional protein. Motivated from single nucleotide polymorphisms, another potential application of LaTcOm is to be used as a functional annotation tool for next generation sequencing (NGS) analysis. A major challenge in NGS is predicting, among thousands of discovered variants, which are candidates to cause a disease. A typical exome sequencing pipeline analysis, includes as a last step the annotation of possible functional consequences of potential SNPs and indels in most cases for non-synonymous changes. Two widely used functional annotation tools of this type are SIFT (Ng and Henikoff, 2003) and Polyphen (Adzhubei et al., 2010). SIFT is based on sequence homology, whereas Polyphen relies on structural evidence. Nevertheless, there is no method dealing with synonymous changes that may be causing disease. A change in the normal translational profile of a protein would possibly be an indication that the protein is not functional. The demand of computational predictions for the impact of variants is still growing and we believe LaTcOm can serve as a potential tool for that purpose.

A possible extension of LaTcOm in the near future would be the implementation or inclusion of other detection methods such as spatial statistics approach (Ponala 2010), the

inclusion of tAi measure as an alternative scale, the availability of codon usage of more species, the report of domain boundaries along with position of rare codon clusters and many more functional and structural annotation e.g. ribosomal profiles, Shine-Dalgarno putative sites, in order to compare different attributes along with RCCs graphically.

3 Rare codon cluster analysis in the *E. coli* coding genome

3.1 Background

In the following analysis we used the standalone LaTcOm tool in order to benchmark the RCC detection methods and then analyze the RCC positions in the coding genome of *E. coli*. As far as we know this is the first effort to benchmark methods for RCC detection. Analysis on the distribution of RCCs in the *E. coli* genome have already been described elsewhere using different approaches and methodologies (Clarke and Clark, 2010; Zhang et al., 2009). Nevertheless, here we describe the application and results using LaTcOm methods. Additionally, we take this analysis a step further and demonstrate our effort to analyze the correlation of the distance of genes in operons with the existence of RCCs at 3' or 5' terminals.

3.2 Data and Methods

3.2.1 Collection of Data and RCC detection

Taking as dataset the whole coding genome of *E. coli*, RCCs were detected with LaTcOm, using the three RCC identification methods that are available (%MinMax, RiboTempo and MSS). Sequence and annotation data of *E. coli* K12 strain MG1655 were downloaded from NCBI GenBank[†] (4141 protein coding sequences). Relevant file formats are presented in Appendix 1 (Figure 37, Figure 38, Figure 39). The GI numbers (column PID in the U00096.ptt file) may differ based on the date of retrieval of the datasets[*]. All coding genes available in the U00096.ffn file were used for RCC detection. Each gene in this file is discriminated based on the position on the genome. These coordinates are available also in the U00096.ptt file as first column from which we can get the GI number ids. In order

[†](ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779 – 19/11/2013).

[*]Current release is NC_000913.ptt at :
ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779.
We have made a cross matching between the different GIs (data not shown) based on the synonym column.

to map and get the proper GI for each gene in U00096.ffn, the script *get_unique_ids.pl* was developed. Finally, a FASTA formatted file of the coding genes was generated.

RCCs of the genes encoded in the *E. coli* genome were identified with the standalone version of LaTcOm (package *batch_latcom.pl*). RCC detection was performed with the three methods %MinMax, RiboTempo and MSS using selected parameters (Table 3). The number of sequences analysed with each method and duration of each run can be seen in Appendix 1 -Table 35. Some sequences were discarded due to window size parameter limitation on sliding window depended methods (Appendix 1 -Table 36). The LaTcOm tool filters out sequences with length smaller than the window threshold, because technically the sequence has to be at least equal to the window for the sliding window procedure to be estimated properly. Moreover, there are sequences with in-frame stop codons, which are known to encode selenocysteine in vivo as described in the introduction. Such cases are not handled by the current version of LaTcOm, therefore these sequences were discarded (Appendix 1 -Table 37).

Table 3 : Parameters used for LaTcOm cluster analysis of the *E. coli* coding genome

Method	MSS	RiboTempo	%MinMax (z)	%MinMax (cu)
Cluster length/ window size	15	19	19	19
Scale	*E. coli* tRNA (Zhang et al., 2009)	*E. coli* tRNA (Zhang et al., 2009)	*E. coli* tRNA (Zhang et al., 2009)	*E. coli* codon usage Codon Usage database[6]
Transformation	Linear	Linear+inverse	None	None
Output	Text	Text	Text	Text

3.2.2 Module for reading LaTcOm results

LaTcOm results are generated and reading these results was made possible with the development of the *read_LaTcOm.pm* module in Perl. Through this module, the subroutines *get_clusters* and *read_ptt* are used in several scripts developed further for the analysis. After careful consideration, it was decided that RCCs that were found on w/2 extremities (see Methods in Chapter 2 for detailed explanation) may bias the results, therefore these clusters were excluded from the analysis. (This exclusion is optionally made possible in the *get_clusters* function). Initial screening (not shown), revealed that there are artifacts with window based methods found on these sites, therefore this might be

[6] http://www.kazusa.or.jp/codon/

affecting the benchmarking of the results. In order to be consistent we excluded extremities from MSS results as well, even though this is not a window based approach. Another option is the selection of only statistically significant clusters (see Statistical validation in Chapter 2) for further analysis. Having the data set, in-house software tools were developed (unless mentioned otherwise) for analysing statistically the detected RCCs from *E. coli* sequences.

3.2.3 Benchmark approaches for LaTcOm methods

Quantifying correlations with Mathews correlation coefficient (Matthews, 1975)

Initially, we compared the output results of the RCC detection methods using the Matthews correlation coefficient (MCC) (Matthews, 1975). To achieve this, the *transform_files_for_SOV_MCC.pl* Perl script was developed that reads the text output result from LaTcOm (as in Figure 10 of Chapter 2) and transforms the RCC positions to a FASTA format file of sequences with strings of Cs and Hs (Cs are for the clusters, Hs are for the non-cluster positions). An example of the format of this transformed file is shown at Appendix 1 - Figure 40.

Moreover, the *MCC_caclulator.pl* Perl script reads the transformed file from two RCC analyses - a reference file and a compared file. For example, as a reference we used the

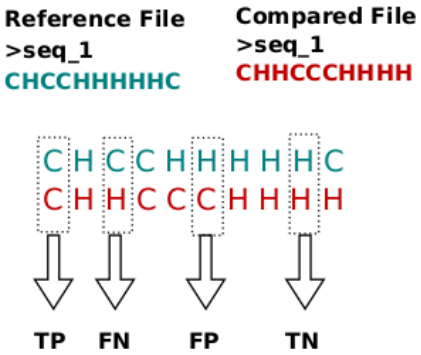

Figure 14: Calculation of MCC.
Graphical representation for true positive (TP), false negative (FN), false positive (FP) and true negative (TN) values calculated in MCC analysis.

RiboTempo results and as a compared the results i) from MSS and ii) from %MinMax. Sequences reported in both files were used further in the analysis. The program calculates five values. The true positive value (TP), the true negative (TN), the false positive (FP), the false negative (FN) and finally the Mathews correlation coefficient (MCC). The criteria for the assignments are described in Figure 14.

These values are used to estimate the MCC directly from the confusion matrix using the formula:

$$MCC=\frac{(TPxTN)-(FPxFN)}{\sqrt{(TP+FP)x(TP+FN)x(TN+FP)x(TN+FN)}}$$

If TP=0, FN=0 and FP=0 (that means that there are no cluster positions “C” in the compared sequences) then we assign MCC=1. This is explained by the fact that both methods did not detect any clusters in the sequences, therefore they correlate perfectly. Moreover, if any of the four sums in the denominator equals to 0, then we assign the denominator value equal to 1 by definition and the MCC is set to zero.

The first measure was the calculation of the MCC per gene and then averaged. Secondly, the overall MCC is calculated taking the sequences as a single string. If the output value of a MCC comparison equals to 1, this is interpreted as perfect (positive) correlation. If the output equals to -1, then this is a perfect (negative) anticorrelation whereas if MCC equals to 0, then this is random (no-correlation).

The MCC analysis was further divided in two sections. The first analysis was done as described above and is referred as MCC_v1. In the second analysis (MCC_v2), with *MCC_calculator_v2.pl*, the following genes were excluded:

a) Genes where none of the compared methods detected any clusters. That is the case described above where TP=0, FP=0 and FN=0 avoiding the over-representation of MCC =1

b) Genes where the compared methods detected clusters only in the reference or only in compared gene. With this we avoid the over-representation of MCC=0 for these cases.

For the analysis where MCC was calculated for each gene, the functions *summary ()* and *sd()* for standard deviation in the R statistical environment (R Development Core Team, 2008) were used in order to obtain descriptive statistics regarding the distributions of MCC values.

In order to evaluate the MCC distributions, random MCC distributions were generated with the script *generate_random_sequences.pl*. The script reads the FASTA files of the reference and compared files (as described above) (Appendix 1 - Figure 40) and creates shuffled FASTA files with the same composition. The random pair sets were then given to *MCC_calculator_v2.pl* in order to create the distributions. The statistics with *summary()* and *sd()* functions in R statistical environment (R Development Core Team, 2008) were also calculated for these random distributions.

Ultimately, we compared the MCC distributions with random MCC distributions using the wilcoxon rank test in the R statistical environment (R Development Core Team, 2008) with the function *wilcox.test()*.

Quantifying correlations with segment overlap measure (SOV) (Zemla et al., 1999)

An alternative benchmarking approach was the comparison with a program designed initially to compare secondary structure elements named segment overlap measure (SOV) (Zemla et al., 1999). The SOV program was downloaded from: (http://proteinmodel.org/AS2TS/download_area/) and a scipt sov.pl was developed in order to create the input for SOV program. For the input creation, the transformed files were used as described in the previous section (Appendix 1 - Figure 40). Moreover, two analyses were performed as described in the previous section: SOV version 1 in which all sequences are included and SOV version 2, were sequences in comparison which are both either all 'Hs' or all 'Cs' are excluded. The shuffled FASTA format files created in the previous section were also used in order to generate random SOV distributions. The statistics with *summary()* and *sd()* functions in the R statistical environment (R Development Core Team, 2008) were generated for random and non-random distributions and the *wilcox.test()* was used to compare distributions.

3.2.4 Methodology to analyze LaTcOm results

In order to estimate the statistical properties of the RCCs detected, the script *statistical_RCC.pl* was developed. The script reads the LaTcOm results and gives as output a table with measurements described below (i-x) as well as data files to be used for downstream Gene Ontology enrichment analysis (see following section):

i) Number of sequences (NS) analysed (those that passed the control thresholds)
ii) NS with no clusters
iii) NS and percentage of sequences from total with at least one cluster
iv) Average coverage of codons in RCCs per gene ("Codon coverage")
v) "Overall codon coverage" which is computed as the fraction of the total codons in the dataset that is detected in RCCs
vi) The average number of RCCs per sequence
vii) The percentage of sequences (with length >=200 codons) that have RCCs at 5' and 3' termini. Regarding 5' sites, if the start of an RCC was within the bin limits (<=20, <=40, <=60 <=100 codons), then the sequence was considered to have a cluster. For

3' termini, if the end of the RCC was within the bin limits, then the sequence was considered to have a cluster at the appropriate bin.

viii) The distribution of RCCs at the 5' and 3' terminals of the sequences.

ix) The percentage of sequences that have none, exactly one, two or three and greater or equal to four RCCs.

x) The length distribution of RCCs keeping in mind the threshold of RCC length and window size that might bias the results.

Next, the script get_*distributions.pl* was developed, in order to estimate the following:

a) Detailed RCC length distribution.

b) Distance distribution of RCCs from the 5' and 3' termini.

c) Distance distribution of first or last RCC from 5' and 3' termini.

d) The distribution of distances between adjacent RCCs for sequences with 2 or more clusters.

The distributions b and c are estimated for sequences with >=200 codons. Wilcoxon rank tests in the R statistical environment (R Development Core Team, 2008) using the *wilcox.test()* function were used to compare statistically the distributions (b-d) with the different methods and between the two termini. A distance threshold for visualization purposes was set for the closest RCC to 5' or 3' terminal to 300 codons for b-c.

3.2.5 RCC analysis in multigene operons of *E. coli*

For this analysis the distance in base pairs (bp) of neighbouring genes that belong to operons was extracted. To achieve this, the following procedure was designed and is described in Figure 15, 16 and 17.

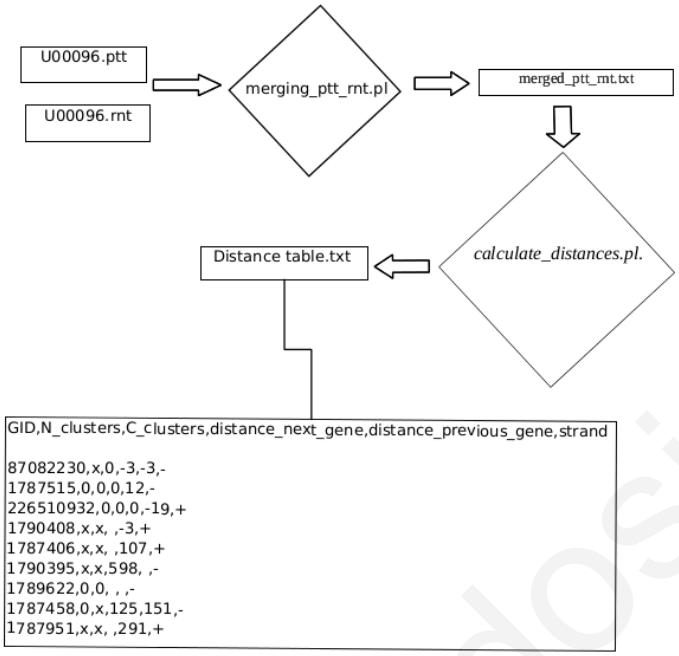

Figure 15: Pipeline for counting distances between of neighboring genes in *E. coli*.
Upstream and downstream distances of neighboring genes were estimated for each gene in *E. coli*. This is the methodological procedure to create an *E. coli* dataset with information regarding existence or not of RCCs at the 5' and 3' terminal sites and distances from neighboring genes. This dataset is created for all *E. coli* genes. Distance_table.txt produced represents a comma delimited file with six columns (1= gid number, 2= x for cluster at 5'; 0 for no clusters at 5', 3= x for cluster at 3'; 0 for no clusters at 3', 4=distance in bp from the next gene, 5=distance in bp from the previous gene, 6=strand).

Figure 15 shows the pipeline and Figure 16 defines graphically the way the software estimates the distance in mutlicystronic genes. Firstly, a merged file is created with sorted coordinate position information for coding genes and RNAs. Then, the distances from the upstream and downstream neighboring genes are calculated taking into account the coding strand information. This information along with the existence of RCCs at the termini is reported in Distance_table.txt.

In order to limit our search for genes in operons, the OperonSet.txt was downloaded for the *E. coli* from RegulonDB (version 8.0 : http://regulondb.ccg.unam.mx/) (Salgado et al., 2013). This file has information for the genes belonging to known operons. The software *read_operon.pl* was developed to extract operons with more than one gene (847 operons). An issue raised from this set, is that it uses the gene name to identify each gene in the operons. However, gene names are not

unique within the U00096.ptt dataset. We identified that the non unique gene names correspond to transposase genes (their gene name starts with "ins"), therefore we excluded all these from further analysis. Moreover, in case of a gene name in Operonset.txt that is not present in the U00096.ptt file, the whole operon was not taken into account. Additionally, computationally predicted operons in the set (that is computationally inferred evidence that support the existence of the operon), are also not taken into account and whole operons are discarded. Only genes that match a gene name in U00096.ptt and belong to operons with 2 or more genes are further analysed (884 genes in 725 operons). The above filtering was achieved with *get_distance_table.pl*.

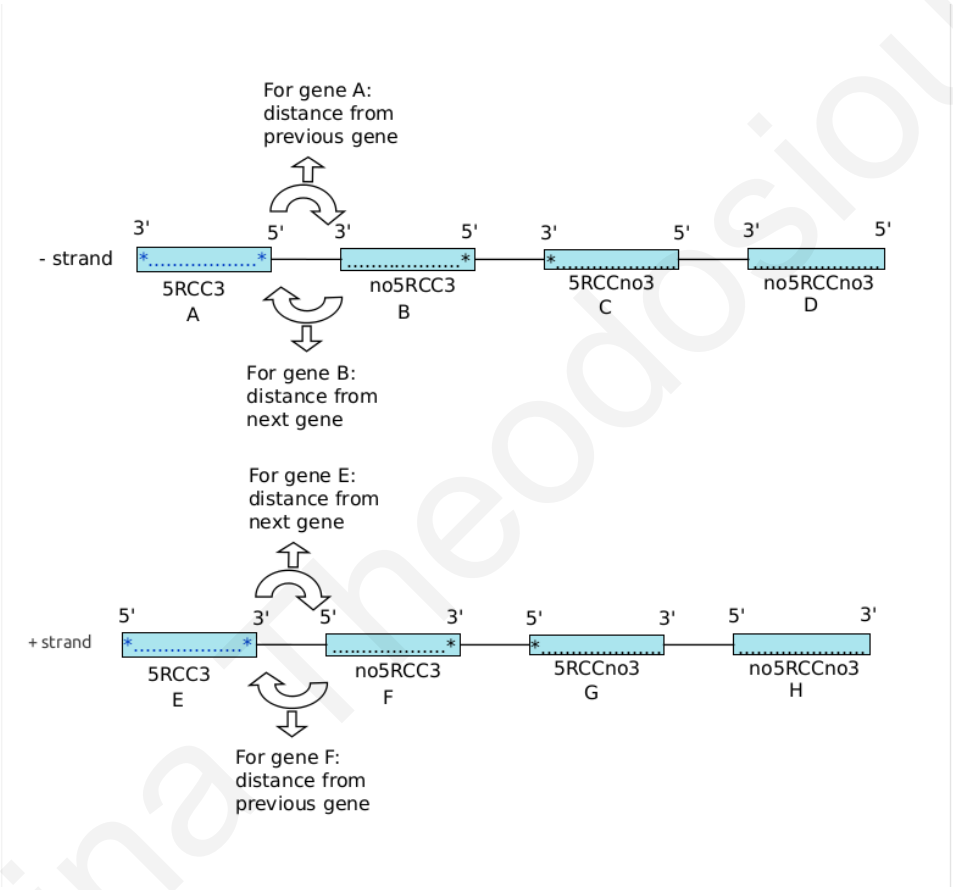

Figure 16: Gene organization in operons and illustration of the distances calculated.
The figure shows a hypothetical scenario of 4 genes on the – strand (A, B, C, D) and 4 genes on the + strand (E, F, G, H). On both strands it is demonstrated how the distances are calculated on genes A and B and on genes E and F. The symbol * is placed on 5' or 3' termini or both when the gene has at least one RCC at these sites. The name "5RCC3" is for genes that have RCC on both sites; "no5RCC3" is for genes that have RCC at the 3' but not at 5' site; "5RCCno3" if for genes that have RCC at the 5' but not at 3' and "no5RCCno3" is for genes that have no RCC on both sites. 5' or 3' termini are defined as the first or last 100 codons respectively.

Furthermore, taking into account the RCCs detected at 5' and 3' termini with LaTcOm (<=100 bp from either end is considered terminal), six gene datasets were generated with *get_distance_table.pl* and *read_distance_table.pl (*pipeline and datasets are shown in

Figure 17):

a) with RCCs at the 5' (dataset 1)

b) with RCCs at the 3' (dataset 2)

c) RCCs on both 5' and 3' (dataset 3)

d) without RCCs at the 5' (dataset 4)

e) without RCCs at the 3' (dataset 5)

f) RCCs in none of 5 'and 3' (dataset 6)

The Wilcoxon Rank Test with function *wilcox.test()* was used in the R statistical environment (R Development Core Team, 2008) for comparing the distribution of distances between neighboring multicystronic genes.

The following strand-specific distance distributions were compared with the Wilcoxon Rank Test (Figure 17 shows the datasets):

a) upstream genes, for genes in operons with dataset 1 versus 4

b) downstream genes, for genes in operons with dataset 1 versus 4

c) upstream genes, for genes in operons with dataset 2 versus 5

d) downstream genes, for genes in operons with dataset 2 versus 5

e) upstream genes, for genes in operons with dataset 3 versus 6

f) downstream genes, for genes in operons with dataset 3 versus 6

An additional analysis was performed to compare the distribution of distances between genes in the following classes:

i) An RCC exist on the 3’ terminal of gene and no RCC exists on the 5’ of the next gene

ii) An RCC exist on the 3’ of a gene and an RCC on the 5’ of the next

iii) No RCC exists on the 3’ but an RCC exists on the 5’ of the next and

iv) No RCCs on the 3’ of the gene and no RCC on the 5’ of the next.

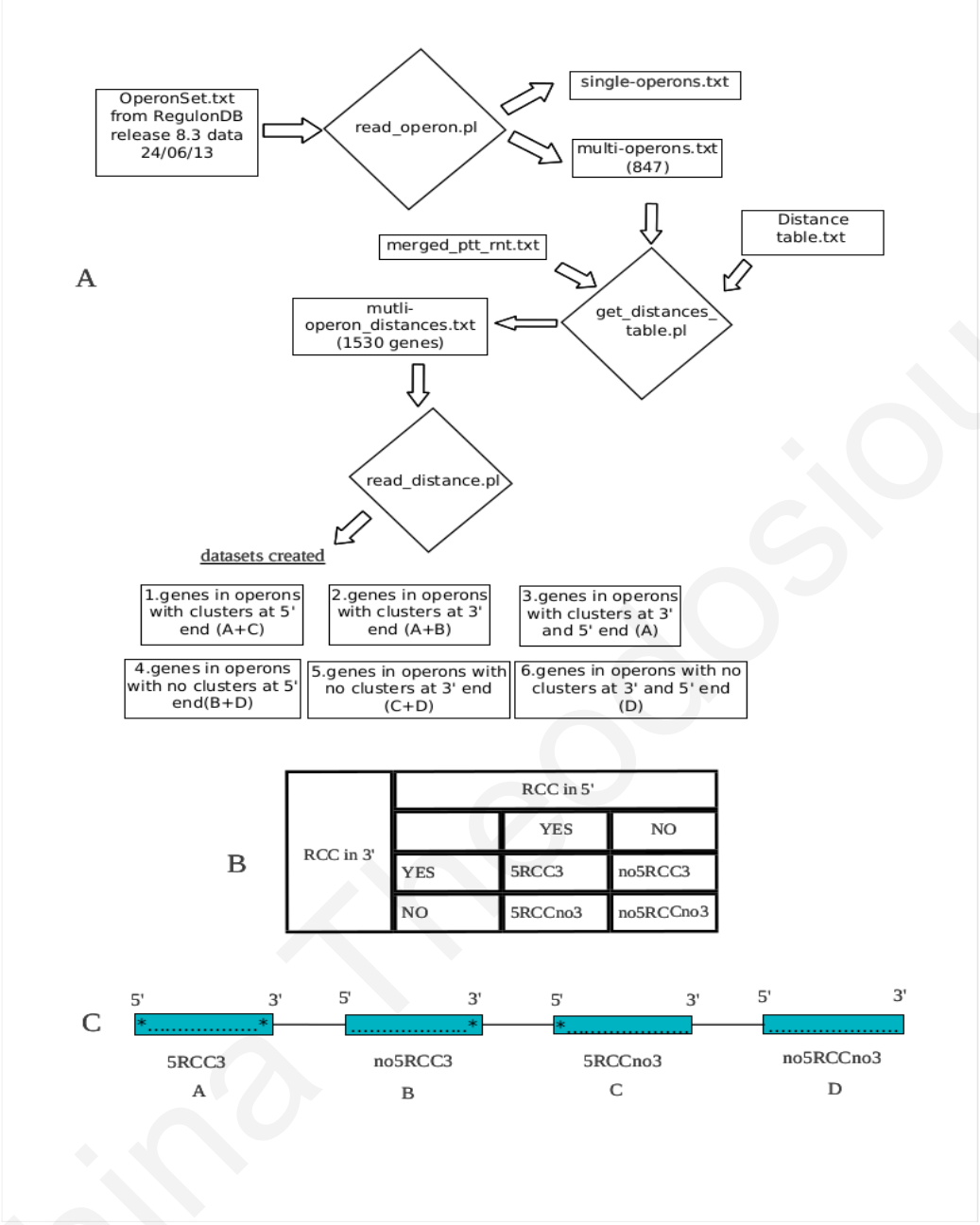

Figure 17: Calculation of distance between neighboring genes of the operons.

3.3 Results and discussion

3.3.1 Benchmarking RCC detection methods in LaTcOm

Searching for possible biological roles of the hidden code of rare codons, RCCs were detected with LaTcOm in the coding genome of *E. coli* K12. In this study, we analyzed the RCCs with the three detection methods implemented in LaTcOm. Initially, we followed a comparative analysis of the methods by applying two measures, the Mathews correlation coefficient (Matthews, 1975) and the segment overlap score (SOV) approach (Zemla et al., 1999). We run two versions of MCC and SOV analysis as already described in Methods. The results of the first version v1 MCC analysis (shown in Table 4 and Figure 18) do not reveal any strong correlation between the RCC detections methods of LaTcOm (all mean MCC values are below 0.5). Nevertheless, taking into consideration all statistical properties we show a mild correlation of MSS and %MinMax (considering both scales used). In the distribution histograms (shown in Figure 18) we identify two strong peeks, the first at MCC=0 and the second at MCC =1. A random MCC distribution creation (from randomly shuffled sequences) which is shown in Appendix 1 - Figure 41 and demonstrates that MCC random values are distributed near these two peeks (0 and 1), showing that these may be affecting the distribution of the results. Moreover, the distribution between MCC v1 values and randomly generated MCC v1 values are not the same with Wilcoxon Rank Test significance $p < 0.01$ (Table 5) discriminating the MCC values from random. The MCC results (v2) in which some genes were excluded (as described in methods) can be seen in Table 6 and Figure 19. In this alternative analysis the median MCC values as well as the overall MCC are higher than in version 1. %MinMax and MSS have a mean 0.448 and overall mean 0.458. The RiboTempo and MSS correlations in this test are the worst showing that the peaks seen in Figure 18 were affecting the results. Again the distributions differ from random distributions of MCC v2 (Appendix 1 - Figure 42 and Table 39).

Table 4: MCC distribution analysis (v1).
%MinMax was used with two different scales: Codon usage from Codon usage database (www.kazusa.or.jp/codon) and the tRNA concentrations from (Zhang et al., 2009). Mean, median and standard deviation were calculated on the distribution of MCC values for each set. The rightmost column represents the overall MCC values as described in Methods. Statistical properties were estimated with R the statistical environment (R Development Core Team, 2008).

Reference	Compared	Mean	Median	SD	Overall MCC
RiboTempo	%MinMax(cu)	0.257	0.000	0.406	0.146
RiboTempo	%MinMax(z)	0.220	0.018	0.363	0.173
RiboTempo	MSS	0.422	0.204	0.479	0.133
%MinMax(cu)	MSS	0.404	0.351	0.408	0.366
%MinMax(z)	MSS	0.385	0.371	0.374	0.422

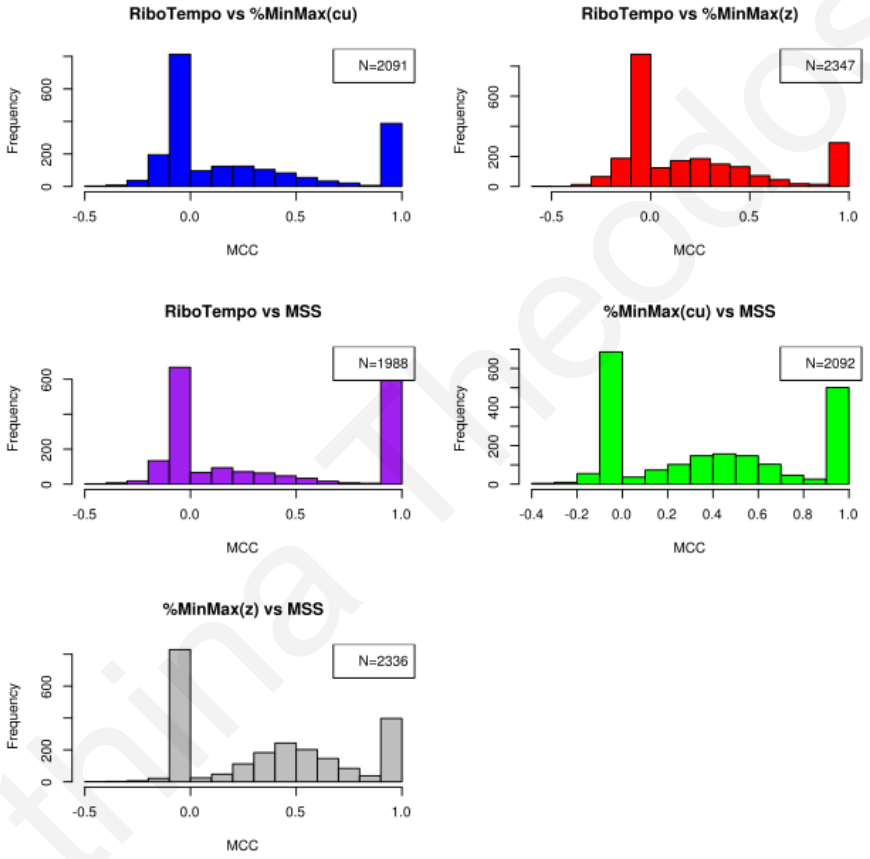

Figure 18: Distribution of MCC values (v1).
MCC values are calculated in the comparisons of reference and predicted set of genes. Analysis was made with all genes as described in the Methods section. The x axis shows the MCC value that can range from -1 to 1, the y axis shows the frequency from N compared MCC values. Graphs were generated with the R statistical environment (R Development Core Team, 2008).

