Université de Montréal

Assessing the robustness of genetic codes and genomes

par
Miguel Sautié Castellanos

Département d’informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l’obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Juin, 2020

Université de Montréal
Faculté des arts et des sciences

Ce mémoire intitulé:

Assessing the robustness of genetic codes and genomes

présenté par:

Miguel Sautié Castellanos

a été évalué par un jury composé des personnes suivantes:

Miklós Csűrös, président-rapporteur
Nadia El-Mabrouk, directeur de recherche
François Major, membre du jury

Mémoire accepté le: . . . .

ABSTRACT

There are two main approaches to assess the robustness of genetic codes and coding sequences. The statistical approach is based on empirical estimates of probabilities computed from random samples of permutations representing assignments of amino acids to codons, whereas, the optimization-based approach relies on the optimization percentage frequently computed by using metaheuristics. We propose a method based on the first two moments of the distribution of robustness values for all possible genetic codes. Based on a polynomially solvable instance of the Quadratic Assignment Problem, we propose also an exact greedy algorithm to find the minimum value of the genome robustness. To reduce the number of operations for computing the scores and Cantelli's upper bound, we developed methods based on the genetic code neighborhood structure and pairwise comparisons between genetic codes, among others. For assessing the robustness of natural genetic codes and genomes, we have chosen 23 natural genetic codes, 235 amino acid properties, as well as 324 thermophilic and 418 non-thermophilic prokaryotes. Among our results, we found that although the standard genetic code is more robust than most genetic codes, some mitochondrial and nuclear genetic codes are more robust than the standard code at the third and first codon positions, respectively. We also observed that the synonymous codon usage tends to be highly optimized to buffer the impact of single-base changes, mainly, in thermophilic prokaryotes.

Keywords: Quadratic assignment problem, Cantelli's upper bound, Generalized linear mixed models, Genetic code, hydrophobicity, thermophiles, codon usage bias.

RÉSUMÉ

Deux approches principales existent pour évaluer la robustesse des codes génétiques et des séquences de codage. L'approche statistique est basée sur des estimations empiriques de probabilité calculées à partir d'échantillons aléatoires de permutations représentant les affectations d'acides aminés aux codons, alors que l'approche basée sur l'optimisation repose sur le pourcentage d'optimisation, généralement calculé en utilisant des métaheuristiques. Nous proposons une méthode basée sur les deux premiers moments de la distribution des valeurs de robustesse pour tous les codes génétiques possibles. En se basant sur une instance polynomiale du Problème d'Affectation Quadratique, nous proposons un algorithme vorace exact pour trouver la valeur minimale de la robustesse génomique. Pour réduire le nombre d'opérations de calcul des scores et de la borne supérieure de Cantelli, nous avons développé des méthodes basées sur la structure de voisinage du code génétique et sur la comparaison par paires des codes génétiques, entre autres. Pour calculer la robustesse des codes génétiques naturels et des génomes procaryotes, nous avons choisi 23 codes génétiques naturels, 235 propriétés d'acides aminés, ainsi que 324 procaryotes thermophiles et 418 procaryotes non thermophiles. Parmi nos résultats, nous avons constaté que bien que le code génétique standard soit plus robuste que la plupart des codes génétiques, certains codes génétiques mitochondriaux et nucléaires sont plus robustes que le code standard aux troisièmes et premières positions des codons, respectivement. Nous avons observé que l'utilisation des codons synonymes tend à être fortement optimisée pour amortir l'impact des changements d'une seule base, principalement chez les procaryotes thermophiles.

Mots-clés: Problème d'Affectation Quadratique, La borne supérieure de Cantelli, Modèles linéaires Généralisés mixtes, Code génétique, hydrophobicité, thermophiles, biais d'utilisation des codons.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PSEUDOCODES

LIST OF APPENDICES

DEDICATION

To my parents

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisor Nadia El-Mabrouk for her support, valuable guidance and patience.

I am also grateful to Emmanuel for his constructive comments at the initial phase of this project.