Table 5: Wilcoxon rank test for MCC v1.
P-values of MCC distributions (v1) and random MCC distributions from shuffled sequences. The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Reference	**Compared**	**p-value**
RiboTempo	%MinMax(cu)	<0.00000002355
RiboTempo	%MinMax(z)	<2.2e-16
RiboTempo	MSS	0.05686
%MinMax(cu)	MSS	<2.2e-16
%MinMax(z)	MSS	<2.2e-16

Table 6: MCC distribution analysis (v2).
Version 2 excludes the comparison of some gene sets as described in Methods. See Table 4 for description.

Reference	**Predicted**	**Mean**	**Median**	**SD**	**Overall MCC mean**
RiboTempo	%MinMax(cu)	0.144	0.118	0.268	0.156
RiboTempo	%MinMax(z)	0.176	0.182	0.269	0.179
RiboTempo	MSS	0.118	0.080	0.254	0.136
%MinMax(cu)	MSS	0.351	0.387	0.277	0.391
%MinMax(z)	MSS	0.448	0.465	0.230	0.458

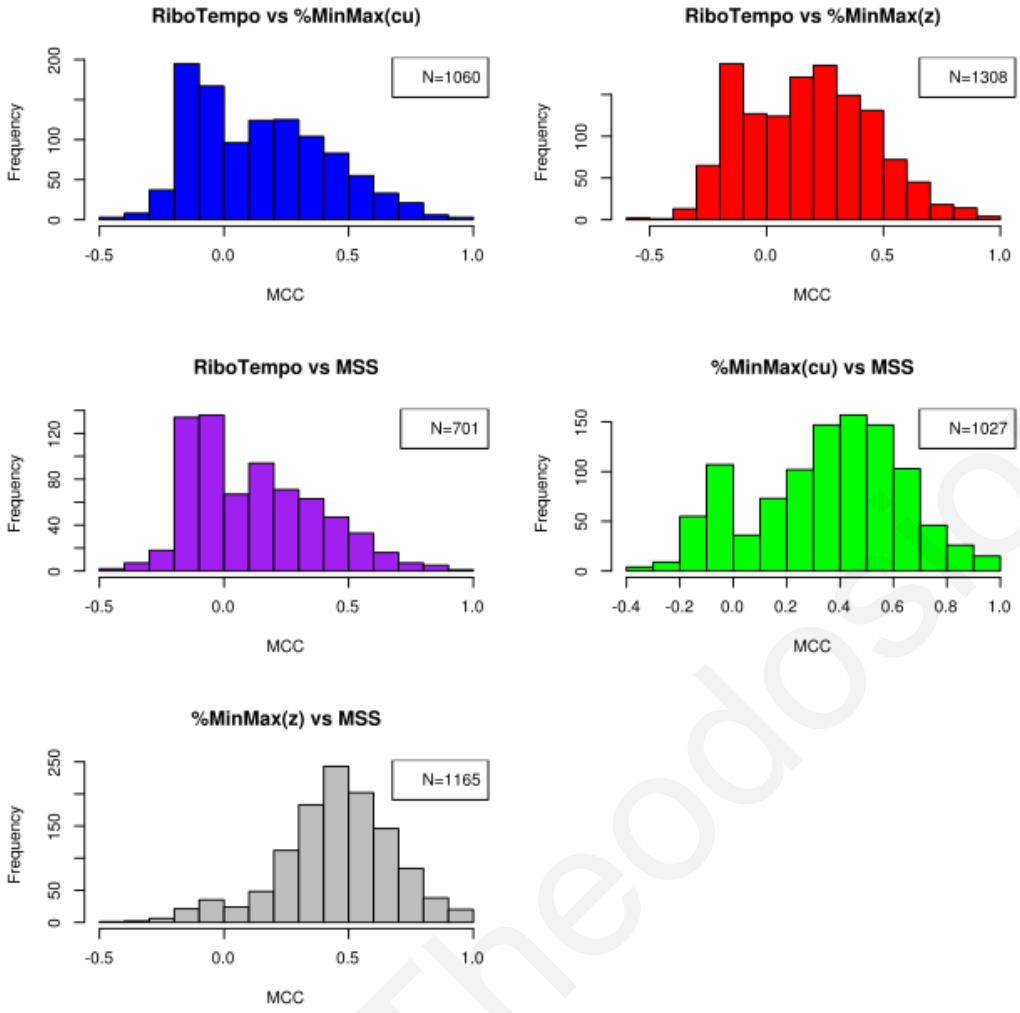

Figure 19: Distribution of MCC values (v2).
Description can be seen in Figure 18.

Table 7 shows the comparison of MCC distributions v2 with random MCC distributions from shuffled sequences using the Wilcoxon rank test, indicating, that the distributions are not the same with significance of p-value < 0.01.

Table 7: Wilcoxon rank test for MCC v2.
P-values for distribution of MCC distributions (v2) and random MCC distributions from shuffled sequences. The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Reference	**Compared**	**p-value**
RiboTempo	%MinMax(cu)	< 2.2e-16
RiboTempo	%MinMax(z)	< 2.2e-16
RiboTempo	MSS	1.09e-7
%MinMax(cu)	MSS	< 2.2e-16
%MinMax(z)	MSS	< 2.2e-16

The SOV analysis gave similar results with MCC. From SOV v1 (Table 8 and Figure 20) we demonstrate that %MinMax (cu) and MSS have higher segmental overlap than any other method comparison (Figure 21 and Table 9 for SOV version 2 analysis). Random distributions of SOV from shuffled sequences are differently distributed (Appendix 1 - Figure 43 and Table 40 for SOV v1 and Figure 44 and Table 41 for SOV v2) with Wilcoxon test significance of $p < 0.01$(as shown in Table 10).

Table 8: Statistical properties for SOV v1 distributions.
Distribution of SOV values (v1) calculated in the comparison of reference and predicted set of genes. Analysis was made with all genes as described in Methods. Statistical properties were estimated with R statistical environment (R Development Core Team, 2008).

Reference	**Predicted**	**Mean**	**Median**	**SD**
RiboTempo	%MinMax(cu)	59.990	55.800	25.089
RiboTempo	%MinMax(z)	54.310	50.100	24.485
RiboTempo	MSS	69.910	67.700	27.245
%MinMax(cu)	MSS	69.300	70.340	23.100
%MinMax(z)	MSS	66.450	65.400	23.728

Table 9: Statistical properties for SOV v2 distributions.
Distribution of SOVvalues (v2) calculated in the comparison of reference and predicted set of genes. Analysis was made by excluding some compared genes as described in Methods. Statistical properties were estimated with the R statistical environment (R Development Core Team, 2008).

Reference	**Predicted**	**Mean**	**Median**	**SD**
RiboTempo	%MinMax(cu)	50.960	49.700	18.126
RiboTempo	%MinMax(z)	47.920	45.900	18.721
RiboTempo	MSS	51.490	49.400	17.398
%MinMax(cu)	MSS	61.370	61.600	18.663
%MinMax(z)	MSS	59.950	59.900	20.293

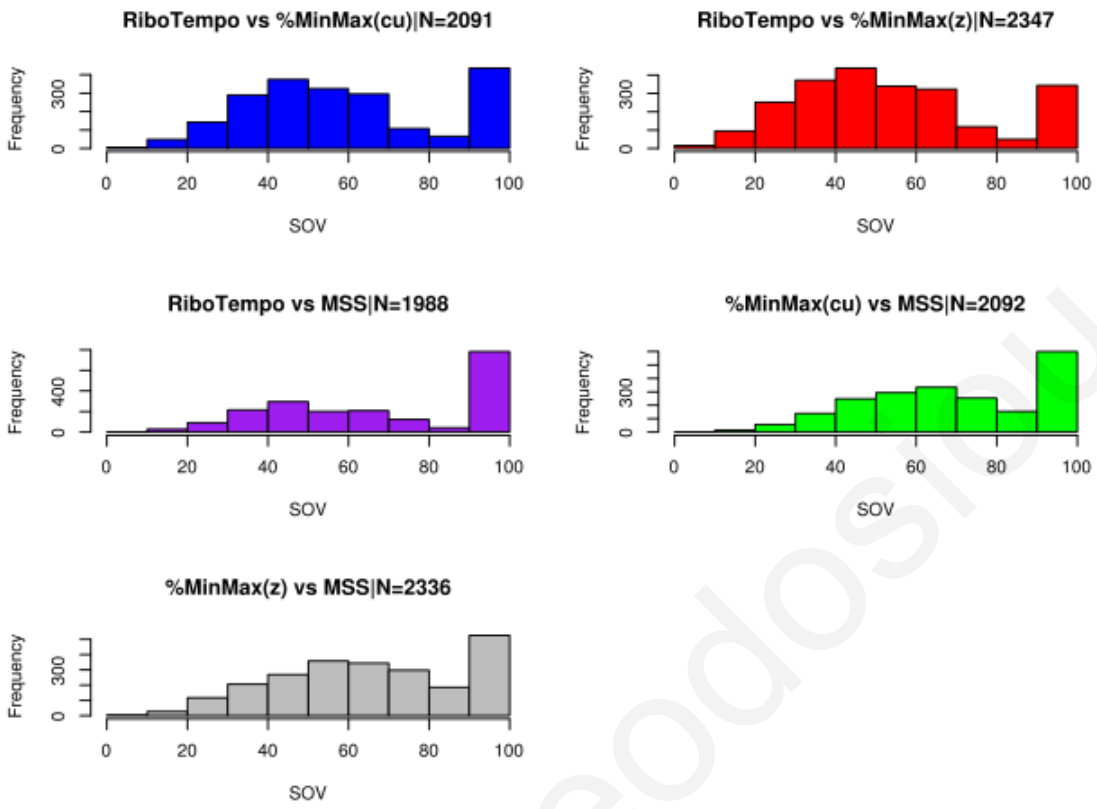

Figure 20: Distribution of SOV values (v1).
Values were calculated from the comparison of reference and predicted set of genes. Analysis was performed with all genes as described in the Methods section. The x axis shows the SOV values for each comparison and the y axis the frequency in data of N comparisons for each histogram. Graphs were generated with R statistical environment (R Development Core Team, 2008).

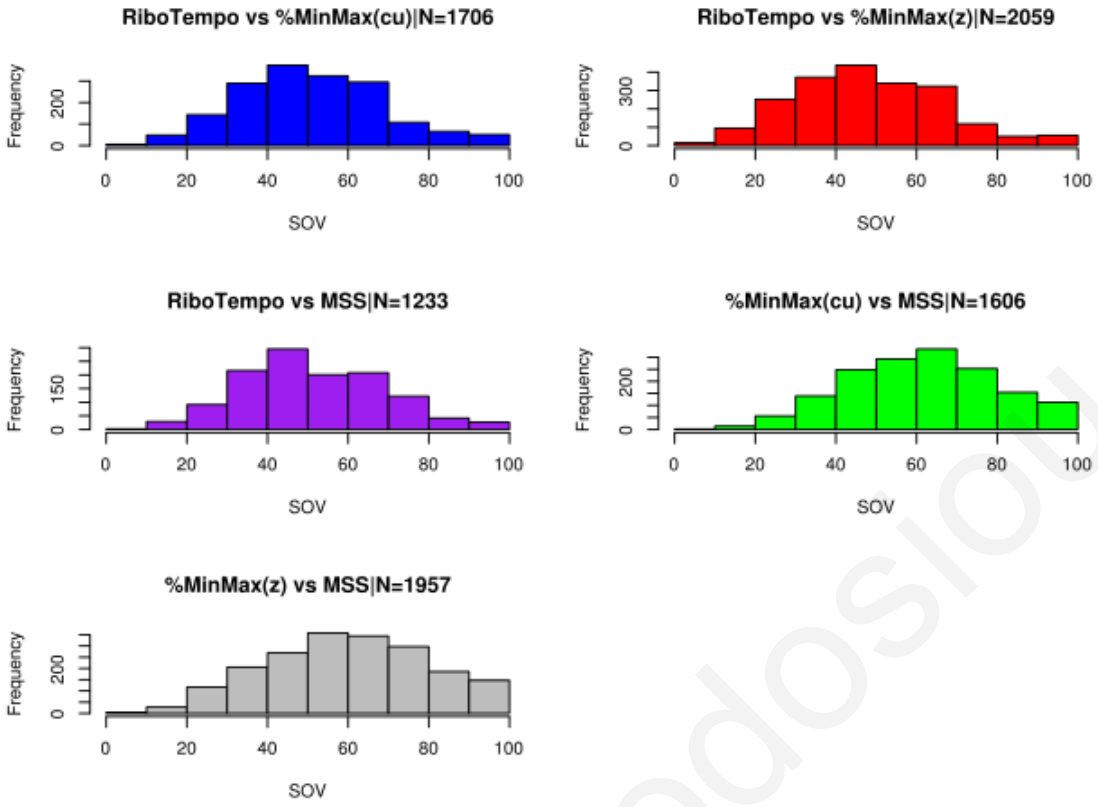

Figure 21 : Distribution of SOV values (v2).
Analysis was made with some compared genes excluded as described in Methods. Description can be seen in Figure 20.

Table 10: Wilcoxon rank test for SOV v1 with shuffled sequences.
P values were estimated for distribution of SOV distributions (v1) and random SOV distributions from shuffled sequences. The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008). The SOV v2 test showed identical results (data not shown).

Reference	**Predicted**	**p-value**
RiboTempo	%MinMax(cu)	< 2.2e-16
RiboTempo	%MinMax(z)	< 2.2e-16
RiboTempo	MSS	< 2.2e-16
%MinMax(cu)	MSS	< 2.2e-16
%MinMax(z)	MSS	< 2.2e-16

As far as we know, this is the first effort to compare new and existing methods for RCC detection. We identified some correlations in the methods but our analysis reveals that there is no clear consistency between the different approaches. We expected RiboTempo and %MinMax to detect similar RCCs since they are both sliding window approaches but surprisingly from our results, %MinMax and MSS seem to have a better segmental overlap than any other comparison. As we clearly see in our results %MinMax (z) over

predicts RCCs, therefore when a number of these were excluded from the comparison (some of the MCC=0) the comparisons were better. Finally, we conclude that the best correlation can be assumed from the overall MCC value and median when extreme cases are discarded (no RCCs identified on both or RCCs identified in only one set) and %MinMax and MSS have the greatest correlation.

3.3.2 General characteristics of RCCs in the *E. coli* coding genome

In order to characterize the results of RCC detection some general statistical properties are provided below. Table 11 shows the actual number of sequences analyzed taking into account the restrictions of each method. When %MinMax is used with the tRNA Zhang et al., (2009) scale, at least one cluster is found in approximately 76% of the sequences. The number of sequences with at least one cluster is reduced to 60% when codon usage is used as in the actual implementation of the method (Clarke and Clark, 2008). RiboTempo and MSS detected at least one cluster in half of the genes analyzed. Although the numbers are similar, this does not mean that the detected RCCs were found in the same sequences. Overlapping IDs with at least one RCC and with no RCCs are shown in the Venn diagrams Figure 22 and Figure 23 respectively.

Table 11: Information for RCCs in *E. coli* analyzed.
Column "All seq" shows the number of sequences that passed the control in LaTcOm and are analyzed further. "Seq_no_RCC" are the number of sequences with no detected RCCs. "Least_one_RCC" shows the number of sequences that have at least one cluster detected. "Seq>=200withRCC" shows the number of sequences with length >=200 and with at least one cluster (total number of sequences with length >=200 is 2834).

Method	**All_seq**	**Seq_no_RCC**	**Least_one_RCC**	**Seq length >=200 & RCCs**
%MinMax (cu*)	4128	1616	2512	2016
%MinMax (z*)	4128	991	3137	2508
RiboTempo	4128	2044	2084	1795
MSS	4136	2051	2085	1694

* z: scale from (Zhang et al., 2009) cu: Codon usage from Codon usage database(www.kazusa.or.jp/codon/)

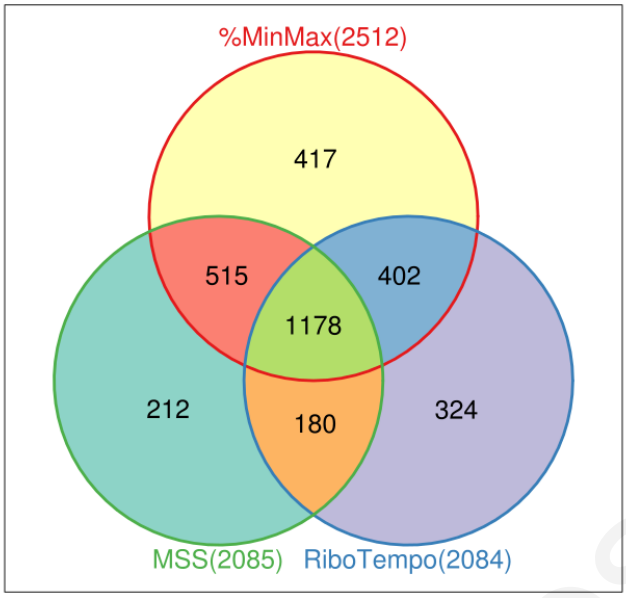

Figure 22: Venn diagram for sequences with at least one RCC.
The diagram illustrates the number of overlapping sequences with at least one RCC detected with the three methods implemented in LaTcOm. Results of %MinMax with the codon usage scale is shown. All sequences were used in this analysis. Diagrams were constructed in the R statistical environment (R Development Core Team, 2008) using the 'Vennerable' package.

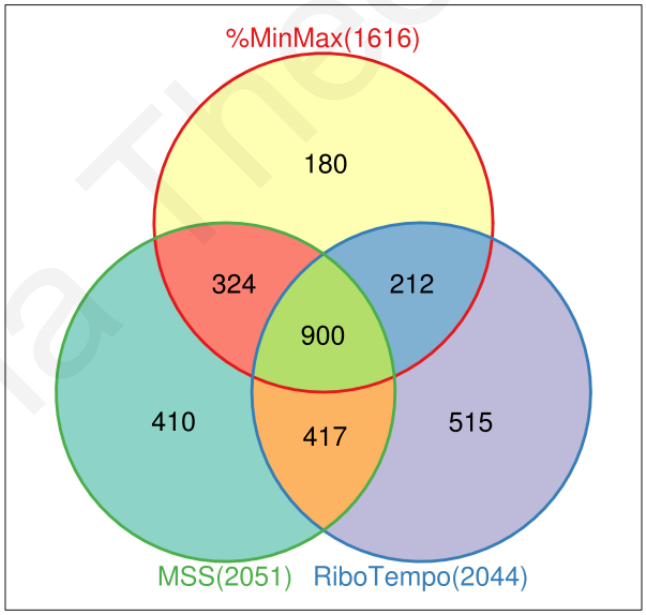

Figure 23: Venn diagram for sequences with no RCCs.
The diagram illustrates the number of overlapping sequences with no RCC detected with the three methods of LaTcOm. Results of %MinMax with the codon usage scale is shown. All sequences were used in the analysis. Diagrams were constructed in the R statistical environment (R Development Core Team, 2008) using the 'Vennerable' package.

All three methods detected at least one RCC in 1178 sequences (Figure 22) and no RCC in 900 sequences (Figure 23). There seems to be a concordance only between half of the genes in *E. coli*. These results were used in the Gene ontology analysis described in a following section.

Furthermore, Figure 24 demonstrates the codon coverage within the RCCs. In this plot it is demonstrated once more that the %MinMax over detects RCCs compared to the other methods when the Zhang et al., 2009 scale is used. %MinMax with codon usage has more comparable results with the other two methods as far as this analysis illustrates. Finally, we propose that when using %MinMax in LaTcOm, the tRNA abundance scale described in Zhang et al., 2009 should be avoided.

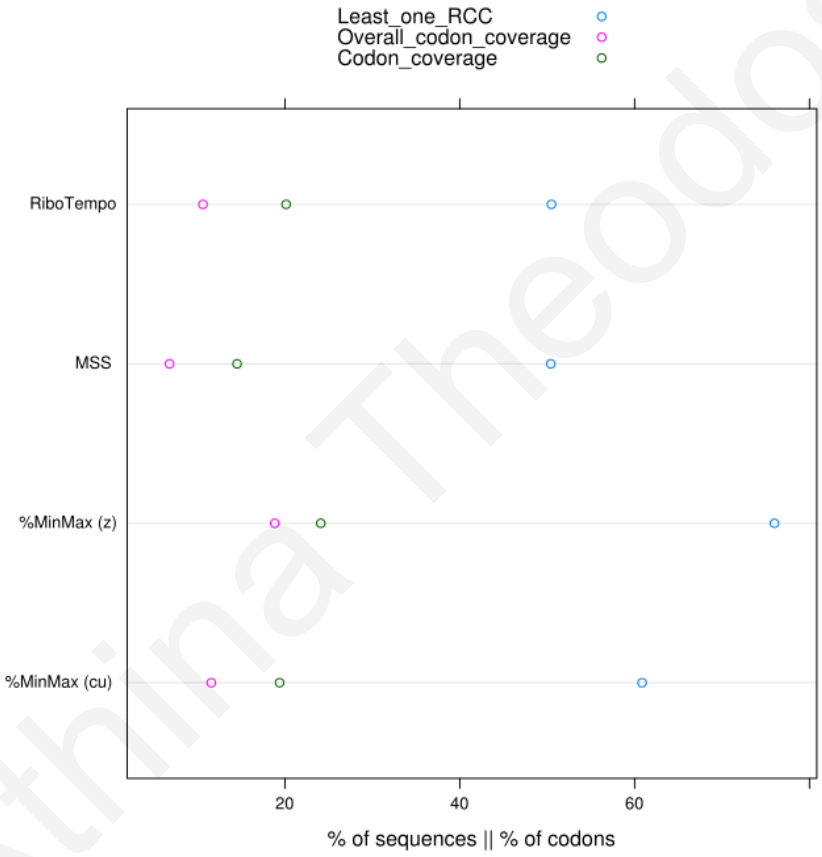

Figure 24: Illustration for sequences with at least one RCC and codon coverage.
The plot demonstrates the percentage of sequences with at least one RCC (in blue- N=4128 for both sliding window methods and N=4136 for MSS), 2). The codon cluster coverage (in green) and 3) the overall codon coverage (in pink). Total codons=1307919 analyzed for sliding window methods and total codons=1308046 for MSS.

In Figure 25 we present graphically informative data regarding the number of clusters found in sequences. For approximately half of the sequences no clusters were detected using MSS and RiboTempo, whereas ~40% of the sequences do not have any RCCs detected with %MinMax. All methods detect approx.. 20-30% sequences with one cluster. All methods detect 15-25% two or three clusters with the exception %MinMax (z) which has more the 30%. Last, 5-10% of the sequences have four or more clusters detected.

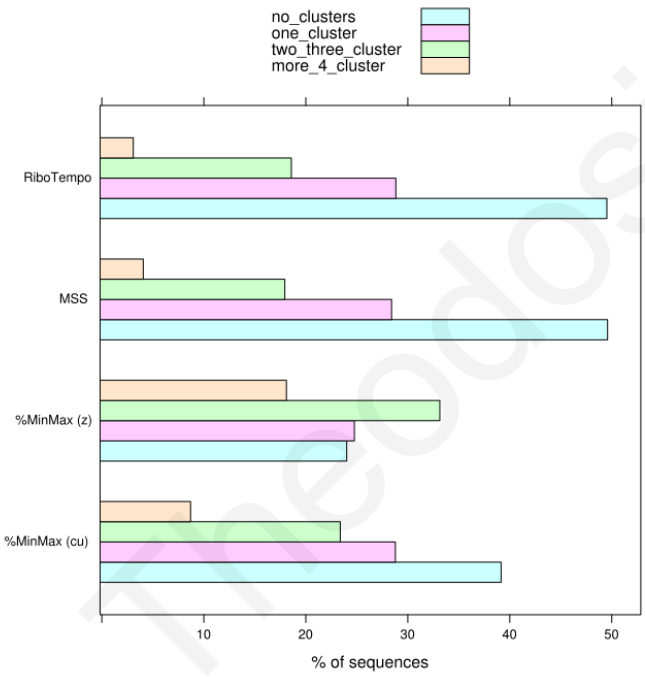

Figure 25: Percentage of sequences with different number of clusters.
Informative illustration for the percentage of *E. coli* sequences with no clusters, with one cluster, with two or three clusters and last with equal to four or more clusters. N=4128 for sliding window methods and N=4136 for MSS.

From this plot it is shown that RiboTempo and MSS are correlated when numbers are concerned. Moreover, once again it is demonstrated that %MinMax with tRNA Zhang et al., 2009 scale, over detects RCCs with approximately half of the sequences detected with >=2 RCCs. The results are closer to the other detection methods when codon usage is used in %MinMax.

3.3.3 Cluster lengths of detected RCCs

With the RCC detection results available from LaTcOm, it was interesting to see if there was a cluster length preference for the different RCC detection methods. In the sliding window approaches (%MinMax and RiboTempo) we had set the window=19 and for MSS we set cluster length=15. For MSS this is the minimum length of reported clusters. It was expected that the thresholds would be preferred among the RCCs. Due to the fact that LaTcOm joins RCCs if they are identified in close regions (see Methods in Chapter 2) we expected also higher lengths to be detected. Indeed as shown in Figure 26 in MSS ~ 56% of clusters have cluster length between 15-20, with 15 being the top percentage. %MinMax detected of 25-29% of cluster at cluster length 19 and 20 whereas 43-44% of clusters are detected with clusterlengths 21-30.

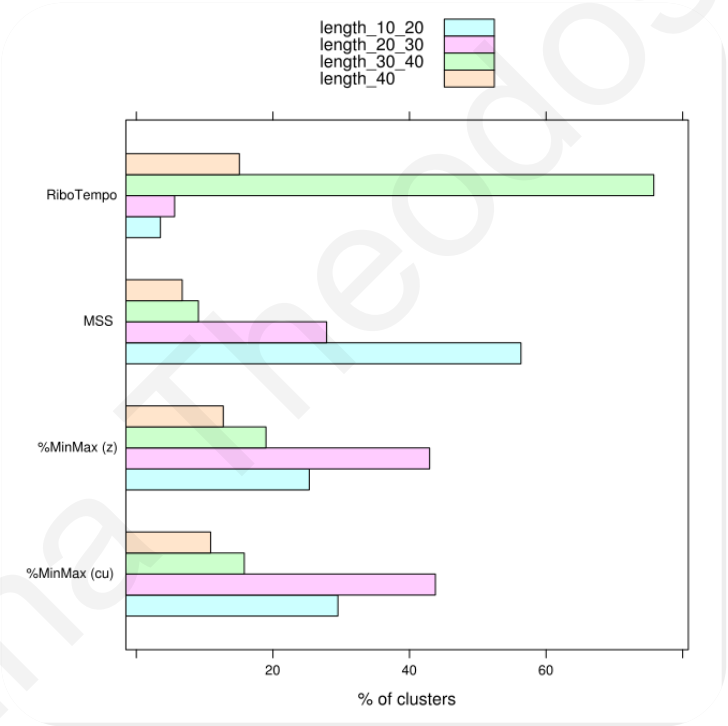

Figure 26: Cluster length distributions.
Percentage of clusters wiith length 10<=length<20; 20<=length<30; 30<= length<40 and length >= 40. Total number of RCCs: N = 5332 for %MinMax (cu), N=8380 for %MinMax (z), N= 3545 for RiboTempo, N =3764 for MSS.

With the exception of RiboTempo, all other detecting schemes show the same length distribution, i.e. monotonically decreasing frequency with increasing length (Figure 27 and Table 12). Interestingly, RiboTempo detected most of clusters at cluster lengths 30-40. In more detail, nearly 71% of the RCCs detected with RiboTempo have cluster length 37. This is most probably caused by two parameters, the window size, which in this case was 19, and the fact that overlapping windows are merged into one by our methodology. Nevertheless, this is not the case with %MinMax which is also a sliding window approach and overlapping windows are also merged. In order to check this we run the sliding window algorithms RiboTempo and %MinMax with different window sizes (see in Appendix 1 - Figure 45 and 46 respectively). These graphs also confirm that with RiboTempo detection, most clusters have cluster length = (2 x window)-1, which means that the algorithm, or the combination of the algorithm and the merging of nearby clusters by LaTcOm is introducing bias. However, %MinMax which is also a window based approach, is only influenced by window size.

The monotone decline in frequency of MSS clusters with cluster length possibly indicates fortuitous clustering of "slow codons" as a major source of RCCs. In the absence of selection, the chance of finding K consecutive slow codons declines exponentially with K.