Special thanks also to my wife and daughter, Mari and Vero, for their supports and understanding.

CHAPTER 1

INTRODUCTION

At the main stages of the flow of genetic information from DNA to proteins, namely, DNA replication and repair, RNA transcription, RNA splicing, translation and post-translational modification, errors are frequent but only the mutations, which occur during DNA replication and repair, can be inherited[1]. The reduction of error and mutation rates and increase of the robustness are two important strategies that usually increase the fitness of an organism[2] [2, 3, 4, 5, 6]. The variations in error and mutation rates stems from genetic variations of some proteins directly involved in the DNA replication and repair systems [5].

The robustness is an intrinsic property of the proteins [7], regulatory gene networks [8, 9], protein interaction networks [8] and natural genetic codes [6, 11]. This property is based on one or two principles, the redundancy and the capacity to buffer the effect of errors [8, 9, 10, 12]. Both principles are present in the standard genetic code. The principle of redundancy can be seen in the fact that each one of most amino acids is encoded by more than one codon in the natural genetic codes. On the other hand, the other principle, that is, the capacity to mitigate the effect of errors, or mutations, is clearly reflected by two characteristics of the natural genetic codes: 1) Most of the codons that specify the same amino acid differ by one single-base change at the third codon position. 2) Similar amino acids are allocated to codons that differ by one single-base change. The first characteristic is due to the wobble codon-anticodon interaction and the resulting degeneracy of the genetic code [13]. The second characteristic, called here the load minimization property, has been considered by some authors as an evidence of evolutionary selection for minimizing the effect of errors and mutations [14,

[1] In DNA replication an incorrect nucleotide is incorporated only once in 10^8–10^{10} events, whereas in transcription and translation, the misincorporation rates are of 1 in 10^4 events and 1 in 10^3–10^4 events, respectively. [1]

[2] The fitness of an organism is the ability to survive and reproduce in a given environment.

15, 16]. However, other authors argue that this property could be an artifact of the evolution of the standard genetic code rather than a selectable feature[3] [17, 18, 19].

The load minimization property is computed as the mean change in a given amino acid property between all the codons that differ by one single-base substitution, henceforth called as mean phenotypic change. The mean phenotypic change is a measure of the robustness of a given genetic code. There are two main approaches to assess the robustness to errors, namely, the “statistical approach” and the “engineering” approach [20, 21]. Both approaches essentially use the same function of mean phenotypic change. The main difference between both approaches lies in the method of assessing the significance or the relevance of the mean phenotypic change values. The first one is based on the estimates of the probability of obtaining random codes as or more robust than a given natural code [14, 22][4]. The second approach is based on the use of optimization algorithms to find at least one code more robust than the standard code. This code is used to calculate the percent of optimization of natural genetic codes [18, 24].

Concerning the statistical approach, since the theoretical null distribution of the mean phenotypic change is unknown, an empirical null distribution is estimated by generating a random sample of codes [14, 25, 26]. The probability of obtaining codes as or more conservative than the natural genetic code is computed from this sampling distribution. For a genetic code representation based on codons, the number of all possible codes is of the order of 10^{89}, whereas for a genetic code representation based on codon blocks[5], the total number of codes is of the order of 10^{18}. As it is impossible to encompass the entire population of all possible codes, a comparatively much smaller sample of codes

[3] There are two types of robustness: 1) The extrinsic robustness which is usually the result of natural selection, for example, the heat shock response. 2) The intrinsic robustness which is probably the side-product of selection for another feature or the result of selection in high mutation rate regimes, for example, the scale free structures of metabolic networks and protein interaction networks, the load minimization property of the genetic code [17].

[4] Some authors used the scores calculated from empirical estimates of mean and standard deviation [22].

[5] Set of synonymous codons.

must be generated. On the other hand, since the unknown probability of generating codes more conservative than a given genetic code could be too small, the size of the random sample should be as large as possible to improve the accuracy of the estimates. Therefore, this method is inaccurate and expensive in terms of running time.