Table 12: Statistical properties for cluster length distributions
Distributions of Figure 27 are displayed rounded to two decimals. Mean, median and standard deviations are displayed.

Method	Mean	Median	Standard deviation
%MinMax (cu)	28.38	24.00	14.34
RiboTempo	39.17	37.00	12.02
MSS	23.58	19.00	13.86

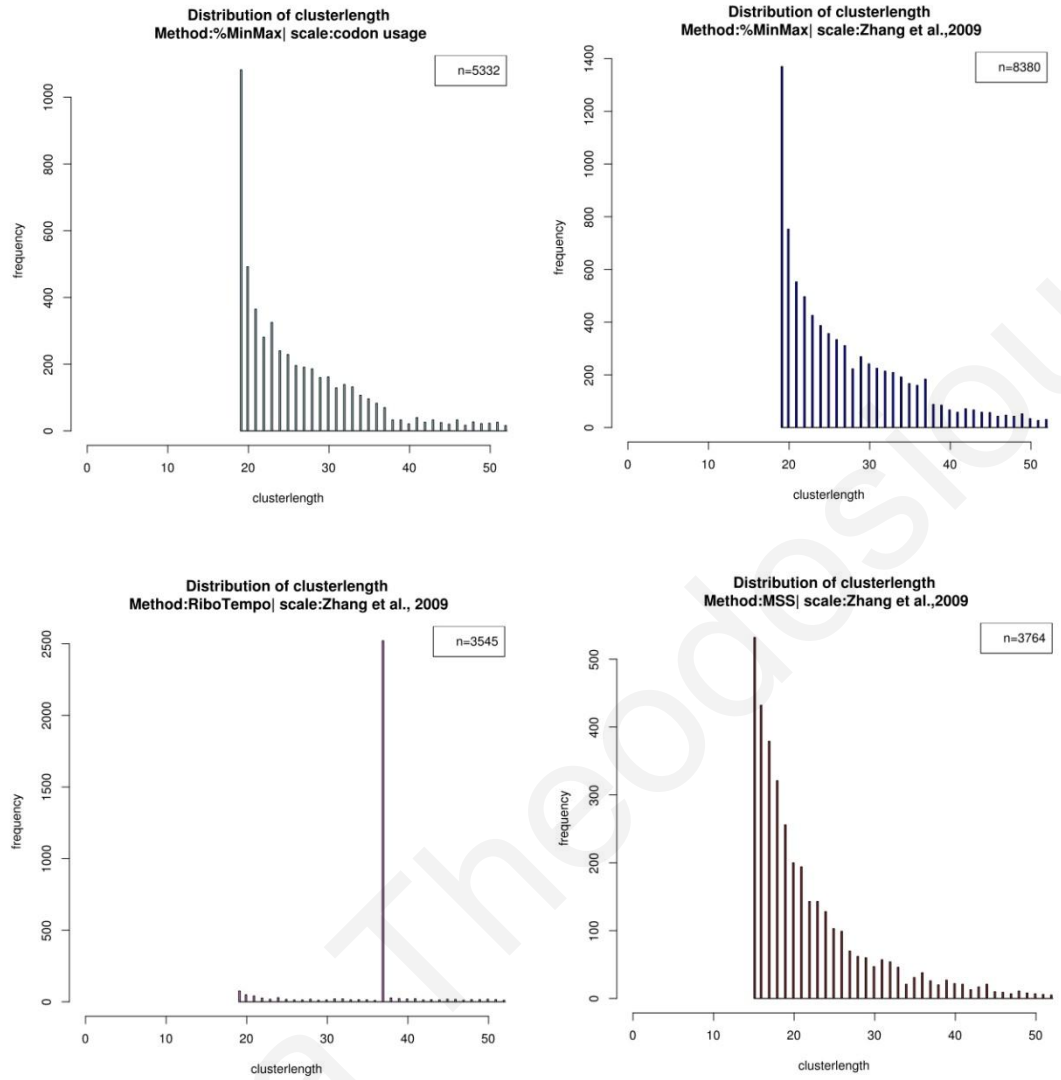

Figure 27: Detailed cluster length distribution.
Histograms show for all methods the frequency of clusters along with specific cluster length. The clusterlength values were truncated to maximum length 50 for visualization purposes; n represents all the clusters analyzed

.

3.3.4 RCCs at the 5' and 3' termini of *E. coli* sequences

The general enrichment of RCCs at the 3' and 5' terminal sites was shown in previous studies. Zhang and coworkers (Zhang et al., 2009) observed a “local minimum” at the 5' end site of sequences in *E. coli*. Moreover, in a more detailed analysis regarding the 5' and

3' end sites", Clarke and Clark (Clarke and Clark, 2010) showed an enrichment of RCCs at the 5' and 3' gene termini of genes from *E. coli* with the %MinMax tool (Clarke and Clark, 2008). In their study, the reported that nearly half of the genes (with more than 268 codons) have a cluster at the first 50 windows. In our study, we considered sequences greater or equal to 200 codons (that is 1694 sequences with at least one cluster in MSS). Considering all RCC detection methods, we confirm previous observations, that 5' and 3' termini are enriched in RCCs (Figure 28). We demonstrate that in 45% of sequences analyzed, with at least a cluster detected, there is a RCC at the first 100 codons (Figure 28 A) and 40% of the sequences have a cluster at the last 100 codons (Figure 28 B). From the total clusters predicted ~35% of the clusters are predicted at the first 100 codons (Figure 28 C) and ~25% at the last 100 codons (Figure 28 D).

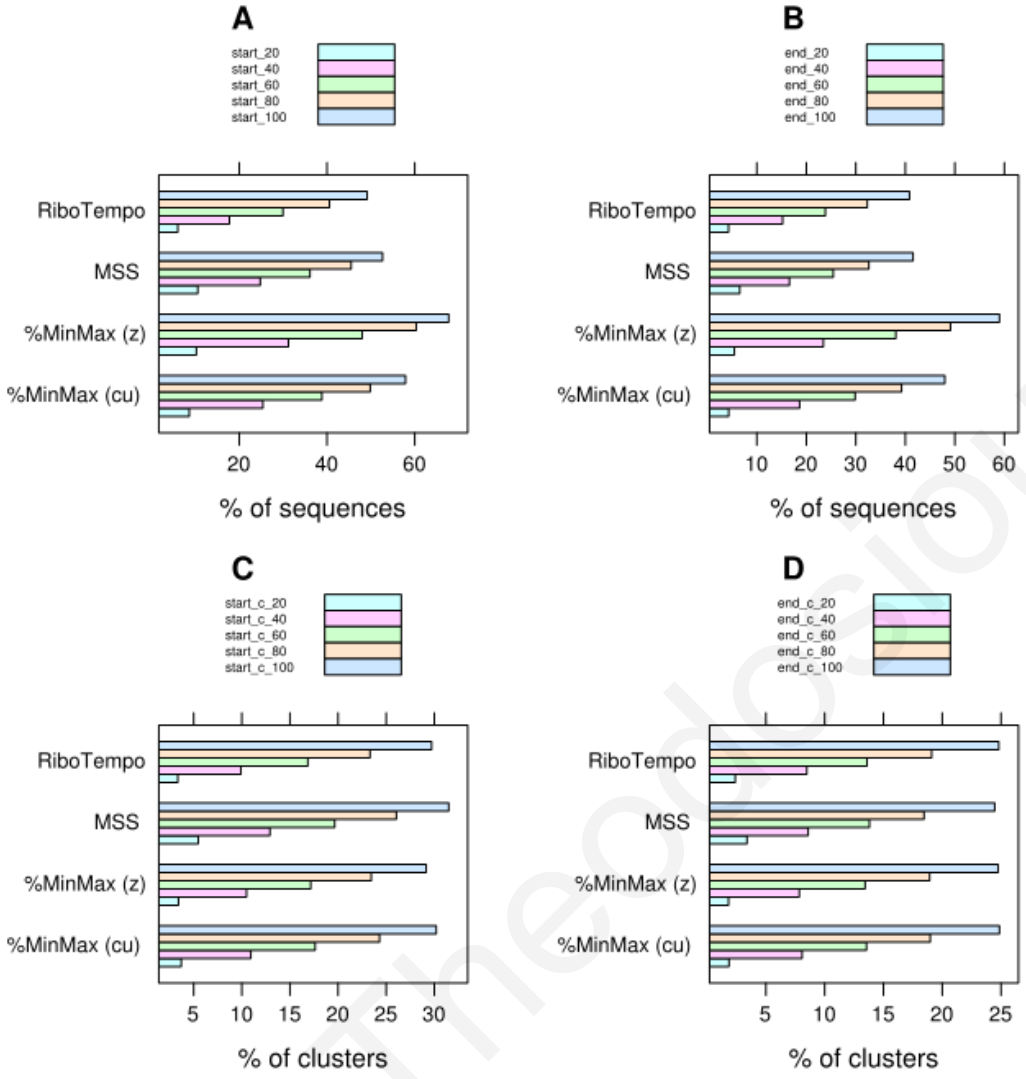

Figure 28: RCCs at the 5’ and 3’ termini.
Bar charts demonstrate the existence of RCCs at the 5' and 3' end sites of sequences in the *E. coli* genome detected with the different methods of LaTcOm. A) Percentage of sequences with RCCs at the 5' end site B) Percentage of sequences with RCCs at the 3' end site C) Distribution of clusters at the 5' end D) distribution of clusters at the 3' end. For all charts only sequences with more or equal than 200 codons were taken into account. Total sequences are sequences with at least one cluster detected. %MinMax(cu) N=2016, %MinMax(z) N=2508, RiboTempo N=1795 and MSS N=1694. Total numbers of detected clusters are: %MinMax(cu) N=4689, %MinMax(z) N=7523, RiboTempo N=3220 and MSS N=3273.

Next, we analyzed the positions of RCCs in sequences with respect to the distance from 5' and 3' terminals of the coding sequence. Figure 29 demonstrates the distribution of the distance of RCCs from the 5’ terminal with MSS method. The distributions with the other methods are similar (data not shown).

The distance is estimated in codons from the start position of an RCC back to the 5' end.

The median distance is estimated to be 230 codons with a higher increase at the first 100 codons.

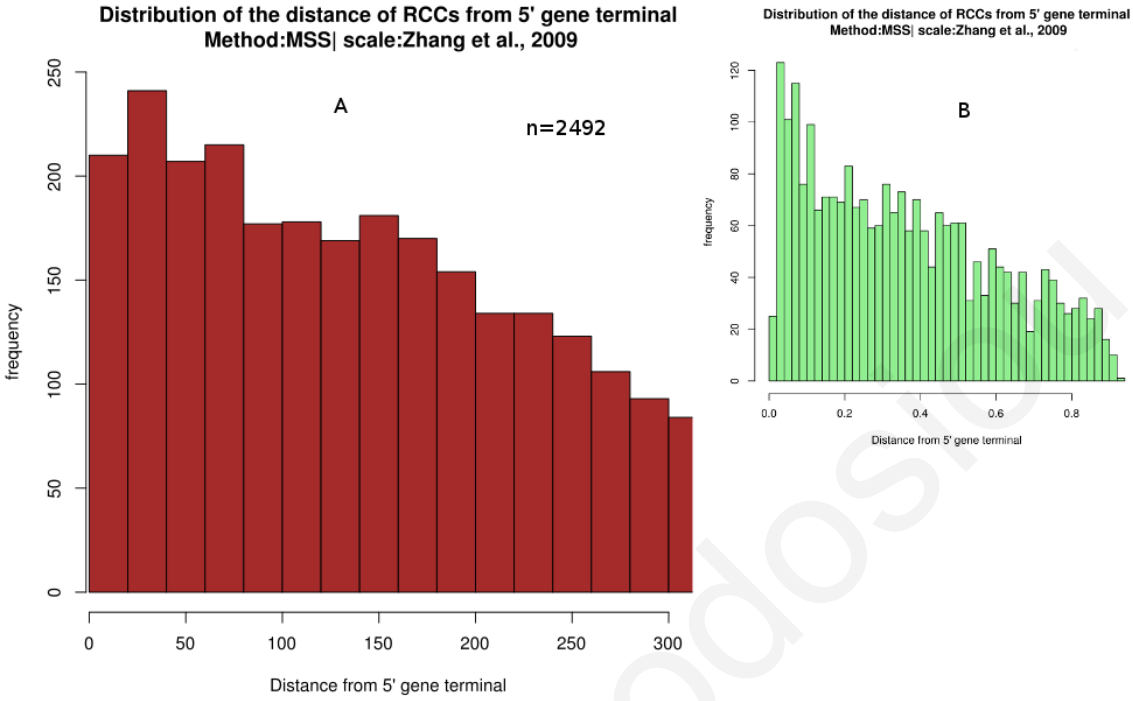

Figure 29: RCC distance distribution from 5' terminal.
The distance is measured in codons from the start point of each RCC. RCCs on the extremities* were excluded. The sequences within the analysis have length more than 200 residues and have at least one cluster detected. Total number of clusters detected for MSS was n=3273 (Median=225.2, Mean=291, and standard deviation=228.82). Total number of clusters with distance 300 or less was n=2492. A) Distribution of clustelength B) Normalized distribution with sequence length. Graphs were generated in R statistical environment (R Development Core Team, 2008).
*see Methods in Chapter 2

We know that the distribution may be biased due to the fact the some coding genes have small length (201-250 since we only take into account sequence greater than 200 codons), therefore for these sequences an RCC that is at 150, may be closer to the 3' and vice versa. Therefore we normalized the distance with sequence length in order to get more accurate distributions.

In Table 13 we see the comparison of the distributions by the different methodologies to identify any potential deviation. From Wilcoxon rank test p-values, none of the comparisons produced significant results ($p < 0.01$) therefore the distributions cannot be considered different. From this point on we choose not to present data from %MinMax

using the scale described in Zhang et al., 2009 due to over detection of RCCs.

Table 13: Wilcoxon rank test for comparing the distance distributions of RCCs with the different methods from 5' terminal.
The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Method	**Method**	**p-value**
%Minmax (cu)	MSS	0.8954
%Minmax (cu)	Ribotempo	0.8456
Ribotempo	MSS	0.9628

The distribution of RCC distances with MSS from the 3' termini can be seen in Figure 30.

The distributions between methods cannot be considered different (Table 14).

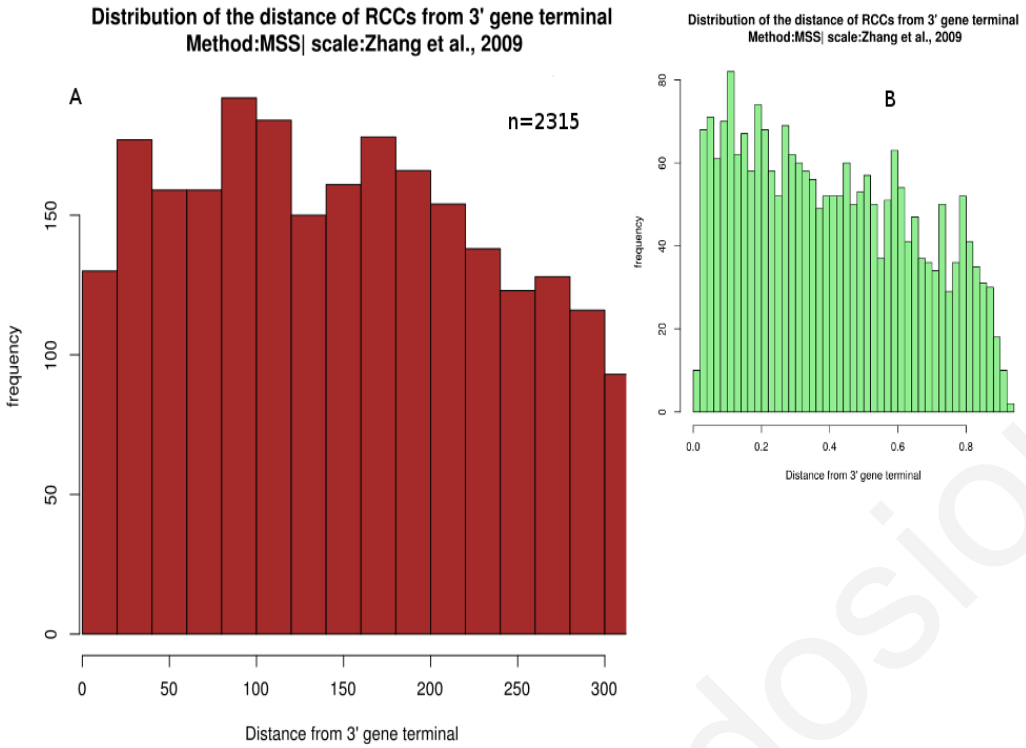

Figure 30: RCC distance distribution from 3' gene terminal.
The distance is measured in codons from the end point of each RCC. RCCs on the extremities* were excluded. The sequences within the analysis have length more than 200 residues and have at least one cluster detected. Total number of clusters detected are n=3273 (Median=252.8, Mean=328, standard deviation=230.97). Total number of clusters with distance from 3' terminal with 300 or less was n=2315. A) The distance distribution B) normalized distance distribution of sequence length.
*see Methods in Chapter 2

Table 14: Wilcoxon rank test for comparing the distance distributions of RCCs from 3' terminal with the different methods.
The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Method	Method	p-value
%Minmax (cu)	MSS	0.982
%Minmax (cu)	Ribotempo	0.912
Ribotempo	MSS	0.888

The distribution median seems to be slightly higher for 3' than for 5' termini placing the RCCs closer to the 5' compared to the 3' termini. To investigate whether there is a different pattern of the distributions between the two termini the Wilcoxon Rank test was applied, but no significant difference was detected (see Appendix 1 - Table 42).

Furthermore, we analyzed the distribution of distances of the first and last RCCs from the terminals as seen in Figure 31 and Figure 32 respectively. The median distance of the first RCCs is estimated at 130 codons from the 5 'end and the median distance of the last is estimated at 170 from the 3' end. The distributions of the methods cannot be considered different based on significance measures (Table 15 and Table 16).

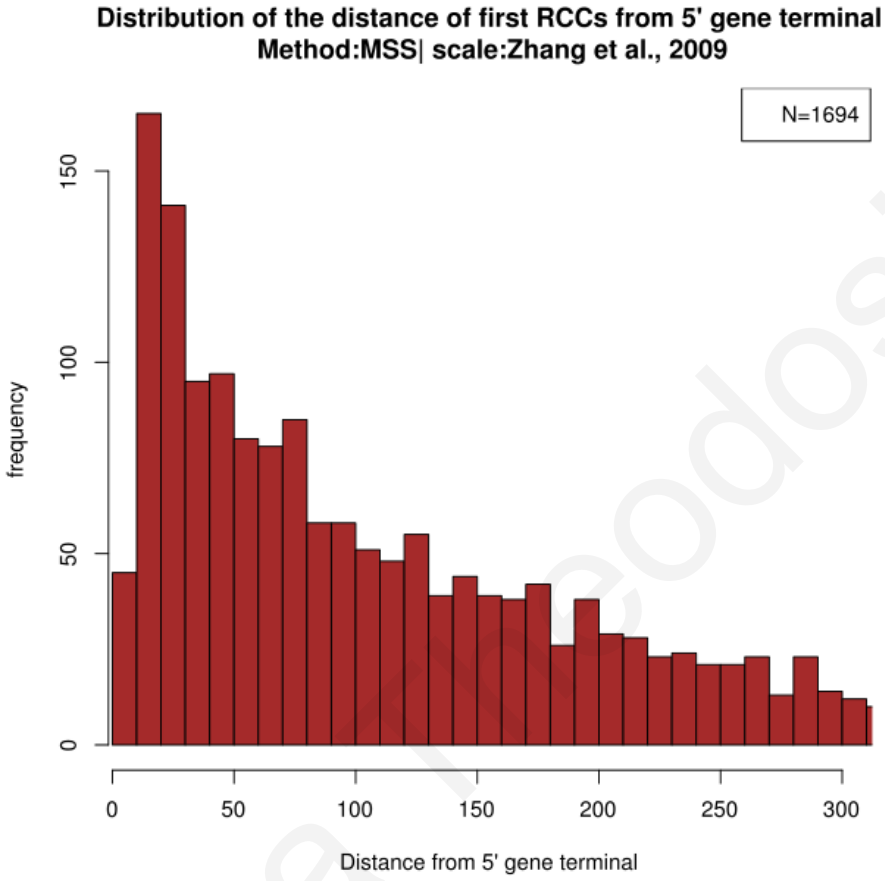

Figure 31: Distance distribution of the first RCCs detected with MSS from 5' terminal.
Total number of clusters detected using MSS N=1694 (Median=130.7, mean=185.8, standard deviation=127.13). Graphs were generated with the R statistical environment (R Development Core Team, 2008).

Table 15: Wilcoxon rank test for comparing the distance distributions of first RCCs from 5' terminal
The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Method	Method	p-value
%Minmax (cu)	MSS	0.767
%Minmax (cu)	Ribotempo	0.382

Ribotempo	MSS	0.243

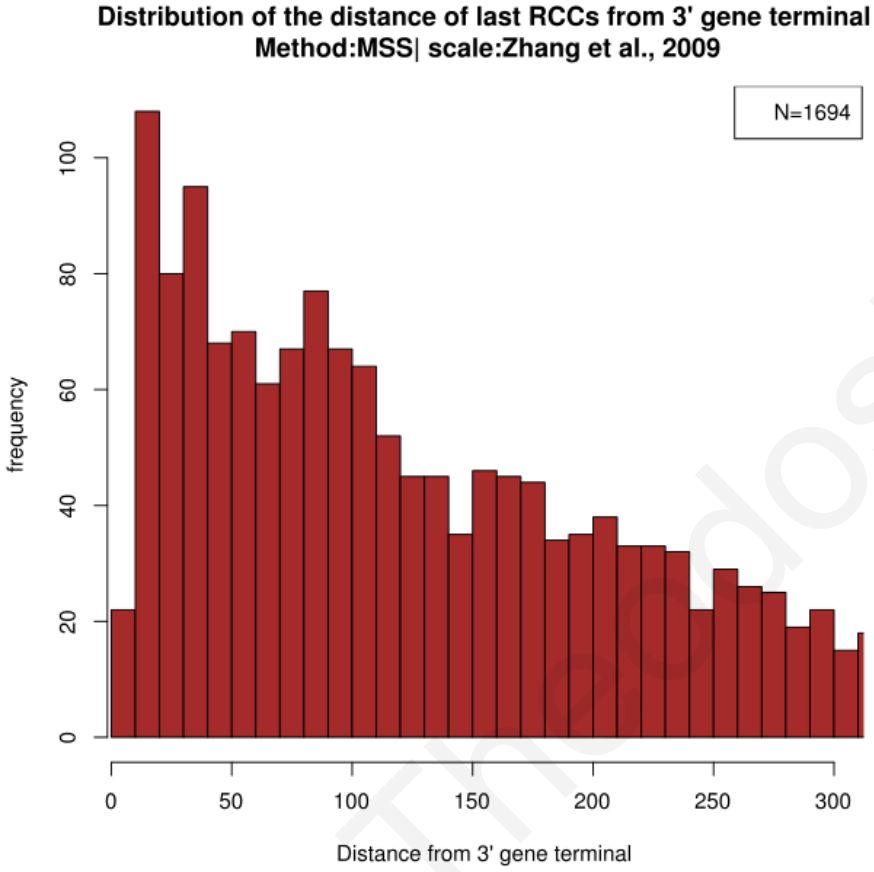

Figure 32: Distance distribution of the last RCCs detected with MSS from 3' terminal.
Total number of clusters detected using MSS N=1694 (Median=169.5; mean=231.8, standard deviation=156.19).

Table 16: Wilcoxon rank test for comparing the distance distribution of last RCC from 3' terminal of the different methods.
The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Method	**Method**	**p-value**
%Minmax (cu)	MSS	0.193
%Minmax (cu)	Ribotempo	0.686
Ribotempo	MSS	0.088

Moreover, distributions for first RCCs from 5' were compared with distributions from last RCCs at 3' and significant results were retrieved for MSS with Wilcoxon test p-value < 0.01 (Table 17). The fact that the first and last median values differ and the difference in distribution for MSS shows that there is a difference in preference regarding the distance from the two termini.

Table 17: P-values from Wilcoxon rank test for the RCC distance distribution between first RCCs at 5' and last RCCS at 3' terminus.
The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Method	**p-value**
%Minmax (cu)	0.080
Ribotempo	0.020
MSS	0.006

In order to evaluate a possible trend towards a specific distance between adjacent RCCs we also analyzed the distribution of these distances. Figure 33 shows the distribution of distance for adjacent RCCs detected with MSS. Most RCC distances are less than 200 codons apart and the median distance was estimated at 59 codons. Moreover, the Wilcoxon rank test showed that the distributions are different between methods with significance at $p<0.01$ (Table 18).

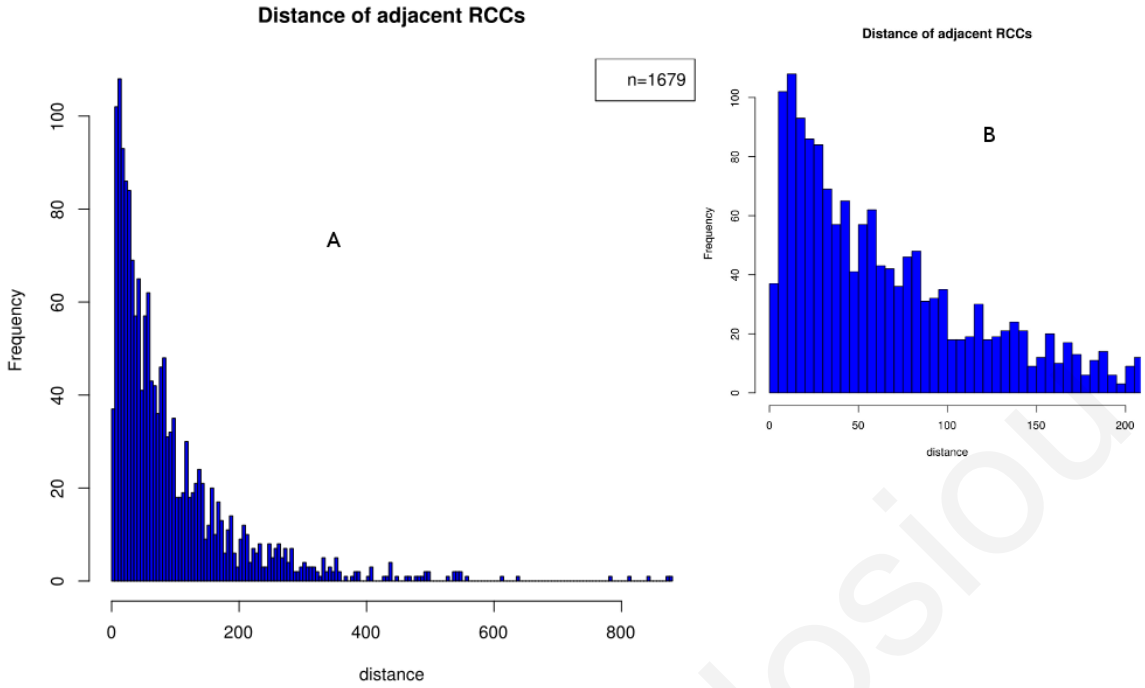

Figure 33: Distribution of distances between adjacent RCCs.
A) Distributions with MSS detections. Total in between distances detected are n=1679. Sequences with less than two clusters detected were discarded (Median=59, mean=92.12, standard deviation=102.76) B) Zooming distribution for illustration purposes.

Table 18: Wilcoxon rank test for comparing the distance distribution of adjacent RCCs as shown in Figure 33.
The wilcox.test () function was used from the R statistical environment (R Development Core Team, 2008).

Method	**Method**	**p-value**
%Minmax (cu)	MSS	<2.2e-16
%Minmax (cu)	Ribotempo	<2.2e-16
Ribotempo	MSS	0.039

From the aforementioned results we reveal a trend for RCCs to lie on both termini. Our results show that the first or last RCC prefer to be located closer to the 5' terminal than in 3' terminal. Nevertheless, the presence of RCCs at 5' and 3' ends of genes may reveal a universal functional role of RCCs or more specific to each site. As previously described, Clarke and Clark (Clarke and Clark, 2010) showed that rare codons are enriched at both 5' and 3' termini of genes. Another recent study has demonstrated with ribosomal profiling a "ramp" at the first 30-50 codons translated with low efficiency (Tuller et al., 2011). Several other studies also reported an enrichment of rare codons at the 5' site in different organisms (Allert et al., 2010; Fluman et al., 2014; Goodman et al., 2013; Pechmann et al.,

2014; Pechmann and Frydman, 2013). (Chartier et al., 2012) performed a large scale analysis on Pfam domains and identified rare codon clusters mainly in the 5' terminal of Pfam domains and not in the 3' terminal. As (Clarke and Clark, 2010) suggested the mechanism of translation is very different in prokaryotes versus eukaryotes therefore the signal in 3' may be prokaryotic specific.