As for the “engineering” approach, the problem of quantifying the genetic code robustness to errors is formulated as an optimization problem. This problem, known as Load minimization problem, consists of finding a code with a minimum value of mean phenotypic change and is equivalent to the Quadratic Assignment Problem (QAP) which is NP-hard in its most general form [27]. To date most used algorithms to solve the Load minimization problem have been metaheuristics and local search algorithms [18, 28, 29, 30, 31]. Consequently, the reported estimates of optimization percent are sub-optimal. Multi-objective and population-based search algorithms have recently been applied to the Load Minimization Problem [32, 33]. Although these algorithms are useful to find more realistic solutions, they do not overcome the shortcomings related to suboptimality.

We focused on finding more efficient methods to assess the mean phenotypic change or robustness values of natural genetic codes and genomes. Concerning the statistical approach, we used the Cantelli’s upper bound and scores as measures of relevance instead of numerically estimated parameters like the probability of obtaining codes more robust than the natural genetic code. The Cantelli’s upper bound and scores were computed by using the equations for the mean and variance of the distribution of the robustness values corresponding to all possible genetic codes, assuming that the assignments of amino acids to codons are random. Since this method does not require generating a random sample of codes of large size, it represents an improvement, mainly, in terms of efficiency with respect to the previous method based on empirical parameter estimates.

We applied this method to compute the robustness of 23 natural genetic codes with respect to 235 amino acid properties. We used three genetic code representations for the whole genetic code and for each codon position and substitution type. To date, most of the studies have been limited to the codon-block representation of the genetic code (i.e. blocks of synonymous codons) and less than 20 amino acid properties [25, 26, 34, 35, 36].

The most relevant amino acid properties obtained by our analysis shed light on possible scenarios where a genetic code with the load minimization property could have emerged and evolved. Moreover, this analysis provides a comparative view of the relevance of each of the 235 properties of amino acids not only for the ability of the natural genetic codes to mitigate the impact of single-base changes but also for the adaptive evolution of proteins [38]. We corroborated that the hydrophobicity/polarity is the most relevant property [25, 39, 40]. More exactly, the most significant amino acid property scales were the Miyazawa's hydrophobicity, Polar requirement and Kyte's hydropathy index. Overall, we observed that there are other properties, besides the Hydrophobicity/polarity, linked to relevant values of genetic code robustness, namely, the average long-range contacts, Flexibility, Solvent Accessible surface area and transmembrane helix and small-linker propensities.

We observed that when considered the whole set of single base changes, the standard code turned out to be among the first three most robust codes of the 23 chosen natural genetic codes. However, the mitochondrial and nuclear genetic codes tend to be more robust than the canonical code at the third and first codon positions, respectively. These results indicate that natural codon reassignments increasing code robustness at these codon positions could be important factors in the recent evolution of the standard genetic code.

As for the "engineering" approach, we have found that the weight matrices of a genetic code representation used for computing the genomic robustness share some characteristics with the coefficient matrices of one instance of the Quadratic Assignment Problem with polynomial solution. Actually, only the weight matrix corresponding to the codon-based representation (i.e. synonymous codons not grouped into blocks) has these characteristics. In this case, we would apply a greedy algorithm to find the global minimum of the genomic robustness. For computing the optimization percentage, this value was used together with the mean of the genomic robustness values for all possible codes[6]. Therefore, the improvement with respect to previous methods is twofold: the estimates of optimization percentage are accurate and more efficiently computed.