A possible explanation of the enrichment of RCCs at the gene start is to keep the ribosome binding site free from stable mRNA structures (Bentele et al., 2013). Another possible suggestion is a functional role in secretion of secretory sequences (Burns and Beacham, 1985; Power et al., 2004), to allow correct folding of pre-secretory proteins (Zalucki and Jennings, 2007). Recent experimental evidence shed light in the strategies that prokaryotes and eukaryotes used to cause the arrest of the ribosome at the initiation site in order to correctly target membrane proteins to the translocon (Fluman et al., 2014; Pechmann et al., 2014). Nevertheless, we find this pattern not only in secretory sequences. We suggest that there is an additional biological role for the existence of RCCs at the gene start. It may serve as a regulator for possible attraction of other regulating factors specific for each protein.

Although the role of 3' RCCs is less discussed as also indicated in (Clarke and Clark, 2010), it has been suggested that RCCs at this side may pause the ribosome to allow further associations of the newly synthesized polypeptide with interacting molecules such as chaperones or factors involved in targeting and degradation (Hayes et al., 2002).

Our results possibly demonstrate that, the position of RCCs concerning the two termini have similar distributions. Nevertheless the distances of the first RCC from 5' and the last RCC from 3' have different distributions, indicating that RCCs at the gene termini may serve different purposes, at least as far as *E. coli* is concerned. The existence of RCCs at the 3' terminal may be preferred in multicystronic operons and since these are transcribed and translated together there might be a correlation of the RCC at the 3' termini and distance to the next genes. This hypothesis was addressed in the next experiments.

An alternative model might involve the existence of overlapping codes at the two terminals. In particular, such a model could account for extra transcriptional instructions coded in the DNA. In this case, it constrains the choice of codons and may lead to the choice of slow codons. Such a 'secondary message' may work from a small distance and

may be needed at the locus between genes (but may end up falling on one side but not the other).

3.3.5 RCCs in mutlicystronic *E.coli* operons

A hypothesis stemming from the previous analysis is that the existence of RCCs at the 3' site may be preferred in multicystronic *E. coli* operons and may be related to the distance of the next neighboring gene. Overlapping genes are common in prokaryotes (Normark et al., 1983) and have been proposed as a "shrinking" mechanism in order to fit maximum information in the minimum possible space. In the classical operon model, multiple genes are transcribed into a single polycistronic mRNA (Jacob and Monod, 1961). However, recent evidence supports internal transcription initiation or termination sites (Koide et al., 2009). In this work, we searched for a potential regulatory mechanism by examining the correlation of the existence of RCCs at terminal sites and the intergenic distance of genes within the same operon. Table 19 shows the number of genes in total and in operons that have RCCs at their terminal sites.

Table 19: Number of genes from total genes passing filtering and from operon genes that passed filtering that have a) RCC at 3' and b) RCCs at 5' according to MSS detections.

	Genes with RCC at 3'	**Genes with RCC at 5'**
Total genes (1694)	704	892
Operon genes (884)	352	447

Furthermore, we tested the differences of distributions with the Wilcoxon rank test regarding distances from upstream or downstream genes with respect to the existence of RCCs at the terminal. Nevertheless no significant correlation could be detected at $p < 0.01$ as shown in Table 20.

Table 20: Wicloxon Rank test p-values for comparison of the intergenic distance distribution.
The different datasets are detailed described in Methods (Figure 17). Datasets compared: 1: Distance of next gene between genes with RCC at 3' and between genes with no RCCs at 3'; 2: Distance of previous gene between genes with RCC at 3' and between genes with no RCC at 3'; 3: Distance of next gene between genes with RCC at 5' and between genes with no RCCs at 5'; 4: Distance of previous gene between genes with RCC at 5' and between genes with no RCCs at 5.'

	Datasets			
Method	**1**	**2**	**3**	**4**
%MinMax(cu)	0.55	0.12	0.18	0.55
RiboTempo	0.70	0.13	0.61	0.06
MSS	0.89	0.03	0.88	0.99

Furthermore we searched for a correlation between neighboring genes in operons. The question in search was what if a gene has an RCC at 3' terminal what does happen at the 5' site of the next gene? In Table 21 we demonstrate the number of the different cases. We applied Fisher exact test on the data but no significant correlation (<p 0.01) could be identified (p=0.21 for first column in Table 21 and p=0.03 for the operons in the second column).

Table 21: Number of distances comparisons calculated.

Comparison data with RCCs detected with MSS	**Comparisons**	
	In all genes (1137)	**Only in multi-gene operons (710)**
RCC at 3' and RCC at 5' of the next gene	128	84
RCC at 3' and not RCC at the 5' of the next gene	354	196
No RCC at 3' but RCC at 5' of the next gene	152	98
No RCC at 3' and no RCC at 5' of the next	503	332

We also searched for correlations between the distances of the genes and the distance distribution of these datasets are given in Appendix 1 - Figure 47. Wilcoxon rank test was applied to compare the four distributions but no significant difference was discovered demonstrating that distance of genes is not related with RCCs at the terminals. A recent work by (Quax et al., 2013) showed with comparative genomics analysis that differential translation is key determinant for gene expression of genes in operons and that codon bias shows the unequal protein production. We mapped RCCs detected with MSS, with the 3 model operon complexes of Quax et al., 2013 as shown in the publications Figure 1. Nevertheless, we could not reveal any correlation of the positions of RCCs in the operons (Appendix -Figure 48).

We additionally tested the correlation of strand and existence of RCCs in *E. coli* with no significant outcome (data not shown).

Based on our results regarding genes in multicystronic operons, our initial hypothesis is rejected. The RCCs found at the 3' terminal sites do not correlate with the distance to the next or previous gene. Nevertheless, an interesting finding was that in most cases at the absence of an RCC at the 3' site of a gene, then an RCC is also absent at the 5' site of the next gene. In fact, this observation holds not only for operon genes but also for all genes under study. In the cases where we find RCCs at the 3' site we hypothesize that the role of RCCs may be more general for the translational regulation of genes as the ones suggested by (Clarke and Clark, 2010) that translational pauses occur at the 3' termini to assist potential interaction with chaperones or other factors.

Additionally, it is important to notice that in prokaryotic organisms transcription and translation are coupled, thus the one process affects the other. One hypothesis could be that RCCs at the 3' terminal site may serve as a sign for the joined regulation of transcription and translation since both procedures are combined, although we cannot rule out that this potential signal may also be 5' terminal specific. Mechanistically, the 3' signal may be more easily associated with chaperone interaction, or preparing the ribosome to release the current synthesized polypeptide and continue with the next one in line. Nevertheless, more work needs to be done to unravel possible functional implications of RCCs at the 3' terminal site.

4 Correlating functional and structural properties with RCCs

4.1 Background

Our initial hypothesis is that the location of RCCs in coding genes correlates well with topological and structural characteristics/properties, while their existence may reveal higher level functional features. Moreover, it is intriguing to clarify whether a coupling exists between rare codon-mediated ribosomal attenuation and the biogenesis of αHTMPs, since most of them are integrated into the bilayer co-translationally (Osborne et al., 2005; Rehling et al., 2003). To explore these correlations we analyzed the positions of RCCs with respect to different topological, structural and functional groups for coding genes in *E. coli*.

4.2 Data and Methods

4.2.1 Collection of functional and structural data and correlation with RCCs

In order to ascertain whether *E. coli* RCCs are associated with specific features of the respective proteins, we examined the co-occurrence of a multitude of characteristics with RCCs. For this purpose, a number of gene sets were defined and the procedure to do so is described separately for each data set below.

Disordered proteins

Firstly, we downloaded all disordered proteins from the DisProt database ((Vucetic et al., 2005); http://www.dabi.temple.edu/disprot/index.php) (04/02/2013). This dataset is a FASTA formatted file with experimentally verified disordered regions mapped to protein sequences from different species. An example of the file is shown in Appendix 2 - Figure 49. With a custom script (*map_dis_uni.pl.)*, we extracted only the UniProt Accession numbers from *E. coli* that are available in each sequences' header in the DisProt FASTA file. Moreover, we mapped the UniProt Accession numbers with the corresponding GI

numbers with the ID mapping tool in UniProt (http://www.uniprot.org/). The *mapping_table_uniprot_gi.tab* was downloaded, which is a tab delimited table with the UniProt Accesion numbers as first column and the mapping GI number as second column. There are multiple GI numbers for a single accession in UniProt. Therefore, the GI number selected was the one that is available in U00096.ptt file (with the *map_back.pl* script). The final datasets were two lists of GI numbers, the disordered sequences in *E. coli* (41 genes) and the ordered (4099 genes). Here we make the simplified assumption that proteins with no characterized disordered regions are ordered. A pipeline implementing the aforementioned procedure is schematically depicted in Figure 34.

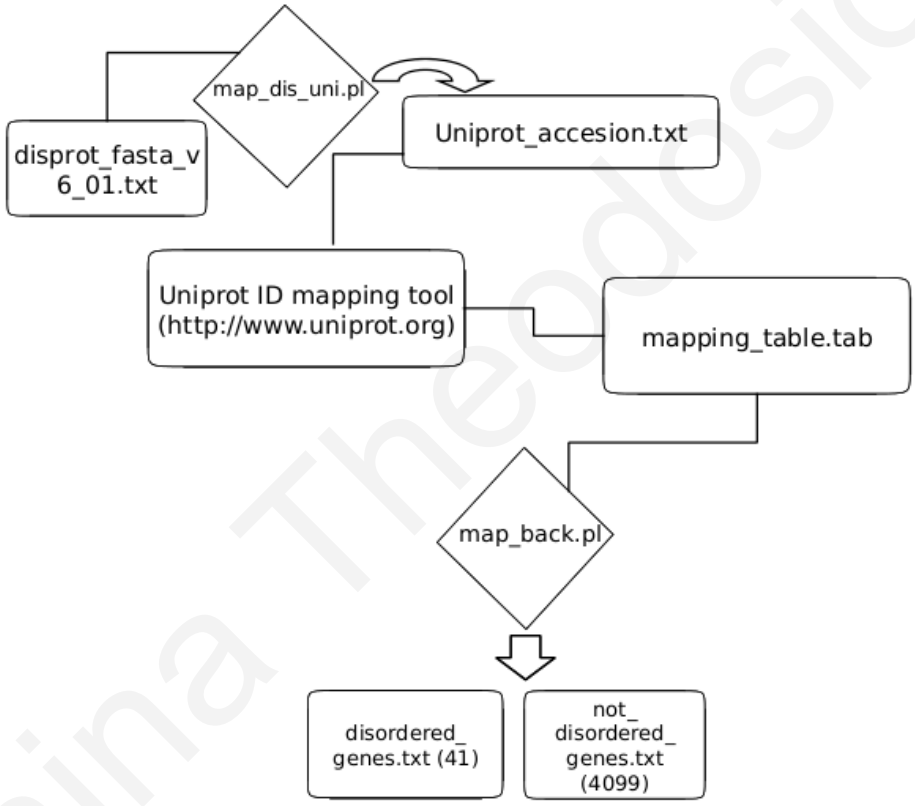

Figure 34: Pipeline to extract disordered IDs.
Extract genes with experimentally determined disordered regions (disordered genes) and IDs corresponding to genes with no experimentally determined regions (ordered genes) from the *E. coli* K12 MG1655 dataset.

2 x 2 contingency tables were created to display the number of disordered genes that have an RCC, the number of disordered genes that do not have an RCC, the number of ordered genes that have an RCC and the number of ordered genes that do not have an RCC. The

datasets disordered.txt (41 genes), not_disordered_genes.txt (4099 genes) along with datasets least_one.txt and *no_RCC.txt* were used to construct the tables with script i*nput_for_Fisher-test.pl*. The last two text files are datasets produced by *statistical_RCC.pl* (described in Chapter 3 for the measurement of statistical properties) and have the GI number IDs for sequences with at least one RCC (*least_one.txt*) and sequences with no RCCs (*no_RCCs.txt*). These two datasets were produced separately for each LaTcOm method (%Minmax (cu), %MinMax (Z), RiboTempo and MSS). Ultimately, four 2 x 2 contingency tables were created, one for each method.

A validation test on the disordered sequences was the creation of contingency tables for 41 randomly selected ordered genes. In more detail, ten datasets were created, each composed of 41 randomly picked genes from the not_disordered_genes.txt set with the script *get_random_IDs.pl*. The GI number that was picked randomly from the ordered set, had to be included in the *least_one.txt* or *non_RCC.txt* file data. If a sequence was excluded from LaTcOm analyses for not satisfying the minimum length criterion, another GI number was chosen. Contingency tables were then created with the i*nput_for_Fisher-test.pl* script. From the ten contingency tables created, a contingency table with mean values was created with *mean_estimation.pl* script.

All contingency tables described above were given as input to *fisher.test*() in the R statistical environment (R Development Core Team, 2008), with default parameters and p-values were estimated.

Spatial correlation of RCCs and disordered regions

The positions of RCCs were placed on sequence strings with the *mapped_on_seq.pl* script. The script reads LaTcOm results and creates a FASTA formatted file in which the sequences are created with the letters “X” and “M”. The X is placed at positions of RCCs whereas the M at non RCC positions of the sequences. The information regarding the exact position of disordered regions in sequences was extracted from disprot_fasta_v6_01.txt file (Appendix 2 - Figure 49) from header information (positions marked with “#” symbol) with *map_disordered.pl* and a transformed FASTA file was created (mapped_disordered.txt), similar to the one with RCCs described previously. Last, the overlapping codons of disordered regions and RCCs were calculated with *final_correlations.pl.*

α-Helical Transmembrane proteins

Transmembrane protein data was extracted from Table S1 in Supplementary data of (Daley et al., 2005). The table is a summary of the experimentally determined topologies of 738 membrane proteins of *E. coli*. These sequences encode proteins longer than 100 residues and with at least two predicted transmembrane helices. The gene names of this table (column 1) are from the Colibri database (Medigue et al., 1993) and some entries are clones from (White, 2004). After filtering those gene names that do not exist in the U00096.ptt dataset (there are 150 gene names missing from the U00096.ptt file possibly because they refer to different strains) - the final dataset included 588 inner membrane proteins of *E. coli* K12 (with the script *get_IDs.pl*) (membrane_IDs_corrected.txt). In order to create 2x2 contingency tables (as described in disorder section above), the non transmembrane dataset was extracted (3510 genes) (non-membrane_IDS.txt) and tables were created with *input_for_Fisher-test.pl* using again least_one.txt and no_RCCs.txt which hold information regarding the detection of RCCs. Additionally, 10 randomly selected datasets of 588 non transmembrane sequences were created (with *get_random_ID.pl*) from which a mean contingency table (with *mean_estimation.pl*) was computed as described with disordered genes using the same scripts. As with disorder, all contingency tables were given as input to *fisher.test* function in the R statistical environment (R Development Core Team, 2008), with default parameters and p-values were estimated.

Taking the data from (Daley et al., 2005), regarding the number of transmembrane helices predicted, we divided the data into sequences with less than 6 helices (191), sequences with more or equal to 6 helices (305) and those not defined (92) with get_tm_helices.pl script. Moreover, we applied the Fisher Exact test regarding the existence or not of RCCs.

Sequences with signal peptides

The standalone version of signalP 4.1 ((Petersen et al., 2011); http://www.cbs.dtu.dk/services/SignalP/) was used in order to predict signal peptides in the *E. coli* K12 proteome. The NC_000913.faa file, which has all *E. coli* protein coding sequences, was used as input to the signalP tool. Example of output is available in

Appendix 2 - Figure 50). The *split_IDs.pl* script is used to divide data into two files the sequences with signal peptide (426) and the sequence with no signal peptide (3676).

As described in disordered and transmembrane sections, 10 randomly selected datasets of 426 gene IDS encoding proteins without a predicted signal peptide. Contingency tables were created with the tools previously described and the Fisher test in R was performed.

Transmembrane and sequences with signal peptides

With s*cript get_seq_and_tm.pl* we combined the transmembrane and signal peptide sets. In more detail, the first file created (*TM_and_secreted_IDS.txt)* included proteins that are either transmembrane or have a signal peptide (998 proteins). The second file (*nonTM_and_nonSec_IDS.txt*), included proteins that are neither transmembrane nor secreted (3100 proteins). We did not perform random sets for this group (since the results are not affected by unequal data sets in Fisher test), but used directly the sets to create the contingency tables for Fisher test.

Peripheral inner membrane proteome

Papanastasiou and colleagues defined the peripheral membrane proteins (i.e. non integral to the membrane) that face the cytoplasmic layer of the *E. coli* plasma membrane (Papanastasiou et al., 2013). Table S1 (Table 1A) from the supplementary material of this work gives the annotation for 278 peripheral inner membrane proteins for *E. coli* K12. All UniProt accession numbers were given to the UniProt mapping tool (www.uniprot.org/uploadlists) in order to retrieve GIs. Nevertheless, only 247 UniProt accession numbers were successfully mapped to a GI, due to the fact that some UniProt IDs were repeated in that list because it was also referring to a different entry names for *E. coli* BL21. A further reduction was made for three genes that did not have a match in *E.coli* K12. EcoGene (www.ecogene.org) classifies these three as pseudogenes (Appendix 1 - Table 43). Ultimately 244 genes were retrieved. Along with LaTcOm results, we used this set to create the contingency tables for Fisher test as previously described.

Single- versus multi-domain proteins

In order to get sequences that have a known 3D structure, we searched UniProt for all *E. coli* K12 proteins that are cross referenced in PDB (http://www.rcsb.org) and identified all the proteins (1310) (date of retrieval - 01/04/14). The corresponding sequences were downloaded in FASTA format. We also mapped UniProt IDs to the respective GenBank GI numbers with the UniProt mapping tool. Next, we downloaded the complete mapping of PDB chains to UniProt entries from the PDBSWS server ((Martin, 2005); www.bioinf.org.uk/pdbsws/).

The annotation regarding domain boundaries was taken from the dir.cla.scope.2.03-stable.txt within the SCOP database ((Fox et al., 2013);http://scop.berkeley.edu-SCOPe *extended* 2.03 release). This is a tab-delimited file with information on the domain coordinates and an example can be seen in Appendix 2 - Figure 51. With the script *get_s_m_domains.pl* that was developed, the *E. coli* single and multi-domain proteins were separately retrieved, along with the annotation on domain boundaries for multidomain proteins and mapping of PDB chains to GIs. We identified in total 213 multidomain and 714 single domain proteins. The pipeline described above in order to retrieve these datasets is shown in Figure 35. Contingency tables were constructed with *input_for_Fisher-test.pl* and Fisher exact test was performed as described above for single and multidomains along with RCCs.

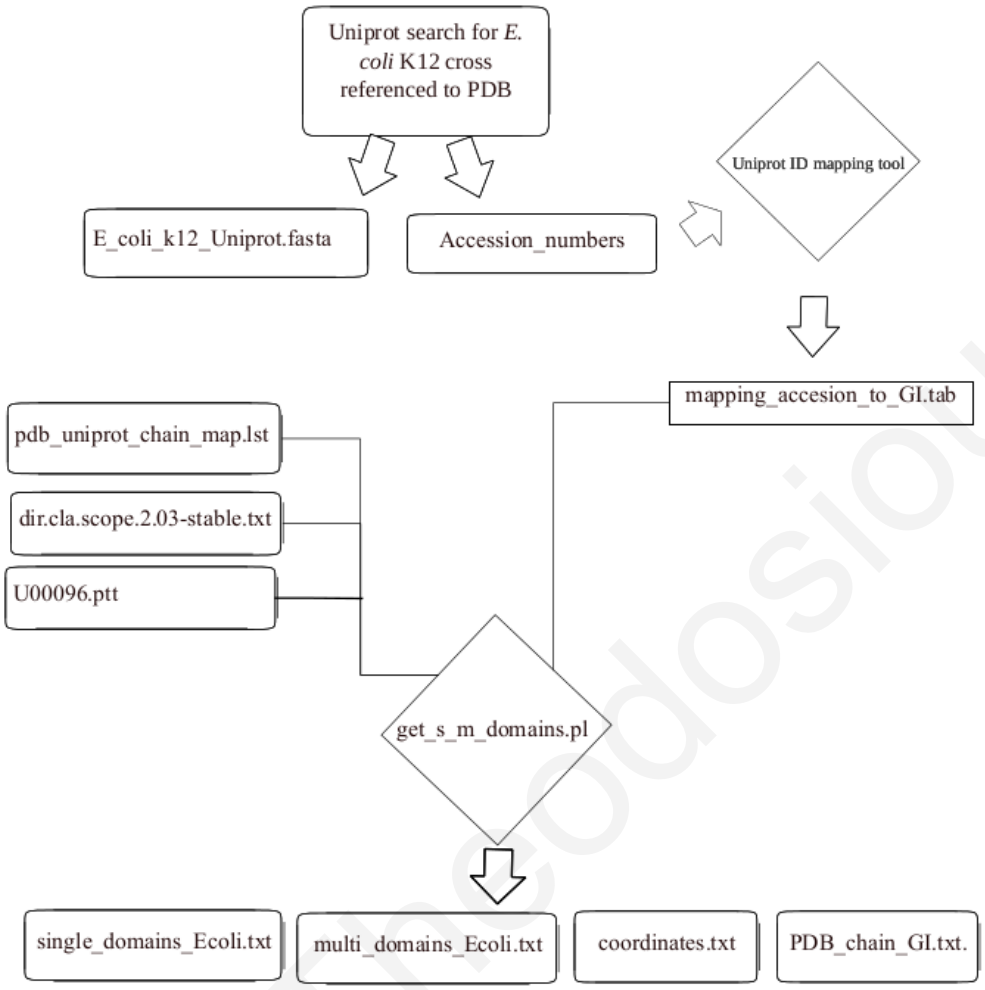

Figure 35: Pipeline to retrieve multidomain proteins of *E. coli* K12.

Mapping domain coordinates to RCCs

The following pipeline describes the procedure in order to get the exact matching cDNA for the mutlidomain chains extracted from PDB (http://www.rcsb.org). The protein sequences of chains were downloaded from the PDB database (www.rcsb.org) in FASTA format. All chains for a single PDB entry are extracted with this procedure therefore, the exact chain in search was selected with get_correct_fasta.pl script. The chains were then given as query to standalone BLAST (Altschul et al., 1990) (blastall version 2.2.25 - blastp). The database in search used was the protein sequences in *E. coli* K12 MG1655. The database was formatted with formatdb and blastp was run with no filtering option on.

The output was then parsed with script *parse_blast.pl*. in order to get the top hits with >85% identity matches. The last file was used in *get_dna.pl* in order to extract the whole length cDNA and the corresponding whole length proteins. Next, the *1_parse_blast_3_updated.pl* script read the blast output and created the exact matching cDNA file that corresponds to the PBD chain. This was then given to standlone LaTcOm and RCCs were detected with the parameters and methods described in Table 3. Two multidomain chains were rejected (with GI 3868712 and 3868719 - see Appendix 1- Table 37), since they contain in-frame 'stop codons', which are actually translated in selenocysteine. As previously discussed, this is not handled by LaTcOm yet. Ultimately, RCCs were detected for 211 multidomain chains.

In order to map RCC positions with domain boundaries, the *map_RCC_domains.pl* script was developed. The cDNA described before was identified based on the recent version of NC000913.ptt, therefore, LaTcOm prediction were done based on the GI numbers of this file. However, the domain coordinates were predicted based on U00096.ptt. In order to get correct mapping of the genes we compared (map_GI_numbers.pl) the two files based on the synonymous code in ptt files. The genes that did not have a matching in the two files are shown in Appendix 2 - and . Nevertheless, these genes do not encode multidomain proteins.

Ultimately, we redefined domain boundaries as the middle of the distance between the end and the start of two successive domains (+/- 10) and last, calculated the distance of RCCs (the middle of the coordinates) from the closest domain boundary (RCCs_dis_from_boundaries.txt) for each LaTcOm result. Overlapping RCC and boundaries are assigned with 0 distance, RCCs downstream the boundary are assigned with negative distance whereas RCCs upstream the domain boundary are assigned positive distance.

The distance distributions were analysed with the R statistical environment (R Development Core Team, 2008) and the *summary ()* and *sd()* functions were applied for estimating the descriptive statistics of the data.

Outer membrane β-barrel sequences

243 β-barrel outer membrane proteins of *E. coli* K12 MG1655 were downloaded from the TMBB-DB database ((Freeman and Wimley, 2012); http://beta-barrel.tulane.edu/). Two

genes in the above dataset (145698324 and 16129928) are not referenced in NC_000913.ptt and were excluded. With the script *get_bbarrels_and_not_bbarrels.pl* 241 β-barrel proteins were identified and 3899 not β-barrels. Contingency tables were constructed with *input_for_Fisher-test.pl* and Fisher exact test was performed as described before, for β-barrels membrane proteins and not β-barrels along with RCCs.

Dataset summary

Summarising, for all the contingency tables that were created for each functional or structural group as described above, we employed the Fisher's exact test (function fisher.test ()) as implemented in the R statistical environment (R Development Core Team, 2008). This was done in order to determine if there are non-random associations between the two categorical variables under study. The null hypothesis for all cases in Fisher exact test is that there is no association between a functional or structural category and the existence of RCCs. The following groups for associations were estimated between RCC/non-/rcc vs:

(i) discorded/non-disordered
(ii) membrane/non-membrane
(iii) secreted/nonsecreted
(iv) secreted or membrane/non secreted and non membrane
(v) cytoplasmic inner membrane peripheral proteins
(vi) single/multidomain
(vii) β-barrels/non β barrel

4.2.2 Gene ontology (GO) enrichment analysis

As mentioned before, the script *statistical_RCC.pl* described in Chapter 3, generated several data files for the GO analysis and for each detection method of LaTcOm. The files contain GIs of *E. coli* K12 proteins and were divided based on the following characteristics related to the detection of RCCs:

i. No RCCs
ii. At least one RCC

iii. Only one RCC

iv. More than four RCCs

v. with RCCs at the 5' terminus

vi. with RCCs at the 3' terminus

vii. with RCCs on both 5' and 3' terminals

viii. Based on cluster length distributions

An in-house Perl script (*E-term_finder.pl*) was used and adjusted to our results (*E-term_finder_pdb.pl)* that makes use of several modules for GO analysis in order to determine whether these sets of sequences share any over-represented GO-term. The *E-term_finder.pl* was developed by Vasilis J. Promponas and Eleni Mytilineou and uses the TermFinder module among others ((Boyle et al., 2004); module version 0.86: http://search.cpan.org/dist/GOTermFinder/lib/GO/TermFinder.pm). This module is a group of object-oriented Perl modules that can be used to determine the significance of a GO annotation to a list of genes. Bonferroni correction was used to correct for multiple testing. For the analysis the following files were downloaded:

i. gene_association.ecocyc (www.geneontology.org/GO.downloads.annotation.shtml-27.06.14) which is a filtered annotation file and

ii. gene_ontology.obo (http://www.geneontology.org/ontology/gene_ontology.obo-27.06.14 format-version: 1.0 / version: releases/2014-05-27) which contains the ontology structure.

GIs were converted to gene names (*convert_gi_to_geneneame.pl*) in order to be associated with the gene_association.ecocyc file and a threshold of $p<0.01$ was set for extracting the statistically significant results. The script *read_GO_output.pl* was developed to produce a user-friendly output of the results.

4.2.3 *E. coli* integral inner-membrane proteins with experimentally determined atomic structures

In this section, we focused our research in a more detailed structural analysis of αHTMP

sequences with experimentally determined structures and their correlation with RCCs. TM sequences were reported to frequently contain RCCs (Zhang et al., 2009) and their possible impact of RCC location in TM protein folding and topology has not been investigated so far. In Chapter 3 we identified that RCC existence is enriched within TM proteins of *E. coli*, therefore it was of great importance to investigate in more depth whether there exists an important pattern connecting RCCs and TM helical topology. We designed and implemented a number of computational tools to facilitate mapping and correlating RCCs to TM protein topology (including cytoplasmic - periplasmic regions) and TM helical packing patterns (for polytopic subunits).

α-helical chain retrieval

Initially, a simple text search was performed in the UniProt/SwissProt database (http://www.uniprot.org/; (Magrane and Consortium, 2011)), which currently contains the most thoroughly annotated protein sequence dataset available. We searched for *E. coli* proteins (OS : *Escherichia coli* (strain K12)) which are characterized in the sequence annotation as TM (FT : transmembrane) and are cross referenced to the PDB database (121 entries were identified with this search). Next, we mapped these UniProt entries to PDB identifiers in order to prepare a structural dataset (553 structures matched). It is important to know that there may be several structures available that represent the same protein e.g. several mutated forms of the same polypeptide, or the same subunit in different complexes/stoichiometries may exist in structural database.

Retrieval of non-redundant cDNA chains and RCCs

In order to create a non-redundant dataset, the PISCES standalone program was used (http://dunbrack.fccc.edu/Guoli/pisces_download.php#BLASTDB; Wang and Dunbrack, 2003). The program removes redundancy according to a threshold and ultimately leaves only one representative from each protein chain. From the PISCES package we used the *Cull_for_UserSEQ.pl* script for filtering with a 40% sequence identity threshold. Ultimately, 46 TM chains were identified (The list with the pdb chain IDs is provided in Appendix 2 -. Next, we identified the cDNA for each TM chain. The procedure is not straightforward, since the protein chains may not be the full length protein sequences. The

description of the procedure on how we extracted the matching cDNA sequence has already been described in a previous section (in Data and Methods of Chapter 3 (section: *Mapping multidomain coordinates with RCCs*)). The pdb chain '3udca' was rejected from further analysis because it belongs to a different strain (*E. coli* strain C43) and it was not identified within the *E. coli* K12 protein sequences. Therefore, RCCs were detected with the LaTcOm standalone tool in 45 PDB chains with all methods with the parameters described in Table 3.

Mapping TM helices

A complete list of the TM proteins is difficult to extract from PDB because the annotation is not reliable (Tusnády et al., 2004). Moreover, TM annotations in UniProt are sometimes inferred from homology relationships. For these reasons, we decided it is more appropriate to use the PDBTM database (http://pdbtm.enzim.hu/; (Tusnády et al., 2004)), which is a comprehensive database collection of TM proteins extracted from PDB. The database relies on the structure-based computational identification of TM regions within protein structures. Each chain record from the PDBTM database contains one or more topological region records, which locates the chain segment in the space relative to the membrane. In αHTMPs we are interested in "side1"/"side2", TM helices and dipping loops (re-entrant regions). Suitable transformations, with regard to sequence position, were performed, to bring all annotations to a common "coordinate" system. This step is necessary to overcome several peculiarities observed in PDB data. For each chain we had mapping information stored for further analysis. Additional work was performed to analyse geometrical features of TM helices such as their pairwise distances. For this, PDB structures (39 available; 1y8s structure is a theoretical model and, therefore, was not used for further analysis) were downloaded from PDB (http://www.rcsb.org).

Assigning enriched topological information

Unfortunately, no annotation exists for the cytoplasmic or periplasmic regions within PDB or PDBTM. Knowing that the environment surrounding TM proteins is highly asymmetric, it is possible that topological information may be of importance in our

analysis. We firstly used the information, from reliable experimentally derived data (ExToPoDB.flat) deposited in the ExTopoDB database (http://bioinformatics.biol.uoa.gr/ExTopoDB/; (Tsaousis et al., 2010)), which is the most comprehensive, manually curated and recently updated database (G. Tsaousis, personal communication). An example of the format of this file can be seen in Appendix 2 - Figure 52. Topological information exists for only 30 PDB chains (from total 44 analyzed[7]) in ExTopoDB. We also explored information based on high throughput experimental mapping of C-terminal regions made by (Daley et al., 2005) in order to map the topology for sequences that we did not find evidence in ExToPoDB. The authors in this paper have established the periplasmic or cytoplasmic locations of C-termini for 601 inner membrane proteins with appropriate reporter fusion constructs. We used this information to manually crosscheck the topological placement for all non-TM regions predicted in ExTopoDB (The topology between the two sets was the same for 27 from 30 chains - the remaining three chains did not have a topological assignment in (Daley et al., 2005)). From the same data we additionally identified the C-terminal topology for 9 more chains. Nevertheless, for the remaining (14) chains that their topology was not found in ExTopoDB we followed a semi-automatic procedure and performed consensus topology predictions based on TOPCONS single (Bernsel et al., 2009). The authors of this web service suggest that their method performs better than any of the other topology prediction methods tested. We identified topological information for 13 PDB chains[8]. These chains were again crosschecked with (Daley et al., 2005) data and topology agreed. Finally, we gathered information of cytoplasmic and periplasmic locations for 43 chains.

In this activity our first attempt was to correlate the topological features (including cytoplasmic - periplasmic regions) and structural features of each chain with RCCs. Therefore, we designed and implemented software (parseXMLpdbtm_helices_or_loops.pl) that correlated the RCC coordinates, taking as input the files of TM topology, cytoplasmic-periplasmic information and structural information (TM interactions from 3D structures) for each sequence. A distinct module was designed to include correlations with pairwise distances for interaction between successive in sequence TM helices (parseXMLpdbtm_tm_interactions.pl). We used a

[7] 1y8s is a theoretical model and was not found in ExToPoDB and pdbsws; 3udca chain originates from a different *E. coli* strain

[8] 4iffa does not have TM helices, therefore it was rejected for topological predictions with TOPCONS

distance threshold of 5.5Å to define interacting pairs of adjacent in sequence TMHs, with the additional condition that more than 5 residue contacts should be present for any TMH pair to be considered as interacting.

All these aimed to:

- Estimate the correlation of the position of independent RCCs with TM/nonTM regions (irrespective of their cytoplasmic or periplasmic topology).
- With the topology data for noTM regions available, we designed and implemented a methodology in order to correlate the relevant information regarding the topology of transmembrane segments, C-terminus, N-terminus and the detected RCCs. We aligned the topology of each protein against its codon rareness/non-rareness cluster and
- We also expanded the already implemented script for the estimation of interactions between transmembrane helices in order to correlate interaction between transmembrane segments and position of RCCs in loop regions.

For validating the presence of rare codon clusters in helices or loops and in cytoplasmic or periplasmic loops we used the measures of Positive Predictive Value (PPV), Negative Predictive Value (NPV), Accuracy, Sensitivity and Specificity (Guggenmoos-Holzmann and van Houwelingen, 2000):

Sensitivity: Sensitivity measures the fraction of the actual positives which are correctly predicted:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity: Specificity denotes the fraction of the actual negatives which are correctly predicted:

$$Specificity = \frac{TN}{TN + FP}$$

PPV: The positive predictive value (PPV) is the fraction of the predicted positives which are correct:

$$PPV = \frac{TP}{TP + FP}$$

.

NPV: The negative predictive value (NPV) stands for the fraction of the negative predictions which are correct:

$$NPV = \frac{TN}{TN + FN}$$

For the purposes of the former analysis, true positives are regions with identified RCCs and general loops. False positives are regions with RCCs and TM helices. False negatives are regions with no RCCs and loops whereas true negatives are regions with no RCCs and TM helices. For the latter analysis regarding the topology of loops, true positives are periplasmic loops with RCCs, false positives are regions of cytoplasmic loops and RCCs. False negatives are regions of periplasmic loops and no clusters whereas true negatives are cytoplasmic loops with no RCCs.

For the third analysis, in which we took into account connecting loops and interacting helices, true positives are connecting loops with non-interacting helices and RCCs. False positives are connecting loops with interacting helices and RCCs. False negatives are connecting loops with non-interacting helices and no RCCs whereas true negatives are connecting loops with helices that interact and no RCCs. An additional parameter taken into account was the strength of the length of the loop connecting the successive helices.

4.3 Results and discussion

4.3.1 Correlation of RCCs with functional and structural features

An increased amount of evidence demonstrated recently that the existence of RCCs has a functional impact in proteins. We have previously discussed the presence of rare codons in secretory sequences that was shown to exist in *E. coli* and *Salmonella typhimurium* (Burns and Beacham, 1985; Power et al., 2004; Zalucki and Jennings, 2007). These slowly translated regions were also shown to be preferred in β-strands and coils than in α-helices (Thanaraj and Argos, 1996b). Additionally, positive correlation between hydrophobic stretches and RCCs was shown in membrane sequences suggesting a functional role in membrane targeting or insertion (Dessen and Képès, 2000). Moreover, the correlation between RCCs and domain boundaries was observed in many studies (Komar and Jaenicke, 1995; Krasheninnikov et al., 1991; Purvis et al., 1987; Thanaraj and Argos, 1996b) although debated in more recent publications (Brunak and Engelbrecht, 1996; Saunders and Deane, 2010). Recently, in a single case study Zhang et al., 2009 showed experimentally that slow codons found in domain boundaries are actually necessary for proper protein folding.

In this work, we applied a bioinformatics analysis and we investigated the association of RCCs detected by LaTcOm with functional and structural features of *E. coli* sequences.

E. coli proteins with structural disorder regions and RCCs

Firstly, we searched for correlations between intrinsically disordered proteins or regions along with the existence of RCCs. However, there are few annotated proteins as disordered in *E. coli* and our data set is imbalanced. In general, disordered proteins are more prevalent in eukaryotic than in bacterial proteomes (Pancsa and Tompa, 2012). In another study of the *E. coli* K12 proteome, 5% of the proteins were shown to be mostly disordered, whereas 20% had at least one disordered segment longer than 30 residues (Oldfield et al., 2005). Nevertheless, the study relied on a consensus of predictions based on charge−hydropathy distribution and disorder prediction score distribution. In our analysis, we used only experimentally verified disordered segments or regions, thus we

gathered only 41 *E. coli* sequences. We have not identified any significant correlation between disordered proteins and the existence of RCCs in these sequences. Table 22 gives the p-values (notably all > than 0.01) of the Fischer Exact Test for disordered and RCCs and Table 47 in Appendix 2 the p-values using random sequence datasets of non disordered proteins.

Table 22: P-values from Fisher Exact Test with contingency table for disordered genes and existence of RCCs.

Method	**p-value**
%MinMax (cu)	0.8725
RiboTempo	0.04
MSS	1

The number of overlapping disorder codons and RCCs can be seen in Table 23. These results additionally demonstrate that no correlations exists in *E. coli* K12 between these two properties as far as this analysis is concerned. However, is important to clarify that we had few sequences to analyse (only 41 experimentally verified disordered sequences). Further analyses based on predictions or analysis in other organisms with higher percentages of disorder regions may be more appropriate to reveal if there is a real connection of disordered property with RCCs.

Table 23: Number of codons in RCC mapping to disordered regions
First row demonstrates the number of disordered codons found in RCCs and the second row the total codons in RCCs detected with each method.

Codons	**%MinMax (cu)**	**RiboTempo**	**MSS**
Disordered mapped RCCs	409	242	200
Total_RCC:	1307	895	735

RCCs are correlated with TM and secreted sequences

Previous analyses demonstrated a relationship between TM proteins and RCCs (Zhang et al., 2009). We confirm these findings by showing that the structural membrane property correlates well with the existence of RCCs with all methods with statistical significance ($p < 0.01$) (Table 24). Exclusion of RCCs that lay at the 5' sites (<100) did not alter the results (Fisher exact test p=0.001) with MSS.

Table 24: P-values of Fisher Exact Test for membrane/non membrane genes versus RCCs/non RCCs. Random data are described in Data and Methods.

Method	**p-value**	**p-value (from random data)**
%MinMax (cu)	0.001	0.011
RiboTempo	<2.2e-016	1.14e-010
MSS	0.003	0.012

In order to check if the correlation differs based on the number of TM helices we performed a correlation analysis with membrane sequences with less than 6 helices and membrane sequences with 6 and more helices. Fisher Extact test p-values can be seen in Table 25. The RCCs seem to correlate better with TM sequences with greater number of helices.

Table 25: P-values of Fisher Exact Test for membrane with less than 6 helices /membrane with more than 6 versus RCCs/non RCCs.

Method	**p-value**
%MinMax (cu)	0.003
RiboTempo	2.995e-05
MSS	0.01

Next, we searched for correlation of secreted sequences with RCCs (Table 26). The p-values are significant, demonstrating a correlation of secreted proteins and RCCs. Exclusion of RCCs that lay at the 5' sites (<100) altered the results with higher p-value but the relationship still remains significant (Fisher exact test p=0.001 with MSS).

The same analysis on the combined TM/secreted dataset resulted in smaller p-values (Table 27).

Table 26: Fisher Exact Test for secreted/non secreted genes versus RCCs/non RCCs.

Method	p-value	p-value (random data)
%MinMax (cu)	0.004	0.026
RiboTempo	0.011	0.032
MSS	1.985e-07	0.00004

Table 27: P-values for Fisher Exact Test for secreted or TM/non secreted and not TM genes versus RCCs/non RCCs.

Method	P-value
%MinMax (cu)	4.482e-06
RiboTempo	7.159e-07
MSS	1.15e-09

The peripheral inner membrane proteome did not show any significant correlation to existence of RCCs for $p < 0.01$. (Table 28).

Table 28: P-values for Fisher Exact Test for peripheral (cytoplasmic) inner membrane proteome and the rest of the proteins of *E. coli* K 12 versus RCCs/non RCCs.

Method	P-value
%MinMax (cu)	0.46
RiboTempo	0.04
MSS	0.03

Multidomain proteins showed interestingly a high significant association with RCCs with two out of the three methods. (Table 29).

Table 29: P-values for Fisher Exact Test for multidomain/single domain versus RCCs/non RCCs.

Method	P-value
%MinMax (cu)	4.163e-05
RiboTempo	4.785e-07
MSS	0.199

Previous studies have demonstrated that there is a substantial correlation between domain boundaries and positions of rare codon clusters on mRNA (Komar and Jaenicke, 1995; Krasheninnikov et al., 1988; Purvis et al., 1987; Thanaraj and Argos, 1996a; Zhang et al., 2009). Nevertheless, (Saunders and Deane, 2010) as well as (Brunak and Engelbrecht, 1996) did not show such correlation in their findings, while correlating rare codons and domain boundaries on a larger scale. Herein we confirm, that at least for multidomain protein of *E. coli*, RCCs identified with LaTcOm methods are correlated well with domain boundaries. summarizes the distributions of distances of RCCs from domains boundaries of mutlidomain proteins with all methods and demonstrates this signal. The mean distances with all methods are located upstream of domain boundaries.

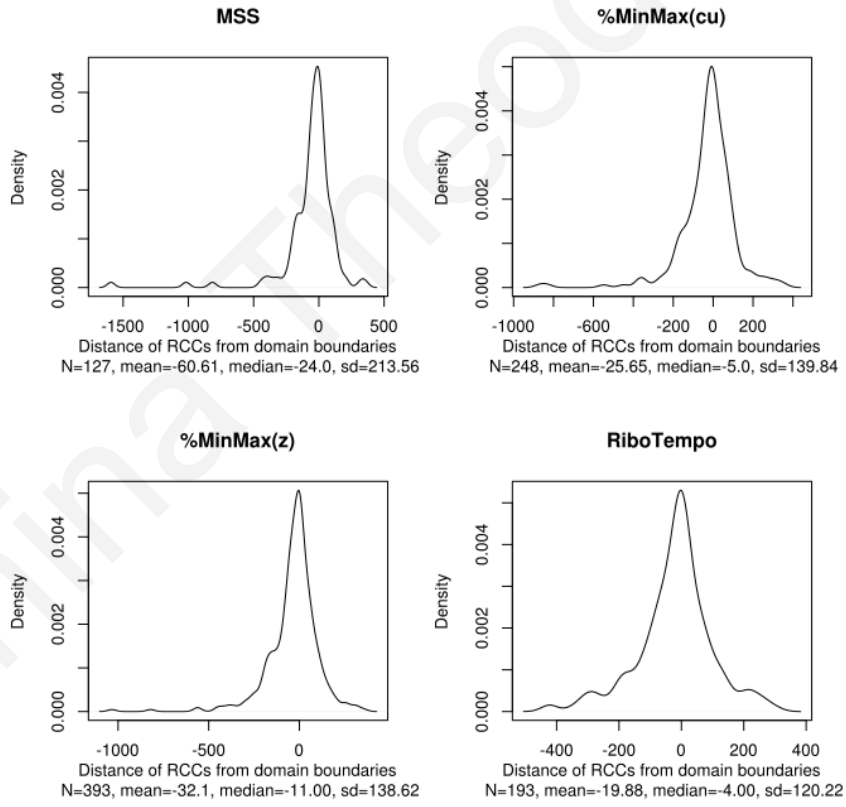

Figure 36: Distance distribution of RCCs detected with each method from domain boundaries.

RCCs correlated with β-barrel outer membrane proteins

(Thanaraj and Argos, 1996b) observed that rare codons preferentially code for β-strands. Additionally a more recent study on global Pfam domains showed that the three top topologies with rare codons included β-strands and β-barrel outer membrane proteins (Chartier et al., 2012). However, rare codons were suggested to be necessary for α-helices to fold (Zalucki and Jennings, 2007).

Table 30: P-values for Fisher Exact Test for β-barrel/non β-barrels versus RCCs/non RCCs.

Method	**P-value**
%MinMax (cu)	3.629e-13
RiboTempo	2.512e-08
MSS	<2.2e-16

β-barrel proteins are found exclusively on the outer membrane of gram-negative bacteria and in the outer membranes of mitochondria and chloroplasts (Cavalier-Smith and Cavalier-Smith, 2000; Elofsson and Heijne, 2007; Gray et al., 1999). In this study we demonstrate that there is a statistically significant correlation of β-barrel outer membrane proteins of *E. coli* and RCCs as highlighted in Table 30. Exclusion of 5' terminal RCCs (those below position 100) did not alter the results with MSS (Fisher test p-value <2.2e-16).

Gene ontology (GO) enrichment analysis

The GO analysis revealed several interesting signals. In Table 31 we provide the results with MSS detections and the rest of the analysis with other methods is provided in Appendix 2 -Table 48 and Table 49. Observing the results from MSS, many of the sequences with no RCCs are cytoplasmic, involved in RNA or protein binding and correlated with ribosome activity. Sequences that have RCCs at the 5' and 3' termini are shown to be involved in signaling and related with the membrane. We further investigated the gene ontology enrichment of sequences that are found to have RCCs or not in all methods (1319 sequence with at least one RCC as shown in Venn diagram of Figure 22 and 900 sequences with no RCC shown in Figure 23). The results (not shown) come into agreement with the above observations.

Table 31: Gene ontology enrichment analysis results filtered for p<0.01.
Bonferroni correction was used to correct multiple testing p-values. Column "Data" describes the dataset that was used for the analysis. "P" is for process, "C" is for cellular component and "F" is molecular function. For example "P_end_80" dataset represents genes that have an RCC at the last 80 codons and have a process related described in column "Term".

Data	Term	P-value	Num_annotations
P_end_80	signal transduction	0.000372831	32 of 466 in the list, versus 105 of 4141 in the genome
P_end_80	Signaling	0.000473251	32 of 466 in the list, versus 106 of 4141 in the genome
P_end_80	single organism signaling	0.000473251	32 of 466 in the list, versus 106 of 4141 in the genome
C_end_80	external encapsulating structure	0.000378754	51 of 466 in the list, versus 225 of 4141 in the genome
C_end_80	cell envelope	0.000433483	50 of 466 in the list, versus 220 of 4141 in the genome
C_end_80	Envelope	0.000496612	50 of 466 in the list, versus 221 of 4141 in the genome
C_end_80	external encapsulating structure part	0.000646482	47 of 466 in the list, versus 205 of 4141 in the genome
F_no_RCC	structural constituent of ribosome	3.14171857099062e-13	54 of 1755 in the list, versus 56 of 4141 in the genome
F_no_RCC	structural molecule activity	3.52980228377893e-09	65 of 1755 in the list, versus 77 of 4141 in the genome
F_no_RCC	rRNA binding	5.72042321061098e-07	42 of 1755 in the list, versus 47 of 4141 in the genome
F_no_RCC	RNA binding	8.99308375857488e-05	107 of 1755 in the list, versus 160 of 4141 in the genome
F_no_RCC	protein binding	0.000138146	352 of 1755 in the list, versus 628 of 4141 in the genome
C_no_RCC	Intracellular	1.09981506652571e-21	615 of 1755 in the list, versus 1024 of 4141 in the genome
C_no_RCC	intracellular part	2.38611289633266e-20	593 of 1755 in the list, versus 989 of 4141 in the genome
C_no_RCC	Cytoplasm	5.44952143946664e-18	542 of 1755 in the list, versus 904 of 4141 in the genome
C_no_RCC	cytoplasmic part	4.18036104582858e-17	202 of 1755 in the list, versus 280 of 4141 in the genome
C_no_RCC	intracellular non-membrane-bounded organelle	3.22387559259071e-15	82 of 1755 in the list, versus 93 of 4141 in the genome
C_no_RCC	intracellular organelle	1.5478828204672e-14	84 of 1755 in the list, versus 97 of 4141 in the genome

C_no_RCC	Cytosol	6.92302458229666e-14	175 of 1755 in the list, versus 245 of 4141 in the genome
C_no_RCC	cytosolic part	2.86857350610084e-13	55 of 1755 in the list, versus 58 of 4141 in the genome
C_no_RCC	ribonucleoprotein complex	5.61805493846237e-13	57 of 1755 in the list, versus 61 of 4141 in the genome
C_no_RCC	non-membrane-bounded organelle	2.2621269889335e-12	97 of 1755 in the list, versus 121 of 4141 in the genome
C_no_RCC	Ribosome	2.32186943382099e-12	55 of 1755 in the list, versus 59 of 4141 in the genome
C_no_RCC	Organelle	5.21920264038167e-12	99 of 1755 in the list, versus 125 of 4141 in the genome
C_no_RCC	cytosolic ribosome	1.0592342213317e-11	50 of 1755 in the list, versus 53 of 4141 in the genome
C_no_RCC	ribosomal subunit	1.0592342213317e-11	50 of 1755 in the list, versus 53 of 4141 in the genome
C_no_RCC	intracellular organelle part	6.97908715494093e-11	60 of 1755 in the list, versus 68 of 4141 in the genome
C_no_RCC	organelle part	1.99244736919464e-08	72 of 1755 in the list, versus 91 of 4141 in the genome
C_no_RCC	large ribosomal subunit	2.41356885562577e-07	30 of 1755 in the list, versus 31 of 4141 in the genome
C_no_RCC	cytosolic large ribosomal subunit	2.41356885562577e-07	30 of 1755 in the list, versus 31 of 4141 in the genome
C_no_RCC	macromolecular complex	2.76058207677343e-05	189 of 1755 in the list, versus 312 of 4141 in the genome
F_end_20	integrase activity	0.000254665	4 of 98 in the list, versus 5 of 4141 in the genome
F_end_100	sequence-specific DNA binding transcription factor activity	0.000166038	60 of 598 in the list, versus 204 of 4141 in the genome
F_end_100	signal transducer activity	0.000288814	32 of 598 in the list, versus 85 of 4141 in the genome
F_end_100	nucleic acid binding transcription factor activity	0.000414579	60 of 598 in the list, versus 209 of 4141 in the genome
F_end_100	molecular transducer activity	0.000707735	32 of 598 in the list, versus 88 of 4141 in the genome
P_end_100	signal transduction	0.000188926	38 of 598 in the list, versus 105 of 4141 in the genome
P_end_100	Signaling	0.000250974	38 of 598 in the list, versus 106 of 4141 in the genome
P_end_100	single organism signaling	0.000250974	38 of 598 in the list, versus 106 of 4141 in the genome

F_end_80	signal transducer activity	0.000535866	27 of 466 in the list, versus 85 of 4141 in the genome
P_no_RCC	Translation	3.26477843283877e-10	89 of 1755 in the list, versus 111 of 4141 in the genome
P_no_RCC	cellular protein metabolic process	0.000148823	179 of 1755 in the list, versus 291 of 4141 in the genome
P_no_RCC	protein metabolic process	0.000816375	221 of 1755 in the list, versus 376 of 4141 in the genome

4.3.2 RCCs in αHTMPs with solved 3D structure

Among the initial drives for designing and implementing this project was to elucidate whether RCCs have a fundamental role in the biogenesis of α-helical proteins. As already discussed in the introduction, the exact mechanisms of co-translational insertion and folding of αHTMPs still remain unclear. Nevertheless, while our work was in its final stages, two recent works have provided evidence of local pause of translational elongation at distinct sites to facilitate targeting membrane proteins to the translocon (Fluman et al., 2014; Pechmann et al., 2014). The two studies used ribosomal profiling data to estimate the translational rates and found that these non-optimal sites to be an additional parameter for the SRP arrest. On top of this we believe that RCCs play also an important role in the disposition of the proteins in the membrane bilayer, and their packing and assembly into the final 3D structure.

In our work, using topology data from experimentally determined structures of *E. coli* αHTMPs, we demonstrate that statistically significant RCCs detected with LaTcOm methods, are preferentially located in loops compared to TM helices. Moreover, the loops mapping to detected RCCs, are mostly periplasmic, a finding demonstrated with almost all tested RCC detection methods of LaTcOm (Table 32, 33 and 34). From these results we show that the RCCs detected, are preferentially located in loops (TP) than in helices (FP) and positive predictive values reach 71% (the highest among the methods found for %MinMax (cu)). Moreover, the second row in all tables shows that the RCCs, among those located in loops, prefer to be located in the periplasmic region with positive predictive value reaching 72% (again the highest found in %MinMax (cu)).

Table 32: Correlations regarding the positions of RCCs detected with %MinMax (cu) in TM helices or in loops.
TP are loops with detected RCCs, FN are loops with no detected RCCs, FP are helices with RCCs and TN are helices with no RCCs. RCCs which span both helices and loops were discarded from further analysis.

Analysis	**PPV**	**NPV**	**ACC**	**SEN**	**SPE**	**TP**	**FP**	**FN**	**TN**
loops/helices	70.97	47.49	48.68	6.77	96.82	22	9	303	274
periplasmic/cytoplasmic	72.73	55.78	56.92	10.67	96.57	16	6	134	169

Table 33: Correlations regarding the positions of RCCs detected with RiboTempo in TM helices or in loops.
For description see Table 32 above.

Analysis	PPV	NPV	ACC	SEN	SPE	TP	FP	FN	TN
loops/helices	54.55	46.55	47.13	7.36	92.93	24	20	302	263
periplasmic/cytoplasmic	70.83	55.63	56.75	11.26	96	17	7	134	168

Table 34: Correlations regarding the positions of RCCs detected with MSS in TM helices or in loops.
For description see Table 32 above.

Analysis	PPV	NPV	ACC	SEN	SPE	TP	FP	FN	TN
loops/helices	68.18	47.02	47.78	4.6	97.53	15	7	311	276
periplasmic/cytoplasmic	60	54.34	54.6	5.96	96.57	9	6	142	169

Moreover, we explored an additional feature, the interaction between successive in sequence TM helices. Nevertheless, this parameter reduced the number of available data points, thus no statistical significant results could be obtained (Appendix 2 - Table 51-56). However, we notice that in the case were RCCs were detected in connecting loops there is a small preference for successive helices that interact. Such a preference could be substantiated with ribosomal attenuation-driven synchronization of exit into the lipid bilayer of successive TM helices that need to be tightly packed in the membrane.

5 Conclusions

Recently, different procedures for identifying RCCs (as a proxy for estimating a measure of the translational elongation rate) have been proposed. However, no direct comparison of these methods has been systematically conducted. To address this, we implemented existing algorithms, the %MinMax and RiboTempo along with the MSS algorithm in order to have a consistent and reliable way of comparing the results of different methods, as well as providing a user interface for external users. The LaTcOm web server (Theodosiou and Promponas, 2012), as well as the standalone tool, were implemented for carrying out several analyses. The way RCCs are identified, as shown in several publications, is not uniform in the biological community. Different scales are used such as codon usage, tRNA experimental measurements (when available) and tRNA gene copy numbers which is also gaining ground because of its simplicity. All these different algorithms and scales make it difficult for comparisons to be made between studies. The uncertainty of which tool and which scale to use made us implement the existing algorithms and an additional, window-less approach, in order to identify RCCs in an unbiased manner.

As far as we know this is the first effort to benchmark existing methodologies for RCC detection and no similar work has been reported elsewhere. The LaTcOm standalone tool was applied to benchmark the methodologies using the well-annotated complement of protein coding genes in the complete genome of *E. coli* K12. The benchmarking we applied on the *E. coli* set revealed that there is no clear consistency between the different approaches. Nevertheless, the best positive correlation was found between %MinMax and MSS. There is clear bias in window-based methods to predict RCCs with least the length of the window. To avoid window bias issues, we propose that MSS can be alternatively used for detecting rare codon clusters. Nevertheless, if %MinMax is preferentially used in LaTcOm, codon usage should be applied as scale, as in the original publication (Clarke and Clark, 2008). When we experimented with tRNA values, the algorithm %MinMax dramatically over predicted RCCs. Nevertheless, it was initially designed to work with codon usage and fairly fails with other scales.

Our analysis concerning the existence of RCCs at the terminal sites revealed a tendency

for RCCs to be located in the 5' terminal as well as at the 3' terminal. RCCs are found closer to the 5’ terminal than in the 3’ terminal and there were statistical significant differences in their distributions, at least when the MSS method was used in the detection of RCCs. This triggered us to investigate the relation of RCCs at the 3’ terminal of multicystronic operons and the distance to the next gene. Nevertheless, no significant finding was identified.

Furthermore, we revealed that most of the sequences without RCCs are found in the cytoplasm, are involved with the ribosome or with metabolic processes. They are not involved with the membrane, whereas sequences with RCCs are found to be related with secretory proteins, involved with the membrane or the cell envelope. Conducting a detailed analysis on α-helical TM sequences we reveal a small preference for RCCs to be located in connecting loops and not in helices, especially in loops located in the periplasmic regions. For this analysis we relied only in experimentally derived annotations (i.e. mainly 3D structures and topology experiments with reporter fusions). Our work is straightforward to be extended to take into account predicted TM topologies. Such an approach may prove important, especially for non-model species, where experimental evidence is scarce.

In addition, we demonstrated a positive relation between RCCs and multidomain proteins, and revealed a preference for RCCs to be located in domain boundaries. This finding confirms and strengthens previous hypotheses on this matter. We, as others suggest that the slowdown of translation at these sites is most probably necessary for the correct protein folding of the nascent peptide domain.