We applied both approaches to evaluate the genomic robustness values of 324 thermophilic and 418 non-thermophilic prokaryotes according to 84 amino acid properties, three methods to process the stop codons and two weightings, one assigning equal weights to all single base changes (called unbiased weighting) and the other based on translation errors (also called, biased weighting). Both groups of prokaryotes were compared with respect to the scores and optimization percentage used as measures of relevance of genomic robustness values. Several three-level logistic mixed models were built including the scores, optimization percentage or principal components as fixed effects and the binary thermal categories as dependent variables. The most relevant amino acid properties and genetic code representations were chosen according to the probability values for fixed effects. Several studies have explored the relationship between codon usage and codon robustness by using the genomic robustness or, more precisely, the mean phenotypic change weighted with codon usage [35, 40, 41, 42, 43, 44, 45]. On the other hand, significant differences have been reported between thermophiles and mesophiles with respect to codon and amino acid usage [46, 47 ,48, 49]. We observed that the synonymous codon usage in prokaryotic genomes tend to be more robust for the weighting based on translation errors and amino acid hydrophobicity scales. We also detected that

[6] The traditional methods used empirical estimates of the mean based on generating random samples of codes [25, 26].

thermophilic prokaryotes are significantly more robust than non-thermophilic prokaryotes for average long-range contacts, mainly, at the first codon position [51, 52, 53]. These results agree with the selection for error minimization observed in coding sequences as well as with results indicating a significant influence of the mRNA stability and load minimization at the protein level on the synonymous codon usage divergence between thermophilic and mesophilic prokaryotes [41, 42, 43, 44, 50]. The codon block for the amino acid, Arginine, resulted to be the block most optimized to mitigate the effect of errors in thermophiles relative to non-thermophiles, among all codon blocks with at least two codons with different robustness (called here, heterogeneous codon blocks)[7].

We propose two methods to reduce the number of operations for computing the scores and Cantelli's upper bound, one based on applying some transformations to the equations for the first two moments of the distributions of robustness values and the other based on partitioning the graph representing the genetic codes into two components according to four different criteria, namely, a first robustness-based criterion, for which one component contains heterogeneous codon blocks and the other, homogeneous codon blocks[8], a second criterion based on pairwise comparisons between genetic codes, a third criterion based on the distinction between sense codons[9] and stop codons and a fourth one based on the distinction between synonymous and non-synonymous single-base changes .

The thesis is organized as follows. In chapter 2 after introducing the three main theories on the origin and evolution of the standard genetic code, we briefly describe some methods for assessing the robustness of genetic codes and genomes and their applications to test some hypotheses on the evolution of the codon usage bias as well as on the origin and evolution of the standard genetic code.

[7] We observed that for both weightings applied and for any amino acid index without repeated values, the set of codon blocks containing at least two codons with different robustness corresponds to the amino acids, Alanine, Glycine, Lysine, Arginine, Valine, Leucine, Serine, Isoleucine, Proline and Threonine. It is noteworthy that this set is equal to the known set of primitive amino acids except for the amino acids, Arginine and Lysine, that must be replaced for the Aspartic and Glutamic acids [54].

[8] The homogeneous codon blocks are sets of synonymous codons with the same robustness.

[9] The sense codon is a codon that codes for an amino acid.

In chapter 3, we review some elements of the Quadratic Assignment Problem. The problem of finding the most robust genetic code is an application of the QAP. We showed that the problem of finding the code for which the genomic robustness reaches a maximum value is equivalent to a known polynomial instance of QAP. On the other hand, the mean and variance of the distribution of robustness values for all possible assignments of amino acids to codons stem from the statistical applications of the Quadratic Assignment Problem. In chapter 4, we present the statistical and optimization-based methods that we propose to assess the robustness of genetic codes and genomes. We also describe some "shortcuts" to improve the efficiency of calculation and the statistical methods used in data analysis. In the chapters 5 and 6, we report and analyze the results of the application of our methods to compute the genetic code and genomic robustness. We set forth the amino acid properties according to which the genetic codes and genomes showed the most significant values of robustness. Several natural genetic codes are compared according to the whole and partial genetic code representations. Additionally, thermophilic and non-thermophilic genomes are compared according to several amino acid properties and different genetic code representations. The results are statistically analyzed and discussed. In chapter 7, we review the contributions of this project and suggest some possible lines of research to pursue in the future.