What we finally conclude is that there is a preference for RCCs to be located in sequences that are somehow involved with the membrane, whether these are transmembrane that are co-translationally folded or just pass the membrane to be translocated to the cell wall such as TMβbs. The obvious connection would be that many TMβb proteins carry at 5’ terminal the secretory signal peptide. Nevertheless, exclusion of 5’ terminal RCCs from the correlation analysis of TMβbs and αHTMPs, did not alter the results. It would be very interesting to explore further, and in more detail, the position of RCCs in TMβb proteins and compare it with soluble β-barrel proteins.

Last, we propose that LaTcOm can have many applications both in basic and applied research. An interesting follow up would be the study of evolutionary conserved patterns of RCCs in various species, having in mind the different codon usage of each organism. Moreover, it can be used to rationally design heterologous gene expression studies, since translational profiles of proteins may reveal positions of RCCs that should not be altered for successfully folded proteins. Another potential application would be the explorations of applying LaTcOm in annotating studies for next generation sequencing experiments, since synonymous changes may lead to malfunction of proteins that cause diseases. Further on, an extension of LaTcOm should include tAi measurements from various species, information about domain boundaries, ribosomal profiling data, positions of Shine-Dalgarno sequences, annotation about the sequences under study and inclusion of codon usage for more model species or tRNA concentrations from different species.

The complex biological role of RCCs as a signal for translational pause in the mRNA sites has now started to be appreciated. We anticipate that our results will inspire and guide further research towards understanding the fine details of this mechanism and unravel the potential coupling with co-translational folding.

6 References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249.

Akashi, H., 1994. Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. Genetics 136, 927–935.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2007. Molecular Biology of the Cell, 5 edition. ed. Garland Science, New York.

Allert, M., Cox, J.C., Hellinga, H.W., 2010. Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames. J. Mol. Biol. 402, 905–918.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Anfinsen, C.B., 1972. The formation and stabilization of protein structure. Biochem. J. 128, 737–749.

Angov, E., 2011. Codon usage: Nature's roadmap to expression and folding of proteins. Biotechnol. J. 6, 650–659.

Artieri, C.G., Fraser, H.B., 2014. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res. gr.175893.114.

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., Blüthgen, N., 2013. Efficient translation initiation dictates codon usage at gene start. Mol. Syst. Biol. 9, n/a–n/a.

Bernardi, G., Bernardi, G., 1985. Codon usage and genome composition. J. Mol. Evol. 22, 363–365.

Bernsel, A., Viklund, H., Hennerdal, A., Elofsson, A., 2009. TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res. 37, W465–W468.

Bonekamp, F., Dalb?ge, H., Christensen, T., Jensen, K.F., 1989. Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in Escherichia coli. J. Bacteriol. 171, 5812–5816.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G., 2004. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinforma. Oxf. Engl. 20, 3710–3715.

Brunak, S., Engelbrecht, J., 1996. Protein structure and the sequential structure of mRNA: α-Helix and β-sheet signals at the nucleotide level. Proteins Struct. Funct. Bioinforma. 25, 237–252.

Bulmer, M., 1991. The Selection-Mutation-Drift Theory of Synonymous Codon Usage. Genetics 129, 897–907.

Burns, D.M., Beacham, I.R., 1985. Rare codons in E. coli and S. typhimurium signal sequences. FEBS Lett. 189, 318–324.

Cannarozzi, G.M., Schneider, A., 2012. Codon Evolution: Mechanisms and Models. Oxford University Press.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., Barral, Y., 2010. A Role for Codon Order in Translation

Dynamics. Cell 141, 355–367.
Carulli, J.P., Krane, D.E., Hartl, D.L., Ochman, H., 1993. Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics 134, 837–845.
Cavalier-Smith, T., Cavalier-Smith, T., 2000. Membrane heredity and early chloroplast evolution. Trends Plant Sci. 5, 174–182.
Chamary, J.V., Hurst, L.D., 2009. The Price of Silent Mutations. Sci. Am. 300, 46–53.
Charneski, C.A., Hurst, L.D., 2014. Positive Charge Loading at Protein Termini Is Due to Membrane Protein Topology, Not a Translational Ramp. Mol. Biol. Evol. 31, 70–84.
Charneski, C.A., Hurst, L.D., 2013. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol 11, e1001508.
Chartier, M., Gaudreault, F., Najmanovich, R., 2012. Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. Bioinforma. Oxf. Engl. 28, 1438–1445.
Chevance, F.F.V., Le Guyon, S., Hughes, K.T., 2014. The Effects of Codon Context on In Vivo Translation Speed. PLoS Genet 10, e1004392.
Clarke, T.F., Clark, P.L., 2010. Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. BMC Genomics 11, 118.
Clarke, T.F., IV, Clark, P.L., 2008. Rare Codons Cluster. PLoS ONE 3, e3412.
Cortazzo, P., Cerveñansky, C., Marín, M., Reiss, C., Ehrlich, R., Deana, A., 2002. Silent mutations affect in vivo protein folding in Escherichia coli. Biochem. Biophys. Res. Commun. 293, 537–541.
Crick, F.H., 1958. On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163.
Crick, F.H., 1970. Central dogma of molecular biology. Nature 227, 561–563.
Crombie, T., Swaffield, J.C., Brown, A.J.P., 1992. Protein folding within the cell is influenced by controlled rates of polypeptide elongation. J. Mol. Biol. 228, 7–12.
Cuff, A.L., Sillitoe, I., Lewis, T., Clegg, A.B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J., Orengo, C.A., 2011. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res. 39, D420–D426.
Curran, J.F., Yarus, M., 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J. Mol. Biol. 209, 65–77.
Daley, D.O., Rapp, M., Granseth, E., Melén, K., Drew, D., Heijne, G. von, 2005. Global Topology Analysis of the Escherichia coli Inner Membrane Proteome. Science 308, 1321–1323.
Dessen, P., Képès, F., 2000. The PAUSE software for analysis of translational control over protein targeting: Application to E. nidulans membrane proteins. Gene 244, 89–96.
Dittmar, K.A., Goodenbour, J.M., Pan, T., 2006. Tissue-Specific Differences in Human Transfer RNA Expression. PLoS Genet 2, e221.
Dong, H., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. 260, 649–663.
Dos Reis, M., Wernisch, L., Savva, R., 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res. 31, 6976–6985.
Elofsson, A., Heijne, G. von, 2007. Membrane Protein Structure: Prediction versus Reality. Annu. Rev. Biochem. 76, 125–140.
Emilsson, V., Kurland, C.G., 1990. Growth rate dependence of transfer RNA abundance in Escherichia coli. EMBO J. 9, 4359–4366.

Emilsson, V., Näslund, A.K., Kurland, C.G., 1993. Growth-rate-dependent Accumulation of Twelve tRNA Species in Escherichia coli. J. Mol. Biol. 230, 483–491.
Ermolaeva, M.D., 2001. Synonymous codon usage in bacteria. Curr. Issues Mol. Biol. 3, 91–97.
Fluman, N., Navon, S., Bibi, E., Pilpel, Y., 2014. mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. eLife e03440.
Fox, N.K., Brenner, S.E., Chandonia, J.-M., 2013. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. gkt1240.
Freeman, T.C., Jr, Wimley, W.C., 2012. TMBB-DB: a transmembrane β-barrel proteome database. Bioinforma. Oxf. Engl. 28, 2425–2430.
Gloge, F., Becker, A.H., Kramer, G., Bukau, B., 2014. Co-translational mechanisms of protein maturation. Curr. Opin. Struct. Biol., Folding and binding / Nucleic acids and their protein complexes 24, 24–33.
Goldman, E., Rosenberg, A.H., Zubay, G., Studier, W.F., 1995. Consecutive Low-usage Leucine Codons Block Translation Only When Near the 5′ End of a Message inEscherichia coli. J. Mol. Biol. 245, 467–473.
Goodman, D.B., Church, G.M., Kosuri, S., 2013. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. Science 342, 475–479.
Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.
Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.
Gray, M.W., Burger, G., Lang, B.F., 1999. Mitochondrial Evolution. Science 283, 1476–1481.
Guggenmoos-Holzmann, I., van Houwelingen, H.C., 2000. The (in)validity of sensitivity and specificity. Stat. Med. 19, 1783–1792.
Guisez, Y., Robbens, J., Remaut, E., Fiers, W., 1993. Folding of the MS2 Coat Protein in Escherichia coli is Modulated by Translational Pauses Resulting from mRNA Secondary Structure and Codon Usage: A Hypothesis. J. Theor. Biol. 162, 243–252.
Gustafsson, C., Govindarajan, S., Minshull, J., 2004. Codon bias and heterologous protein expression. Trends Biotechnol. 22, 346–353.
Hayes, C.S., Bose, B., Sauer, R.T., 2002. Stop codons preceded by rare arginine codons are efficient determinants of SsrA tagging in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 99, 3440–3445.
Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., von Heijne, G., 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 433, 377–381.
Huang, L., Kulldorff, M., Gregorio, D., 2007. A Spatial Scan Statistic for Survival Data. Biometrics 63, 109–118.
Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.
Ikemura, T., 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. 158, 573–597.
Ikemura, T., 1981a. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 146,

1–21.
Ikemura, T., 1981b. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389–409.
Ikemura, T., Wada, K., 1991. Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res. 19, 4333–4339.
Ingolia, N.T., 2010. Chapter 6 - Genome-Wide Translational Profiling by Ribosome Footprinting, in: Jonathan Weissman; Christine Guthrie and Gerald R. Fink (Ed.), Methods in Enzymology, Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis. Academic Press, pp. 119–142.
Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., Weissman, J.S., 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324, 218–223.
Itakura, K., Hirose, T., Crea, R., Riggs, A.D., Heyneker, H.L., Bolivar, F., Boyer, H.W., 1977. Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. Science 198, 1056–1063.
Jacob, F., Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356.
Jacob, W.F., Santer, M., Dahlberg, A.E., 1987. A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. Proc. Natl. Acad. Sci. 84, 4757–4761.
Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., Ikemura, T., 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53, 290–298.
Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.
Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U. S. A. 87, 2264–2268.
Karlin, S., Mrázek, J., Campbell, A.M., 1998. Codon usages in different gene classes of the Escherichia coli genome. Mol. Microbiol. 29, 1341–1355.
Képès, F., 1996. The "+70 pause": Hypothesis of a translational control of membrane protein assembly. J. Mol. Biol. 262, 77–86.
Kink, J.A., Maley, M.E., Ling, K.-Y., Kanabrocki, J.A., Kung, C., 1991. Efficient Expression of the Paramecium Calmodulin Gene in Escherichia coli after Four TAA-to-CAA Changes through a Series of Polymerase Chain Reactions. J. Protozool. 38, 441–447.
Kleinschmidt, J.H., Tamm, L.K., 2002. Secondary and Tertiary Structure Formation of the β-Barrel Membrane Protein OmpA is Synchronized and Depends on Membrane Thickness. J. Mol. Biol. 324, 319–330.
Knight, R.D., Freeland, S.J., Landweber, L.F., 2001. Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. 2, 49–58.
Koide, T., Reiss, D.J., Bare, J.C., Pang, W.L., Facciotti, M.T., Schmid, A.K., Pan, M.,

Marzolf, B., Van, P.T., Lo, F.-Y., Pratap, A., Deutsch, E.W., Peterson, A., Martin, D., Baliga, N.S., 2009. Prevalence of transcription promoters within archaeal operons and coding sequences. Mol. Syst. Biol. 5, 285.

Komar, A.A., 2009. A pause for thought along the co-translational folding pathway. Trends Biochem. Sci. 34, 16–24.

Komar, A.A., Jaenicke, R., 1995. Kinetics of translation of γB crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. FEBS Lett. 376, 195–198.

Komar, A.A., Lesnik, T., Reiss, C., 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett. 462, 387–391.

Krasheninnikov, I.A., Komar, A.A., Adzhubei, I.A., 1991. Nonuniform size distribution of nascent globin peptides, evidence for pause localization sites, and a contranslational protein-folding model. J. Protein Chem. 10, 445–453.

Krasheninnikov, I.A., Komar, A.A., Adzhubeĭ, I.A., 1988. [Role of the rare codon clusters in defining the boundaries of polypeptide chain regions with identical secondary structures in the process of co-translational folding of proteins]. Dokl. Akad. Nauk SSSR 303, 995–999.

Krüger, M.K., Pedersen, S., Hagervall, T.G., Sørensen, M.A., 1998. The modification of the wobble base of tRNAGlu modulates the translation rate of glutamic acid codons in vivo. J. Mol. Biol. 284, 621–631.

Kudla, G., Murray, A.W., Tollervey, D., Plotkin, J.B., 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324, 255–258.

Lakkaraju, A.K.K., Mary, C., Scherrer, A., Johnson, A.E., Strub, K., 2008. SRP maintains nascent chains translocation-competent by slowing translation rates to match limiting numbers of targeting sites. Cell 133, 440–451.

Li, G.-W., Oh, E., Weissman, J.S., 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484, 538–541.

Lloyd, A.T., Sharp, P.M., 1992. Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. Nucleic Acids Res. 20, 5289–5295.

Lynn, D.J., Singer, G.A.C., Hickey, D.A., 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res. 30, 4272–4277.

Magrane, M., Consortium, U., 2011. UniProt Knowledgebase: a hub of integrated protein data. Database J. Biol. Databases Curation 2011.

Mahlab, S., Linial, M., 2014. Speed Controls in Translating Secretory Proteins in Eukaryotes - an Evolutionary Perspective. PLoS Comput. Biol. 10.

Makhoul, C.H., Trifonov, E.N., 2002. Distribution of Rare Triplets Along mRNA and Their Relation to Protein Folding. J. Biomol. Struct. Dyn. 20, 413–420.

Marin, M., 2008. Folding at the rhythm of the rare codon beat. Biotechnol. J. 3, 1047–1057.

Martin, A.C.R., 2005. Mapping PDB chains to UniProtKB entries. Bioinformatics 21, 4297–4301.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

McKown, R.L., Raab, R.W., Kachelries, P., Caldwell, S., Laurie, G.W., 2013. Conserved Regional 3? Grouping of Rare Codons in the Coding Sequence of Ocular Prosecretory Mitogen Lacritin. Invest. Ophthalmol. Vis. Sci. 54, 1979–1987.

McLachlan, A.D., Staden, R., Boswell, D.R., 1984. A method for measuring the non-random bias of a codon usage table. Nucleic Acids Res. 12, 9567–9575.

Medigue, C., Viari, A., Henaut, A., Danchin, A., 1993. Colibri: a functional data base for the Escherichia coli genome. Microbiol. Rev. 57, 623–654.
Moriyama, E.N., Powell, J.R., 1997. Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. 45, 514–523.
Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H., Bernardi, G., 2001. Translational Selection on Codon Usage in Xenopus laevis. Mol. Biol. Evol. 18, 1703–1707.
Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28, 292.
Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814.
Nørholm, M.H.H., Light, S., Virkki, M.T.I., Elofsson, A., von Heijne, G., Daley, D.O., 2012. Manipulating the genetic code for membrane protein production: What have we learnt so far? Biochim. Biophys. Acta BBA - Biomembr., Protein Folding in Membranes 1818, 1091–1096.
Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P., Olsson, O., 1983. Overlapping Genes. Annu. Rev. Genet. 17, 499–525.
Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R.J., Typas, A., Gross, C.A., Kramer, G., Weissman, J.S., Bukau, B., 2011. Selective ribosome profiling reveals the co-translational chaperone action of trigger factor in vivo. Cell 147, 1295–1308.
Oldfield, C.J., Cheng, Y., Cortese, M.S., Brown, C.J., Uversky, V.N., Dunker, A.K., 2005. Comparing and Combining Predictors of Mostly Disordered Proteins†. Biochemistry (Mosc.) 44, 1989–2000.
Osborne, A.R., Rapoport, T.A., van den Berg, B., 2005. Protein Translocation by the Sec61/Secy Channel. Annu. Rev. Cell Dev. Biol. 21, 529–550.
Pancsa, R., Tompa, P., 2012. Structural Disorder in Eukaryotes. PLoS ONE 7, e34687.
Papanastasiou, M., Orfanoudaki, G., Koukaki, M., Kountourakis, N., Sardis, M.F., Aivaliotis, M., Karamanou, S., Economou, A., 2013. The Escherichia coli peripheral inner membrane proteome. Mol. Cell. Proteomics MCP 12, 599–610.
Pavlov, M.Y., Watts, R.E., Tan, Z., Cornish, V.W., Ehrenberg, M., Forster, A.C., 2009. Slow peptide bond formation by proline and other N-alkylamino acids in translation. Proc. Natl. Acad. Sci. U. S. A. 106, 50–54.
Pechmann, S., Chartron, J.W., Frydman, J., 2014. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nat. Struct. Mol. Biol. 21, 1100–1105.
Pechmann, S., Frydman, J., 2013. Evolutionary conservation of codon optimality reveals hidden signatures of co-translational folding. Nat. Struct. Mol. Biol. 20, 237–243.
Pedersen, S., 1984. Escherichia coli ribosomes translate in vivo with variable rate. EMBO J. 3, 2895–2898.
Percudani, R., Pavesi, A., Ottonello, S., 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J. Mol. Biol. 268, 322–330.
Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786.
Ponnala, L., 2010. Detecting slow-translating regions in E.coli. Int. J. Bioinforma. Res. Appl. 6, 522–530.
Power, P.M., Jones, R.A., Beacham, I.R., Bucholtz, C., Jennings, M.P., 2004. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of Escherichia coli. Biochem. Biophys. Res. Commun. 322, 1038–1044.
Prilusky, J., Bibi, E., 2009. Studying membrane proteins through the eyes of the genetic

code revealed a strong uracil bias in their coding mRNAs. Proc. Natl. Acad. Sci. 106, 6662–6666.

Purvis, I.J., Bettany, A.J.E., Santiago, T.C., Coggins, J.R., Duncan, K., Eason, R., Brown, A.J.P., 1987. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo: A hypothesis. J. Mol. Biol. 193, 413–417.

Quax, T.E.F., Wolf, Y.I., Koehorst, J.J., Wurtzel, O., van der Oost, R., Ran, W., Blombach, F., Makarova, K.S., Brouns, S.J.J., Forster, A.C., Wagner, E.G.H., Sorek, R., Koonin, E.V., van der Oost, J., 2013. Differential Translation Tunes Uneven Production of Operon-Encoded Proteins. Cell Rep. 4, 938–944.

Raine, A., Ullers, R., Pavlov, M., Luirink, J., Wikberg, J.E.S., Ehrenberg, M., 2003. Targeting and insertion of heterologous membrane proteins in E. coli. Biochimie 85, 659–668.

R Development Core Team, 2008. R : A language and environment for statistical computing. R Foundation for Statistical Computing.

Regalia, M., Rosenblad, M.A., Samuelsson, T., 2002. Prediction of signal recognition particle RNA genes. Nucleic Acids Res. 30, 3368–3377.

Rehling, P., Pfanner, N., Meisinger, C., 2003. Insertion of Hydrophobic Membrane Proteins into the Inner Mitochondrial Membrane—A Guided Tour. J. Mol. Biol. 326, 639–657.

Ruzzo, W.L., Tompa, M., 1999. A linear time algorithm for finding all maximal scoring subsequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol. 234–241.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J.S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, A., Porrón-Sotelo, L., Huerta, A.M., Bonavides-Martínez, C., Balderas-Martínez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chávez, V., Hernández-Alvarez, A., Morett, E., Collado-Vides, J., 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 41, D203–213.

Saunders, R., Deane, C.M., 2010. Synonymous codon usage influences the local protein structure observed. Nucleic Acids Res. 38, 6719–6728.

Schauder, B., McCarthy, J.E.G., 1989. The role of bases upstream of the Shine-Dalgarno region and in the coding sequence in the control of gene expression in Escherichia coli: Translation and stability of mRNAs in vivo. Gene 78, 59–72.

Sharp, P.M., Li, W.H., 1989. On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. 28, 398–402.

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Shields, D.C., Sharp, P.M., 1987. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. Nucleic Acids Res. 15, 8023–8040.

Shields, D.C., Sharp, P.M., Higgins, D.G., Wright, F., 1988. “Silent” sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5, 704–716.

Shine, J., Dalgarno, L., 1974. The 3?-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. Proc. Natl. Acad. Sci. U. S. A. 71, 1342–1346.

Silva, R.M., Miranda, I., Moura, G., Santos, M.A.S., 2004. Yeast as a model organism for studying the evolution of nonstandard genetic codes. Brief. Funct. Genomic. Proteomic. 3, 35–46.

Sørensen, M.A., Pedersen, S., 1998. Determination of the Peptide Elongation Rate In Vivo, in: Martin, R. (Ed.), Protein Synthesis, Methods in Molecular Biology. Springer New York, pp. 129–142.

Srinivasan, G., James, C.M., Krzycki, J.A., 2002. Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science 296, 1459–1462.

Steitz, J.A., Jakes, K., 1975. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. Proc. Natl. Acad. Sci. 72, 4734–4738.

Supek, F., Smuc, T., 2010. On Relevance of Codon Usage to Expression of Synthetic and Natural Genes in Escherichia coli. Genetics 185, 1129–1134.

Tamm, L.K., Hong, H., Liang, B., 2004. Folding and assembly of?-barrel membrane proteins. Biochim. Biophys. Acta BBA - Biomembr., Lipid-Protein Interactions 1666, 250–263.

Tanner, D.R., Cariello, D.A., Woolstenhulme, C.J., Broadbent, M.A., Buskirk, A.R., 2009. Genetic Identification of Nascent Peptides That Induce Ribosome Stalling. J. Biol. Chem. 284, 34809–34818.

Thanaraj, T.A., Argos, P., 1996a. Ribosome-mediated translational pause and protein domain organization. Protein Sci. Publ. Protein Soc. 5, 1594–1612.

Thanaraj, T.A., Argos, P., 1996b. Protein secondary structural types are differentially coded on messenger RNA. Protein Sci. Publ. Protein Soc. 5, 1973–1983.

Theodosiou, A., Promponas, V.J., 2012. LaTcOm: a web server for visualizing rare codon clusters in coding sequences. Bioinformatics 28, 591–592.

Trifonov, E.N., 2011. Thirty Years of Multiple Sequence Codes. Genomics Proteomics Bioinformatics 9, 1–6.

Tsaousis, G.N., Tsirigos, K.D., Andrianou, X.D., Liakopoulos, T.D., Bagos, P.G., Hamodrakas, S.J., 2010. ExTopoDB: a database of experimentally derived topological models of transmembrane proteins. Bioinforma. Oxf. Engl. 26, 2490–2492.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpel, Y., 2010a. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell 141, 344–354.

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., Ziv-Ukelson, M., 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 12, R110.

Tuller, T., Waldman, Y.Y., Kupiec, M., Ruppin, E., 2010b. Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. 107, 3645–3650.

Tusnády, G.E., Dosztányi, Z., Simon, I., 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 20, 2964–2972.

Varenne, S., Buc, J., Lloubes, R., Lazdunski, C., 1984. Translation is a non-uniform process: Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. 180, 549–576.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., Newton, C.D., Dunker, A.K., 2005. DisProt: a database of protein disorder. Bioinforma. Oxf. Engl. 21, 137–140.

Wang, G., Dunbrack, R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.

Wang, L., Wessler, S.R., 2001. Role of mRNA Secondary Structure in Translational Repression of the Maize Transcriptional ActivatorLc. Plant Physiol. 125, 1380–1387.

White, S.H., 2004. The progress of membrane protein structure determination. Protein Sci. 13, 1948–1949.

Widmann, M., Clairo, M., Dippon, J., Pleiss, J., 2008. Analysis of the distribution of functionally relevant rare codons. BMC Genomics 9, 207.

Wright, F., 1990. The "effective number of codons" used in a gene. Gene 87, 23–29.

Zalucki, Y.M., Jennings, M.P., 2007. Experimental confirmation of a key role for non-optimal codons in protein export. Biochem. Biophys. Res. Commun. 355, 143–148.

Zeeberg, B., 2002. Shannon Information Theoretic Computation of Synonymous Codon Usage Biases in Coding Regions of Human and Mouse Genomes. Genome Res. 12, 944–955.

Zemla, A., Venclovas, C., Fidelis, K., Rost, B., 1999. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 34, 220–223.

Zhang, G., Hubalewska, M., Ignatova, Z., 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. 16, 274–280.

Zhang, G., Ignatova, Z., 2009. Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. PloS One 4, e5036.

Zinoni, F., Birkmann, A., Stadtman, T.C., Bock, A., 1986. Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 83, 4650–4654.

Appendix 1

```
>gi|545778205|gb|U00096.3|:190-255 Escherichia coli str. K-12 substr. MG1655, complete genome
ATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGA
>gi|545778205|gb|U00096.3|:337-2799 Escherichia coli str. K-12 substr. MG1655, complete genome
ATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTC
TGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCT
GGTGGCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATT
TTTGCCGAACTTTTGACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCG
TCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAG
CATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCG
CGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTA
CCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGAT
GGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCT
GCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATA
```

Figure 37: Part of the content of file U00096.fnn downloaded from GenBank.

DNA coding genes in FASTA format for *E. coli* K12 MG1655 strain.

(ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/).

```
Escherichia coli str. K-12 substr. MG1655, complete genome. - 1..4641652
4141 proteins
Location        Strand  Length  PID     Gene    Synonym Code    COG     Product
190..255        +       21      1786182 thrL    b0001   -       -       thr operon leader peptide
337..2799       +       820     1786183 thrA    b0002   -       -       fused aspartokinase I and homoserine dehydrogenase I
2801..3733      +       310     1786184 thrB    b0003   -       -       homoserine kinase
3734..5020      +       428     1786185 thrC    b0004   -       -       threonine synthase
5234..5530      +       98      1786186 yaaX    b0005   -       -       predicted protein
5683..6459      -       258     1786187 yaaA    b0006   -       -       peroxide resistance protein, lowers intracellular iron
6529..7959      -       476     1786188 yaaJ    b0007   -       -       predicted transporter
8238..9191      +       317     1786189 talB    b0008   -       -       transaldolase B
```

Figure 38: Part of content of file U00096.ptt downloaded from GenBank.
Information regarding the coding genes of *E. coli* K12 MG1655 strain.

```
Escherichia coli str. K-12 substr. MG1655, complete genome. - 1..4641652
176 RNAs
Location        Strand  Length  PID     Gene    Synonym Code    COG     Product
16952..17006    +       55      545778205       sokC    b4413   -       -       -
77367..77593    +       227     545778205       sgrS    b4577   -       -       -
189712..189847  +       136     545778205       tff     b4414   -       -       -
223771..225312  +       1542    545778205       rrsH    b0201   -       -       16S ribosomal RNA of rrnH operon
225381..225457  +       77      545778205       ileV    b0202   -       -       Ile tRNA
225500..225575  +       76      545778205       alaV    b0203   -       -       Ala tRNA
225759..228662  +       2904    545778205       rrlH    b0204   -       -       23S ribosomal RNA of rrnH operon
228756..228875  +       120     545778205       rrfH    b0205   -       -       5S ribosomal RNA of rrnH operon
```

Figure 39: Part of the content of file U00096.rnt downloaded from GenBank.

Information regarding the RNA genes of *E. coli* K12 MG1655 strain.

(ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/).

Table 35: Number of sequence analyzed and duration of RCC detection with each method of LaTcOm for a single run.

Method	No of sequences analyzed	Duration in sec
MSS	4136	403
%MinMax	4128	1323
RiboTempo	4128	732

Table 36: Discarded sequences from LaTcOm analysis.
Sequences in which window size exceeded number of codons with window based methods of LaTcOm: %MinMax and RiboTempo.

GI	Annotation	UniProt accesion	Length
226510957	Small toxic protein IbsB (inferred from homology in UniProt)	C1P608	18
1788329	his operon leader peptide (inferred from homology in UniProt)	P60995	17
308199521	regulatory leader peptide for mgtA (inferred from homology in UniProt)	E2JKY7	18
226510989	Uncharacterized protein YjeV (evidence at the protein level)	C1P621	18
1788950	Phe operon leader peptide (predicted)	P0AD72	16
226510987	Uncharacterized protein IlvX (evidence at the protein level)	C1P619	17
1787519	trp operon leader peptide (evidence at the protein level)	P0AD92	15
1788008	phenylalanyl-tRNA synthetase operon leader peptide (predicted)	P0AD74	15

Table 37: Sequences that were discarded due to "in-frame stop codons".

In-frame stop codons:GI number ID	Gene name	Annotation from Ecogene (http://www.ecogene.org)
3868721	fdhF	Formate dehydrogenase H (The UGA stop codon 140 is translated as selenocysteine in vivo-)
3868720	fdoG	formate dehydrogenase-O, large subunit (The UGA stop codon 196 is translated as selenocysteine in vivo
3868719	fdnG	formate dehydrogenase-N, alpha subunit, nitrate-inducible (he UGA stop codon 196 is translated as selenocysteine in vivo)

```
>Ec_3081
HHHCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCC
CCCCCCCCCCCCC
>
>Ec_3609
HHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCC
CCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

>Ec_2457
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

>Ec_3028
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHH
```

Figure 40: Example of output format of the file created by transform_files_for_SOV_MCC.pl. “C” strings are position of clusters and “H” strings are positions of non-detected clusters

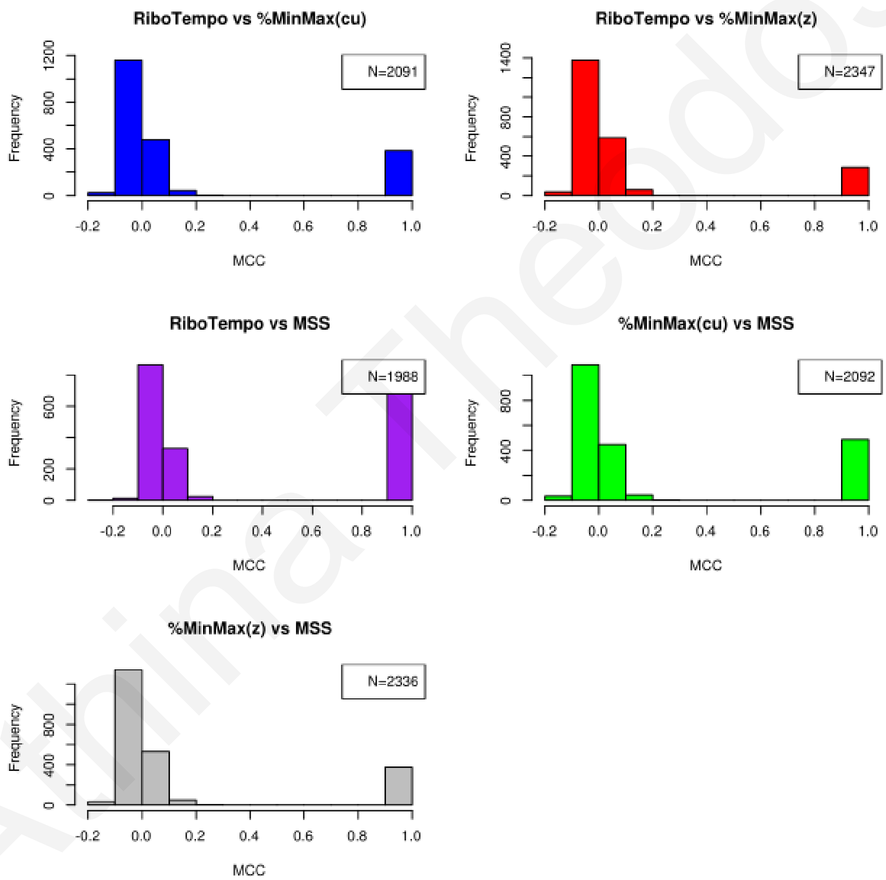

Figure 41: Distribution of MCC v1 values from shuffled sequences.
MCC values are calculated in the comparisons of reference and predicted set of genes. Analysis was made with all shuffled genes as described in the Methods of Chapter 3section. The x axis shows the MCC value that can range from -1 to 1, the y axis shows the frequency from N compared MCC values. Graphs were generated in R statistical environment (R Development Core Team, 2008).

Table 38: Random MCC v1 distribution analysis from shuffled sequences.
Statistical properties were computed in R statistical environment (R Development Core Team, 2008).

Reference	Predicted	Mean	Median	SD	Overall MCC mean
RiboTempo	%MinMax(cu)	-0.001	-0.001	0.065	0.000
RiboTempo	%MinMax(z)	0.000	0.000	0.065	0.000
RiboTempo	MSS	-0.001	-0.001	0.071	-0.002
%MinMax(cu)	MSS	0.001	0.000	0.066	0.001
%MinMax(z)	MSS	-0.001	-0.001	0.065	-0.001

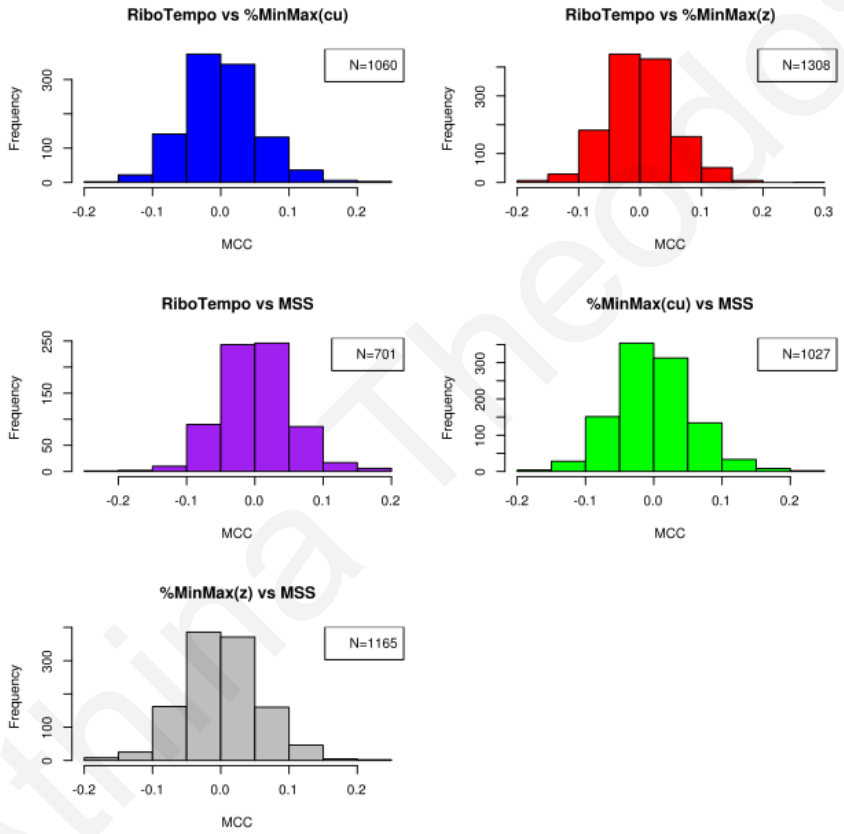

Figure 42: MCC v2 distribution values from shuffled sequences.
Description shown in Figure 41.

Table 39: Random MCC v2 distribution analysis from shuffled sequences.
Statistical properties were computed in R statistical environment (R Development Core Team, 2008).

Reference	Predicted	Mean	Median	SD	Overall MCC mean
RiboTempo	%MinMax(cu)	0.001	-0.002	0.054	0.032
RiboTempo	%MinMax(z)	0.001	0.000	0.055	0.029
RiboTempo	MSS	0.002	0.001	0.052	0.034
%MinMax(cu)	MSS	0.000	-0.003	0.056	0.085
%MinMax(z)	MSS	0.001	0.000	0.056	0.072

Table 40: Random SOV v1 distribution analysis from shuffled sequences.
Statistical properties were computed in R statistical environment (R Development Core Team, 2008).

Reference	Predicted	Mean	Median	SD
RiboTempo	%MinMax(cu)	36.700	32.400	33.211
RiboTempo	%MinMax(z)	31.710	30.700	29.353
RiboTempo	MSS	51.380	38.300	40.080
%MinMax(cu)	MSS	42.610	35.500	34.104
%MinMax(z)	MSS	34.950	30.800	31.573

Table 41: Random SOV v2 distribution analysis from shuffled sequences.
Statistical properties were computed in R statistical environment (R Development Core Team, 2008).

Reference	Predicted	Mean	Median	SD
RiboTempo	%MinMax(cu)	22.4200	27.3500	15.5950
RiboTempo	%MinMax(z)	22.2000	26.8500	15.5300
RiboTempo	MSS	21.6100	26.3000	15.9840
%MinMax(cu)	MSS	25.2500	31.1000	14.7090
%MinMax(z)	MSS	22.4300	26.5000	14.7400

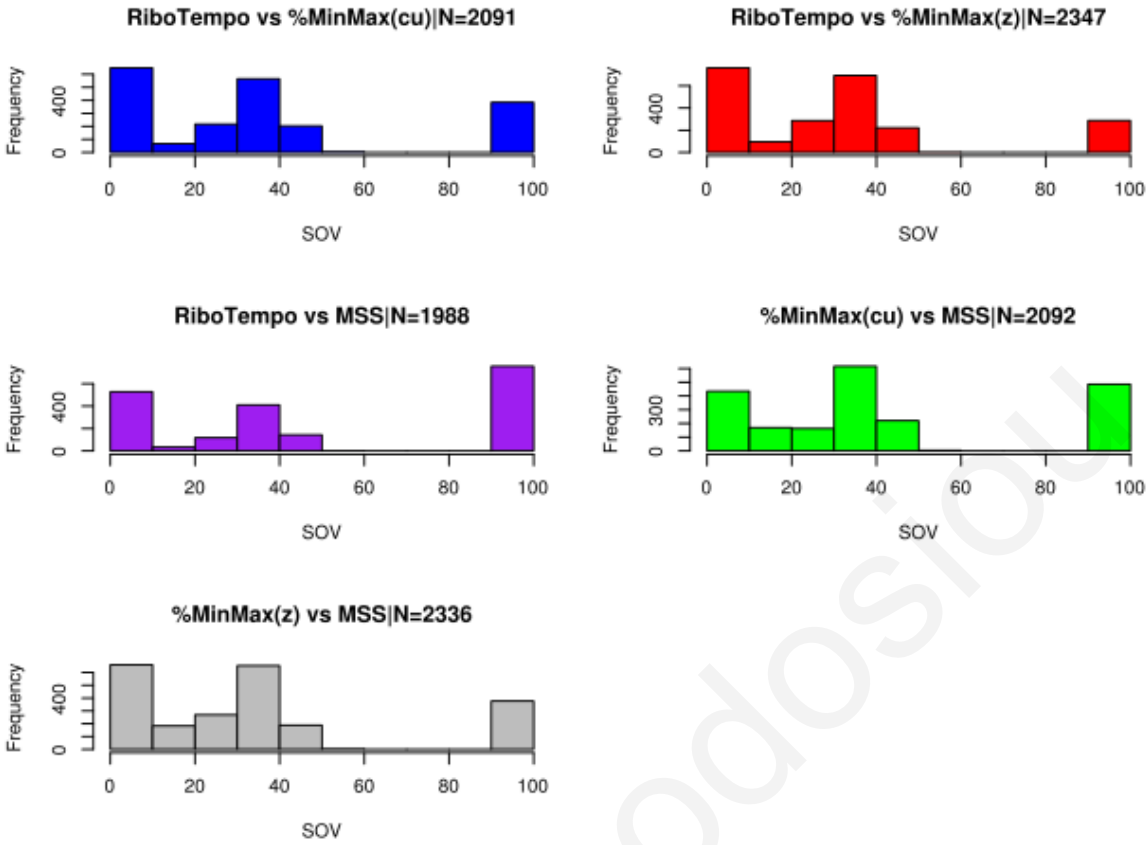

Figure 43: Distribution of SOV v1 values from shuffled sequences.
SOV values are calculated in the comparisons of reference and predicted set of genes. Analysis was made with all shuffled genes as described in the Methods section. The x axis shows the SOV value that can range from 0 to 100 and the y axis shows the frequency from N compared SOV values. Graphs were generated with R statistical environment (R Development Core Team, 2008).

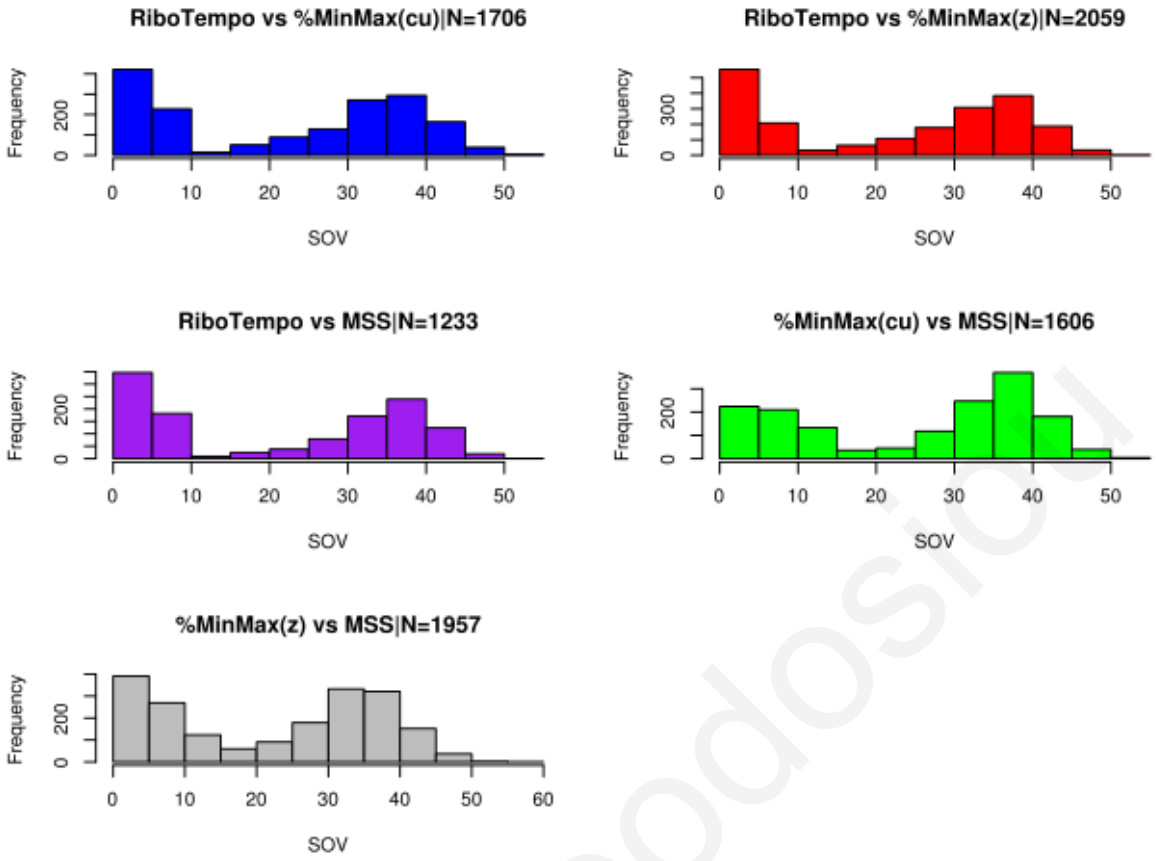

Figure 44: Distribution of SOV v2 values from shuffled sequences.
Read description shown in Figure 43.

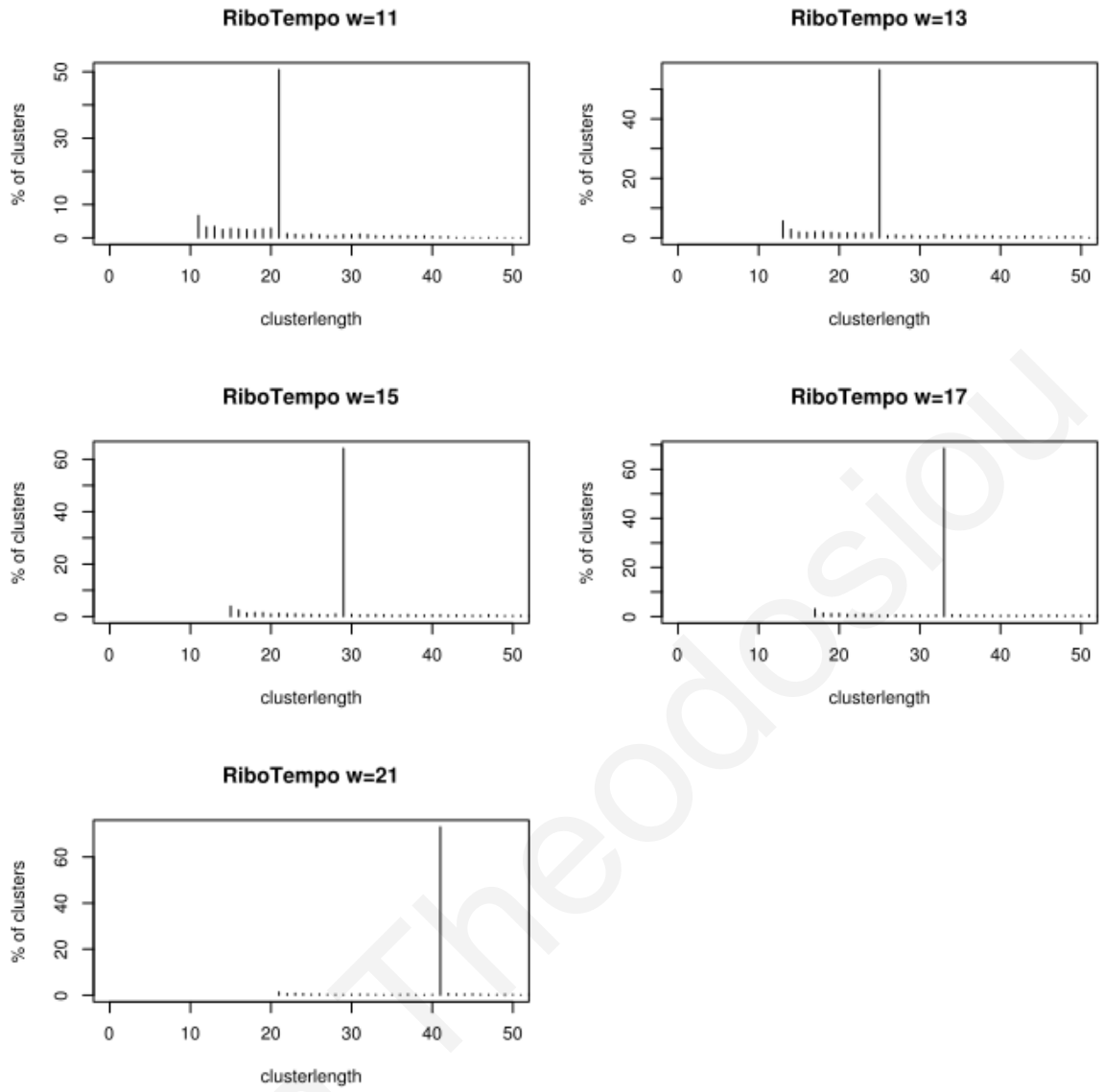

Figure 45: Detailed distribution of clusterlength of RCCs detected with RiboTempo with different window thresholds.
In this analysis window thresholds 11,13,15,17 and 21 were applied. Graphs were produced in the R statistical environment (R Development Core Team, 2008).

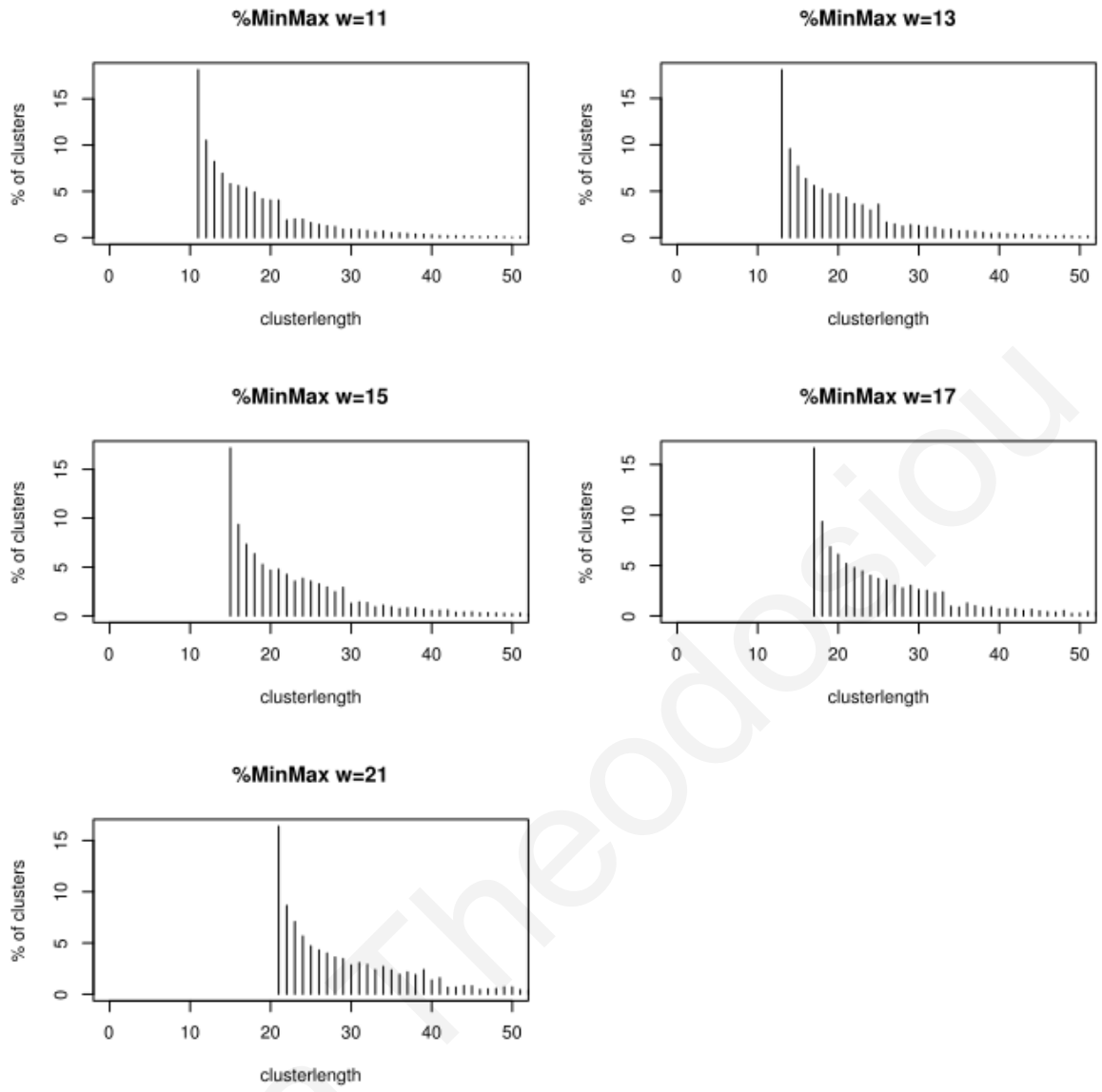

Figure 46: Detailed distribution of clusterlength of RCCs detected with %MinMax with different window thresholds.
In this analysis window thresholds 11,13,15,17 and 21 were applied. Graphs were produced in the R statistical environment (R Development Core Team, 2008).

Table 42: P-values from Wilcoxon rank test of all RCCs distance distribution from 3' compared with distributions from the 5' terminus.
The wilcox.test() function was used from the R statistical environment (R Development Core Team, 2008).

Method	**p-value**
%Minmax (cu)	0.968
Ribotempo	0.904
MSS	0.837

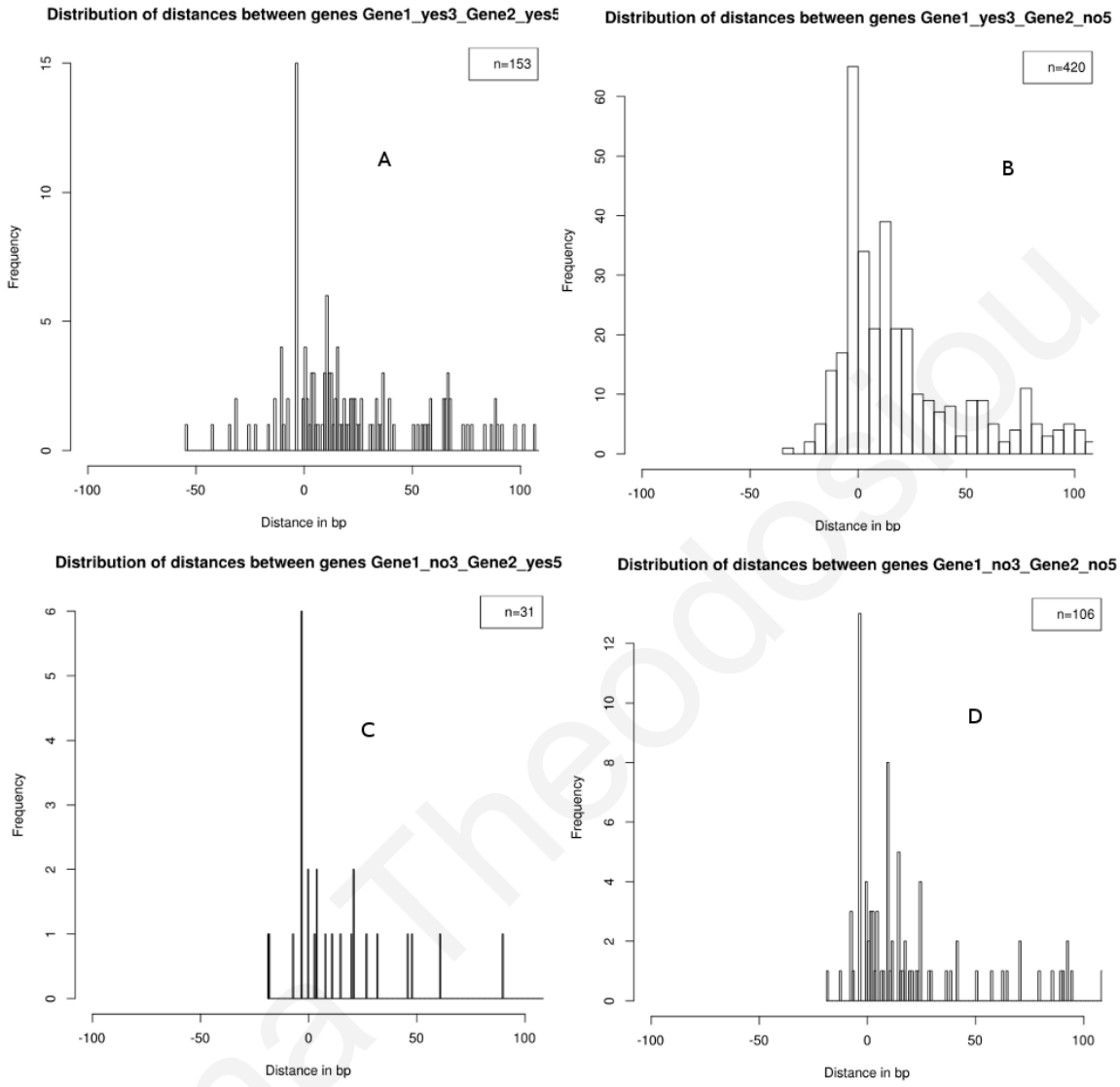

Figure 47: Distribution of the distance in bp between adjacent gene in operons.
A) Cases were the first gene has RCCs at the 3' terminal and the next gene has RCCs at the 5' terminal. B) Cases were the first gene has RCCs at the 3' terminal and the next gene does not have RCC at the 5' terminal. C) Cases were the first gene does not have RCC at the 3' terminal but the next gene has RCCs at the 5' terminal and D) Cases were neither the first has RCC at the 3' terminal nor the next gene has RCCs at the 5'.

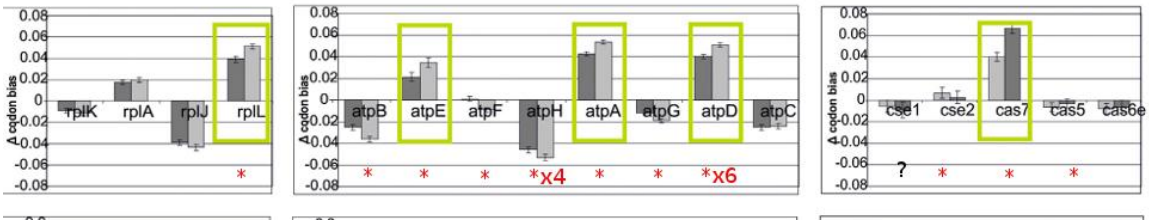

Figure 48: Presence of RCCs in operons.
Figure is adapted from (Quax et al., 2013) and is demonstrating the codon bias in three operon complexes the ribosomal protein operon L7/L12; the F-type ATPase and type I-E Cascade complex. We demonstrate with red asterisk the existence of RCC detected with MSS for the genes of the three operon complexes. The symbol "?" was placed for the gene that could not be identified in our dataset.

Appendix 2

```
>DisProt|DP00001|uniprot|Q9HFQ6|sp|RLA3_CANAL #1-108
MSTEASVSYAALILADAEQEITSEKLLAITKAAGANVDQVWADVFAKAVEGKNLKELLFS
FAAAAPASGAAAGSASGAAAGGEAAAEEAAEEEAAEESDDDMGFGLFD

>DisProt|DP00002|uniprot|P02400|sp|RLA4_YEAST #1-110
MKYLAAYLLLVQGGNAAPSAADIKAVVESVGAEVDEARINELLSSLEGKGSLEEIIAEGQ
KKFATVPTGGASSAAAGAAGAAAGGDAEEEKEEEAKEESDDDMGFGLFD

>DisProt|DP00003|uniprot|P03265|sp|DNB2_ADE05 #401-406 &332-400 &465-529 &407-452 #453-464 #297-331 &174-296
MASREEEQRETTPERGRGAARRPPTMEDVSSPSPSPPPPRAPPKKRMRRRIESEDEEDSS
QDALVPRTPSPRPSTSAADLAIAPKKKKKRPSPKPERPPSPEVIVDSEEEREDVALQMVG
FSNPPVLIKHGKGGKRTVRRLNEDDPVARGMRTQEEEEEPSEAESEITVMNPLSVPIVSA
WEKGMEAARALMDKYHVDNDLKANFKLLPDQVEALAAVCKTWLNEEHRGLQLTFTSNKTF
VTMMGRFLQAYLQSFAEVTYKHHEPTGCALWLHRCAEIEGELKCLHGSIMINKEHVIEMD
VTSENGQRALKEQSSKAKIVKNRWGRNVVQISNTDARCCVHDAACPANQFSGKSCGMFFS
EGAKAQVAFKQIKAFMQALYPNAQTGHGHLLMPLRCECNSKPGHAPFLGRQLPKLTPFAL
SNAEDLDADLISDKSVLASVHHPALIVFQCCNPVYRNSRAQGGGPNCDFKISAPDLLNAL
VMVRSLWSENFTELPRMVVPEFKWSTKHQYRNVSLPVAHSDARQNPFDF
```

Figure 49: Part of file disprot_fasta_v6_01.txt.
The file was downloaded from DisProt database (Vucetic et al., 2005); http://www.dabi.temple.edu/disprot/index.php) (4/02/2013) and includes information on disorder regions from several species.

```
# SignalP-4.1 gram- predictions
# name                   Cmax  pos  Ymax  pos  Smax  pos  Smean   D     ?  Dmaxcut    Networks-used
gi|16127995|ref|NP_414542.1|  0.106  21  0.200  11  0.641   1  0.366   0.278 N  0.570      SignalP-noTM
gi|16127996|ref|NP_414543.1|  0.142  15  0.149  15  0.207   4  0.161   0.155 N  0.570      SignalP-noTM
gi|16127997|ref|NP_414544.1|  0.223  41  0.226  12  0.544  10  0.502   0.356 N  0.570      SignalP-noTM
gi|16127998|ref|NP_414545.1|  0.119  19  0.104  19  0.110   3  0.093   0.099 N  0.570      SignalP-noTM
gi|16127999|ref|NP_414546.1|  0.841  24  0.894  24  0.984  12  0.949   0.920 Y  0.570      SignalP-noTM
gi|16128000|ref|NP_414547.1|  0.102  38  0.146  15  0.450  10  0.202   0.172 N  0.570      SignalP-noTM
gi|16128001|ref|NP_414548.1|  0.113  30  0.122  54  0.186  52  0.109   0.117 N  0.510      SignalP-TM
gi|16128002|ref|NP_414549.1|  0.122  33  0.122  33  0.176   1  0.121   0.121 N  0.570      SignalP-noTM
gi|16128003|ref|NP_414550.1|  0.123  18  0.113  22  0.149   1  0.100   0.107 N  0.570      SignalP-noTM
gi|16128004|ref|NP_414551.1|  0.122  34  0.166  11  0.339   1  0.283   0.209 N  0.510      SignalP-TM
gi|16128005|ref|NP_414552.1|  0.117  24  0.105  57  0.120  53  0.091   0.098 N  0.570      SignalP-noTM
```

Figure 50: SignalP output file taking as input the *E. coli* K12 proteins.

Table 43: Proteins from Papanstasiou et al., 2013 in which the *E. coli* K 12 UniProt accession numbers did not map with GIs that are available in U00096.ptt.

UniProt accession	Gene name	Existence in U00096.ptt	EcoGene description
P75684	yagP	no	Pseudogene
P77481	ycjV	no	Pseudogene
P31450	glvG	no	Pseudogene

```
d1gw0a2 1gw0    A:163-343       b.6.1.3 70624   cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=70624
d1gw0a3 1gw0    A:344-559       b.6.1.3 70625   cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=70625
d1gw0b1 1gw0    B:1-162 b.6.1.3 70626   cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=70626
d1gw0b2 1gw0    B:163-343       b.6.1.3 70627   cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=70627
d1gw0b3 1gw0    B:344-559       b.6.1.3 70628   cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=70628
d3fu9a1 3fu9    A:1-162 b.6.1.3 210143  cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=210143
d3fu9a2 3fu9    A:163-343       b.6.1.3 210144  cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=210144
d3fu9a3 3fu9    A:344-559       b.6.1.3 210145  cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=210145
d3fu9b1 3fu9    B:1-162 b.6.1.3 210146  cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=210146
d3fu9b2 3fu9    B:163-343       b.6.1.3 210147  cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=210147
d3fu9b3 3fu9    B:344-559       b.6.1.3 210148  cl=48724,cf=49502,sf=49503,fa=49550,dm=49557,sp=74873,px=210148
```

Figure 51: Part of parsable file from SCOP with domain coordinate annotation.

Table 44**:** GI numbers from the NC00093.ptt file that do not have a synonymous code in U00093.ptt file.

16131297	157783152	16128784
16129429	16128226	145698234
16130030	145698286	145698305
16128261	16129331	

Table 45: GI numbers in U00096.ptt that did not have a synonymous code in NC00096.ptt.

545778207
308199521
345297179
545778206

Table 46: List with the 46 TM chains identified with PISCES standalone program. (http://dunbrack.fccc.edu/Guoli/pisces_download.php#BLASTDB).

2kea	1q16c	2ksfa	4gbya	1kf6d	4djia	1pw4a	3fwla	4iffa
1kqfc	2nmra	3udca	1y8sa	3o7pa	2o9da	2oaua	1kqfb	2gfpa
3ze3a	2r6gg	4njna	3k07a	4dbla	2gifa	2k73a	2r6gf	ifx8a
4gd3a	4atva	1nekc	2ksda	1nekd	3qe7a	4iu8a	1kf6c	
1a91a	1kpka	2qfia	3dhwa	2wsxa	1b9ua	2y5ya	4gd3q	

```
ExTopoDBID      1
UNIPROT         O04714
DESCRIPTION     F11A17.17 protein
GENE_NAME       GCR1
ORGANISM        Arabidopsis thaliana (Mouse-ear cress);
NCBI_TAXID      3702
SUBCELL_LOC     plasma membrane
DB_REF          Pfam; PF05462
DB_REF          EMBL; U95142|U95143|AC007932|AK228479
DB_REF          InterPro; IPR017981|IPR000848
DB_REF          PROSITE; PS50261
DB_REF          PIR; F96522
TOPO_INFO       105:I; PMID:15155892; Method:gene fusion (Ub)
TOPO_INFO       271:I; PMID:15155892; Method:gene fusion (Ub)
TOPO_INFO       326:I; PMID:15155892; Method:gene fusion (Ub)
SIGNAL          No information available
SEQUENCE        326 AA; Fragment:N
                MSAVLTAGGGLTAGDRSIITAINTGASSLSFVGSAFIVLCYCLFKELRKFSFKLVFYLAL
                SDMLCSFFLIVGDPSKGFICYAQGYTTHFFCVASFLWTTTIAFTLHRTVVKHKTDVEDLE
                AMFHLYVWGTSLVVTVIRSFGNNHSHLGPWCWTQTGLKGKAVHFLTFYAPLWGAILYNGF
                TYFQVIRMLRNARRMAVGMSDRVDQFDNRAELKVLNRWGYYPLILIGSWAFGTINRIHDF
                IEPGHKIFWLSVLDVGTAALMGLFNSIAYGFNSSVRRAIHERLELFLPERLYRWLPSNFR
                PKNHLILHQQQQQRSEMVSLKTEDQQ
HMM-TM          Reliability score:0.924
                OOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMMMMMMMMMMMIIIIIIMMMMMMMMM
                MMMMMMMMMMOOOOOOOMMMMMMMMMMMMMMMMMMMMMMMIIIIIIIIIIIIIIIIIII
                IMMMMMMMMMMMMMMMMMMOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMMM
                MMMIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIMMMMMMMMMMMMMMMMMMMOO
                OOOOOOMMMMMMMMMMMMMMMMMIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
                IIIIIIIIIIIIIIIIIIIIIIII
//
```

Figure 52: Illustration of file format of ExToPoDB.flat file.
Downloaded from (http://bioinformatics.biol.uoa.gr/ExTopoDB/; (Tsaousis et al., 2010)).

Table 47: P-values from Fisher Exact Test with contingency table for disordered genes and existence of RCCs using the random dataset of 41 genes for non disordered identifiers.

Method	**P-value**
%MinMax (cu)	1
RiboTempo	0.3679
MSS	1

Table 48: Gene ontology enrichment analysis results with RCCs detected with RiboTempo filtered for p<0.01.
Bonferroni correction was to correct multiple testing p-values.

Data	Term	P-value	Num_annotations
C_between_2_3	plasma membrane	2.14311066547357e-08	273 of 694 in the list, versus 1104 of 4141 in the genome
C_between_2_3	integral component of membrane	2.00958273668402e-07	248 of 694 in the list, versus 1000 of 4141 in the genome
C_between_2_3	intrinsic component of membrane	3.87564696975815e-07	248 of 694 in the list, versus 1006 of 4141 in the genome
C_between_2_3	membrane part	1.71382553107749e-06	252 of 694 in the list, versus 1040 of 4141 in the genome
C_between_2_3	cell periphery	6.14721076282356e-05	293 of 694 in the list, versus 1287 of 4141 in the genome
F_between_2_3	phosphorelay sensor kinase activity	1.33526623341393e-06	22 of 694 in the list, versus 34 of 4141 in the genome
F_between_2_3	protein histidine kinase activity	1.33526623341393e-06	22 of 694 in the list, versus 34 of 4141 in the genome
F_between_2_3	phosphotransferase activity, nitrogenous group as acceptor	2.31790345647959e-06	24 of 694 in the list, versus 40 of 4141 in the genome
F_between_2_3	signal transducer activity	7.05669795090002e-06	38 of 694 in the list, versus 85 of 4141 in the genome
F_between_2_3	signaling receptor activity	7.7545598427894e-06	23 of 694 in the list, versus 39 of 4141 in the genome
F_between_2_3	molecular transducer activity	2.23774624437778e-05	38 of 694 in the list, versus 88 of 4141 in the genome
F_between_2_3	ATP binding	3.61527890433384e-05	125 of 694 in the list, versus 446 of 4141 in the genome
F_between_2_3	adenyl ribonucleotide binding	4.8081725709482e-05	125 of 694 in the list, versus 448 of 4141 in the genome
F_between_2_3	adenyl nucleotide binding	7.82731819897238e-05	126 of 694 in the list, versus 456 of 4141 in the genome
F_between_2_3	purine ribonucleoside triphosphate binding	8.02836310243538e-05	133 of 694 in the list, versus 488 of 4141 in the genome
F_between_2_3	purine ribonucleoside binding	9.16768220859626e-05	133 of 694 in the list, versus 489 of 4141 in the genome
F_between_2_3	purine ribonucleotide binding	9.69871714248268e-05	134 of 694 in the list, versus 494 of 4141 in the genome
F_between_2_3	purine nucleoside binding	0.000104602	133 of 694 in the list, versus 490 of 4141 in the genome
F_between_2_3	ribonucleoside binding	0.000104602	133 of 694 in the list, versus 490 of 4141 in the genome
F_between_2_3	purine nucleotide binding	0.000150994	135 of 694 in the list, versus 502 of 4141 in the genome
F_between_2_3	nucleoside binding	0.000226945	133 of 694 in the list, versus 496 of 4141 in the genome
F_between_2_3	protein kinase activity	0.000273234	23 of 694 in the list, versus 45 of 4141 in the genome
F_between_2_3	anion binding	0.000393769	178 of 694 in the list, versus 713 of 4141 in the genome
F_no_RCC	structural constituent of ribosome	1.14999757139735e-17	56 of 1680 in the list, versus 56 of 4141 in the genome
F_no_RCC	structural molecule activity	3.88027043310777e-16	71 of 1680 in the list, versus 77 of 4141 in the genome
F_no_RCC	rRNA binding	9.26144114519854e-09	43 of 1680 in the list, versus 47 of 4141 in the genome

F_no_RCC	RNA binding	2.65385830226206e-08	112 of 1680 in the list, versus 160 of 4141 in the genome
P_start_100	single-organism cellular process	0.000141977	488 of 788 in the list, versus 2020 of 4141 in the genome
P_start_100	Transport	0.000570069	206 of 788 in the list, versus 743 of 4141 in the genome
C_no_RCC	Intracellular	1.2908829180807e-20	591 of 1680 in the list, versus 1024 of 4141 in the genome
C_no_RCC	intracellular part	2.1676806969923e-19	570 of 1680 in the list, versus 989 of 4141 in the genome
C_no_RCC	ribonucleoprotein complex	2.70502482404139e-18	60 of 1680 in the list, versus 61 of 4141 in the genome
C_no_RCC	Ribosome	1.35956186529051e-17	58 of 1680 in the list, versus 59 of 4141 in the genome
C_no_RCC	Cytoplasm	1.91204373765015e-17	522 of 1680 in the list, versus 904 of 4141 in the genome
C_no_RCC	cytosolic part	1.12404843383932e-15	56 of 1680 in the list, versus 58 of 4141 in the genome
C_no_RCC	cytosolic ribosome	1.69991061321921e-15	52 of 1680 in the list, versus 53 of 4141 in the genome
C_no_RCC	ribosomal subunit	1.69991061321921e-15	52 of 1680 in the list, versus 53 of 4141 in the genome
C_no_RCC	cytoplasmic part	1.64371689479613e-13	189 of 1680 in the list, versus 280 of 4141 in the genome
C_no_RCC	Cytosol	8.6680953236444e-10	161 of 1680 in the list, versus 245 of 4141 in the genome
C_no_RCC	organelle part	1.55042082163675e-09	72 of 1680 in the list, versus 91 of 4141 in the genome
C_no_RCC	large ribosomal subunit	1.62883057339307e-09	31 of 1680 in the list, versus 31 of 4141 in the genome
C_no_RCC	cytosolic large ribosomal subunit	1.62883057339307e-09	31 of 1680 in the list, versus 31 of 4141 in the genome
C_no_RCC	non-membrane-bounded organelle	2.15818533746871e-09	90 of 1680 in the list, versus 121 of 4141 in the genome
C_no_RCC	intracellular non-membrane-bounded organelle	2.25972442182411e-09	73 of 1680 in the list, versus 93 of 4141 in the genome
C_no_RCC	intracellular organelle part	2.84515466449351e-09	57 of 1680 in the list, versus 68 of 4141 in the genome
C_no_RCC	Organelle	1.23344728965074e-08	91 of 1680 in the list, versus 125 of 4141 in the genome
C_no_RCC	intracellular organelle	1.92141602270566e-08	74 of 1680 in the list, versus 97 of 4141 in the genome
C_no_RCC	macromolecular complex	2.04824086019386e-07	190 of 1680 in the list, versus 312 of 4141 in the genome
C_no_RCC	small ribosomal subunit	6.99019069337628e-05	21 of 1680 in the list, versus 22 of 4141 in the genome
C_no_RCC	cytosolic small ribosomal subunit	6.99019069337628e-05	21 of 1680 in the list, versus 22 of 4141 in the genome
C_start_80	plasma membrane	2.87933421886493e-08	254 of 638 in the list, versus 1104 of 4141 in the genome
C_start_80	cell periphery	9.02135102528147e-07	281 of 638 in the list, versus 1287 of 4141 in the genome
C_start_80	Membrane	7.41145898102118e-06	288 of 638 in the list, versus 1350 of 4141 in the genome
C_start_80	integral component of membrane	9.62375483524683e-06	224 of 638 in the list, versus 1000 of 4141 in the genome
C_start_80	membrane part	1.16708387318954e-05	231 of 638 in the list, versus 1040 of 4141 in the genome
C_start_80	intrinsic component of membrane	1.67284558396561e-05	224 of 638 in the list, versus 1006 of 4141 in the genome

P_between_2_3	signal transduction	7.15433153730498e-08	48 of 694 in the list, versus 105 of 4141 in the genome
P_between_2_3	Signaling	1.08238188005449e-07	48 of 694 in the list, versus 106 of 4141 in the genome
P_between_2_3	single organism signaling	1.08238188005449e-07	48 of 694 in the list, versus 106 of 4141 in the genome
P_between_2_3	phosphorelay signal transduction system	4.34621032436087e-07	39 of 694 in the list, versus 80 of 4141 in the genome
P_between_2_3	peptidyl-histidine phosphorylation	8.08569488878223e-07	23 of 694 in the list, versus 35 of 4141 in the genome
P_between_2_3	peptidyl-histidine modification	8.08569488878223e-07	23 of 694 in the list, versus 35 of 4141 in the genome
P_between_2_3	cell communication	1.06996102601049e-06	62 of 694 in the list, versus 162 of 4141 in the genome
P_between_2_3	signal transduction by phosphorylation	2.95802998119928e-06	22 of 694 in the list, versus 34 of 4141 in the genome
P_between_2_3	single-organism process	0.000108891	490 of 694 in the list, versus 2330 of 4141 in the genome
P_between_2_3	cellular response to stimulus	0.000197051	128 of 694 in the list, versus 466 of 4141 in the genome
P_between_2_3	single-organism cellular process	0.000380661	432 of 694 in the list, versus 2020 of 4141 in the genome
P_between_2_3	protein phosphorylation	0.000549385	25 of 694 in the list, versus 51 of 4141 in the genome
P_between_2_3	Transport	0.000969398	184 of 694 in the list, versus 743 of 4141 in the genome
C_start_100	plasma membrane	2.35553973602955e-09	308 of 788 in the list, versus 1104 of 4141 in the genome
C_start_100	integral component of membrane	3.62699288865021e-07	275 of 788 in the list, versus 1000 of 4141 in the genome
C_start_100	cell periphery	4.52444935050398e-07	339 of 788 in the list, versus 1287 of 4141 in the genome
C_start_100	intrinsic component of membrane	7.30299813767784e-07	275 of 788 in the list, versus 1006 of 4141 in the genome
C_start_100	membrane part	1.05365141022087e-06	282 of 788 in the list, versus 1040 of 4141 in the genome
C_start_100	Membrane	2.77691814948825e-06	349 of 788 in the list, versus 1350 of 4141 in the genome
P_start_60	single-organism cellular process	0.000143626	302 of 466 in the list, versus 2020 of 4141 in the genome
C_start_60	plasma membrane	4.75713854228445e-07	191 of 466 in the list, versus 1104 of 4141 in the genome
C_start_60	Membrane	7.11081180043872e-06	219 of 466 in the list, versus 1350 of 4141 in the genome
C_start_60	cell periphery	1.80144678061232e-05	209 of 466 in the list, versus 1287 of 4141 in the genome
C_start_60	integral component of membrane	0.000118666	167 of 466 in the list, versus 1000 of 4141 in the genome
C_start_60	membrane part	0.000147145	172 of 466 in the list, versus 1040 of 4141 in the genome
C_start_60	intrinsic component of membrane	0.000179876	167 of 466 in the list, versus 1006 of 4141 in the genome
P_no_RCC	Translation	1.60129601319406e-15	94 of 1680 in the list, versus 111 of 4141 in the genome
P_no_RCC	nucleoside monophosphate biosynthetic process	2.63381617469943e-05	41 of 1680 in the list, versus 49 of 4141 in the genome
P_no_RCC	ribose phosphate biosynthetic process	2.74984213192458e-05	48 of 1680 in the list, versus 60 of 4141 in the genome
P_no_RCC	ribonucleoside monophosphate biosynthetic process	0.000104029	37 of 1680 in the list, versus 44 of 4141 in the genome

P_no_RCC	Unannotated	0.000122126	199 of 1680 in the list, versus 341 of 4141 in the genome
P_no_RCC	ribonucleotide biosynthetic process	0.000163453	45 of 1680 in the list, versus 57 of 4141 in the genome
F_start_100	hydrolase activity	0.000534685	198 of 788 in the list, versus 715 of 4141 in the genome

Table 49: Gene ontology enrichment analysis results with RCCs detected with %MinMax (cu) filtered for p<0.01.
Bonferroni correction was to correct multiple testing p-values.

Data	Term	P-value	Num_annotations
P_start_end	signal transduction	0.000410218	30 of 420 in the list, versus 105 of 4141 in the genome
P_start_end	Signaling	0.000513743	30 of 420 in the list, versus 106 of 4141 in the genome
P_start_end	single organism signaling	0.000513743	30 of 420 in the list, versus 106 of 4141 in the genome
F_start_end	signal transducer activity	0.000228646	26 of 420 in the list, versus 85 of 4141 in the genome
F_start_end	molecular transducer activity	0.000480309	26 of 420 in the list, versus 88 of 4141 in the genome
F_start_end	porin activity	0.000784966	13 of 420 in the list, versus 28 of 4141 in the genome
F_start_end	wide pore channel activity	0.000784966	13 of 420 in the list, versus 28 of 4141 in the genome
F_no_RCC	structural constituent of ribosome	5.56175902482921e-09	45 of 1351 in the list, versus 56 of 4141 in the genome
F_no_RCC	structural molecule activity	1.96848828998567e-06	53 of 1351 in the list, versus 77 of 4141 in the genome
P_start_100	signal transduction	0.000236203	53 of 1017 in the list, versus 105 of 4141 in the genome
P_start_100	signaling	0.000351104	53 of 1017 in the list, versus 106 of 4141 in the genome
P_start_100	single organism signaling	0.000351104	53 of 1017 in the list, versus 106 of 4141 in the genome
C_no_RCC	intracellular part	7.77584199715446e-12	453 of 1351 in the list, versus 989 of 4141 in the genome
C_no_RCC	intracellular	1.10110450404268e-11	466 of 1351 in the list, versus 1024 of 4141 in the genome
C_no_RCC	cytoplasm	1.14711454593131e-10	415 of 1351 in the list, versus 904 of 4141 in the genome
C_no_RCC	ribonucleoprotein complex	8.34704720191129e-10	48 of 1351 in the list, versus 61 of 4141 in the genome
C_no_RCC	ribosome	4.15884138055891e-09	46 of 1351 in the list, versus 59 of 4141 in the genome
C_no_RCC	cytosolic part	5.51766525335649e-08	44 of 1351 in the list, versus 58 of 4141 in the genome
C_no_RCC	cytosolic ribosome	7.72001385049827e-08	41 of 1351 in the list, versus 53 of 4141 in the genome
C_no_RCC	ribosomal subunit	7.72001385049827e-08	41 of 1351 in the list, versus 53 of 4141 in the genome
C_no_RCC	organelle part	4.37652452803528e-07	60 of 1351 in the list, versus 91 of 4141 in the genome
C_no_RCC	organelle	9.04373399486398e-07	76 of 1351 in the list, versus 125 of 4141 in the genome
C_no_RCC	non-membrane-bounded organelle	9.95513913232091e-07	74 of 1351 in the list, versus 121 of 4141 in the genome
C_no_RCC	intracellular organelle part	2.31912897755348e-06	47 of 1351 in the list, versus 68 of 4141 in the genome
C_no_RCC	cytoplasmic part	3.05437475644993e-06	144 of 1351 in the list, versus 280 of 4141 in the genome

C_no_RCC	intracellular organelle	4.78351275873119e-06	61 of 1351 in the list, versus 97 of 4141 in the genome
C_no_RCC	intracellular non-membrane-bounded organelle	5.03242427689803e-06	59 of 1351 in the list, versus 93 of 4141 in the genome
C_no_RCC	cytosol	2.45850629416427e-05	126 of 1351 in the list, versus 245 of 4141 in the genome
C_start_end	cell envelope	0.000297789	47 of 420 in the list, versus 220 of 4141 in the genome
C_start_end	envelope	0.000339756	47 of 420 in the list, versus 221 of 4141 in the genome
C_start_end	external encapsulating structure	0.000568736	47 of 420 in the list, versus 225 of 4141 in the genome
F_end_100	sequence-specific DNA binding transcription factor activity	1.3228063523303e-05	80 of 848 in the list, versus 204 of 4141 in the genome
F_end_100	nucleic acid binding transcription factor activity	2.0070926464523e-05	81 of 848 in the list, versus 209 of 4141 in the genome
F_end_100	molecular transducer activity	0.000611848	40 of 848 in the list, versus 88 of 4141 in the genome
F_end_100	signal transducer activity	0.000620539	39 of 848 in the list, versus 85 of 4141 in the genome
P_end_100	signal transduction	0.00027083	47 of 848 in the list, versus 105 of 4141 in the genome
P_end_100	signaling	0.000383102	47 of 848 in the list, versus 106 of 4141 in the genome
P_end_100	single organism signaling	0.000383102	47 of 848 in the list, versus 106 of 4141 in the genome
P_no_RCC	translation	1.48742607545599e-05	69 of 1351 in the list, versus 111 of 4141 in the genome

Table 50: P-values form Fisher Exact Test for β-barrel TM/non TM and the existence of RCCs.

Method	**P-value**
%MinMax (cu)	0.737
RiboTempo	0.09
MSS	1

Table 51: Correlations regarding the presence of RCCs detected with %MinMax (cu) in loops of TM helices that interact or not.
TP are loops with no interacting helices and detected RCCs, FN are loops with no interacting helices and no detected RCCs, FP are loops with interacting helices and detected RCC and TN are loops with interacting TM helices with no detected RCCs.

DISTANCE	LOOPLENGTH	PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN
5.5	10	16.67	76.83	72.73	5	92.65	1	63	5	19
5.5	15	33.33	73.68	70.73	9.09	93.33	3	84	6	30
5.5	20	41.67	66.43	64.47	9.62	93	5	93	7	47
5.5	25	41.67	63.86	62.36	7.69	93.81	5	106	7	60
5.5	30	41.67	61.45	60.21	6.76	94.02	5	110	7	69

Table 52: Correlations of Table 51 divided to cytoplasmic and periplasmic loops.

Cytoplasmic									Periplasmic								
PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN	PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN
0	76.32	74.36	0	96.67	0	29	1	9	20	77.27	71.43	9.09	89.47	1	34	4	10
50	69.23	68.52	5.88	97.3	1	36	1	16	28.57	77.42	72.46	12.5	90.57	2	48	5	14
66.67	60	60.29	7.14	97.5	2	39	1	26	33.33	72	67.86	12.5	90	3	54	6	21
66.67	56.96	57.32	5.56	97.83	2	45	1	34	33.33	70.11	66.67	10.34	91.04	3	61	6	26
66.67	54.88	55.29	5.13	97.83	2	45	1	37	33.33	67.01	64.15	8.57	91.55	3	65	6	32

Table 53: Correlations regarding the presence of RCCs detected with RiboTempo in loops of TM helices that interact or not.
TP are loops with no interacting helices and detected RCCs, FN are loops with no interacting helices and no detected RCCs, FP are loops with interacting helices and detected RCC and TN are loops with interacting TM helices with no detected RCCs.

DISTANCE	LOOPLENGTH	PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN
5.5	10	22.22	80.7	72.73	15.38	86.79	2	46	7	11
5.5	15	23.08	75.68	67.82	14.29	84.85	3	56	10	18
5.5	20	26.67	67.39	61.68	11.76	84.93	4	62	11	30
5.5	25	29.41	65.35	60.17	12.5	84.62	5	66	12	35
5.5	30	29.41	62.96	58.4	11.11	85	5	68	12	40

Table 54: Correlations of Table 53 divided to cytoplasmic and periplasmic loop regions.

Cytoplasmic									Periplasmic								
PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN	PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN
0	84	77.78	0	91.3	0	21	2	4	28.57	78.12	69.23	22.22	83.33	2	25	5	7
25	74.19	68.57	11.11	88.46	1	23	3	8	22.22	76.74	67.31	16.67	82.5	2	33	7	10
25	60	56.82	5.88	88.89	1	24	3	16	27.27	73.08	65.08	17.65	82.61	3	38	8	14
40	56.52	54.9	9.09	89.66	2	26	3	20	25	72.73	64.18	16.67	81.63	3	40	9	15
40	54.17	52.83	8.33	89.66	2	26	3	22	25	70	62.5	14.29	82.35	3	42	9	18

Table 55: Correlations regarding the presence of RCCs detected with MSS (cl=7) in loops of TM helices that interact or not.
TP are loops with no interacting helices and detected RCCs, FN are loops with no interacting helices and no detected RCCs, FP are loops with interacting helices and detected RCC and TN are loops with interacting TM helices with no detected RCCs.

DISTANCE	LOOPLENGTH	PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN
5.5	10	37.5	76.71	72.84	15	91.8	3	56	5	17
5.5	15	33.33	72	68.81	9.68	92.31	3	72	6	28
5.5	20	35.71	64.75	61.76	10.42	89.77	5	79	9	43
5.5	25	40	62.5	59.62	13.56	87.63	8	85	12	51
5.5	30	43.48	60	57.74	14.71	87	10	87	13	58

Table 56: Correlations of Table 55 divided into cytoplasmic and periplasmic loops.

Cytoplasmic									Periplasmic								
PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN	PPV	NPV	ACC	SEN	SPE	TP	TN	FP	FN
0	72.73	70.6	0	96	0	24	1	9	42.86	80	74.47	27.27	88.89	3	32	4	8
0	66.67	65.2	0	96.77	0	30	1	15	37.5	76.36	71.43	18.75	89.36	3	42	5	13
50	58.18	57.6	8	94.12	2	32	2	23	30	70.15	64.94	13.04	87.04	3	47	7	20
57.14	56.25	56.3	12.5	92.31	4	36	3	28	30.77	68.06	62.35	14.81	84.48	4	49	9	23
57.14	53.73	54.1	11.43	92.31	4	36	3	31	37.5	65.38	60.64	18.18	83.61	6	51	10	27