CHAPTER 2

ON THE ROBUSTNESS OF THE GENETIC CODES AND GENOMES

2.1 Early evolution of the standard genetic code

The standard genetic code is clearly structured such that similar amino acids tend to be assigned to codons differing by only a single nucleotide. This non-random arrangement of the genetic code results in its ability to mitigate the phenotypic impact of mutations or translation errors. Three main theories have been proposed to explain why the standard code has this property. These theories, namely, the Adaptive, Stereochemical and Coevolution theories, refer to the ancient evolution of the genetic code at the time of the last universal ancestor. In line with the Adaptive theory, this property is the result of selective forces acting on the ancestral genetic codes to minimize the effect of errors on the protein structure and function [14, 22, 55]. The Stereochemical theories states that this property is not a consequence of natural selection but rather due to the fact that the result that similar amino acids tend to bind to related codons or anticodons. As suggested by this theory, the interactions between ribozymes and amino acids used as cofactors gave rise to the initial genetic code in the context of the RNA-world [21, 39, 56, 57]. As for the coevolution theory, it suggests that the genetic code was expanded from an ancestral form containing a limited set of abiogenically synthetized amino acids, by incorporating the novel amino acids as biosynthesis pathways evolved. The codons specifying the precursor amino acids were reassigned to product amino acids synthetized from them in such a way that the effect of these replacements on protein structures tended to be minimized. According to this theory, the error minimization plays a subsidiary role because the evolutionary advantage that entails the new amino acid introduction outweighs its deleterious effect [59,60,61,62].

In general, there are two specific patterns in the genetic code structure that can be considered as good evidences for the adaptive and coevolution theories. The second codon position tends to group the amino acids with similar properties. More exactly, the codons that share the nucleotide U at second position specify hydrophobic amino acids whereas those that share the nucleotide A at this position specify hydrophilic amino acids. This pattern has been considered consistent with the adaptive theory. On the other hand, the codons starting with the same nucleotide correspond to amino acids that come from the same biosynthetic pathway [21,58]. More precisely, the amino acids of the shikimate, glutamate and aspartate families are encoded by codons starting with U, C and A, respectively. In addition, the codons with G at the first position correspond to the primitive amino acids [54]. This pattern agrees with the coevolution theory. Other alternative evolutionary pathways that could give rise to this structure of the standard genetic code have been put forward, such as, the 2-1-3 model of Massey [58, 64], the four columns theory of Higgs [65] and the ambiguity reduction model [66, 67].

Three main methodologies have been applied to test these theories: 1) The methods to quantify the robustness of genetic codes. (Adaptive and Coevolution theories), 2) The techniques of in vitro selection of RNA ligands, called aptamers, based on their binding strength to the amino acids. (stereo-chemical theory) and 3) Phylogenetic analysis of tRNAs and aminoacyl-tRNA synthetases. (Coevolution theories) [39, 66, 68].

2.2 Recent evolution of the standard genetic code

In addition to the standard code, more than 20 alternative genetic codes have been reported so far. These genetic codes, belonging to organelles and organisms with reduced genomes, have evolved from the standard code through a few codon reassignments. Three types of changes of the standard code are clearly visible, the reassignment of codons, the unassignment of codons and the introduction of new amino acids. Among the 23 alternative natural genetic codes, 10 involve reassignments of only sense codons and 8 also involve reassignments of stop codons [39, 69]. Two non-canonical amino acids have been shown to be encoded by the standard genetic code, namely, selenocysteine and pyrrolysine.

Two main theories explain the origin and evolution of the non-canonical genetic codes, the Codon capture and the Ambiguous intermediate theories. The Codon capture theory states that some codons vanish from genomes subjected to neutral mutational pressure resulting from errors of the DNA repair and replication systems. Later, these codons can eventually reappear by genetic drift. Then, a misreading non-cognate tRNA can capture this codon but reassigning it to a different amino acid. As for the Ambiguous intermediate theory, it states that in a duplicated tRNA gene, a mutation can occur that changes the anticodon identity or the specificity to aminoacyl-tRNA synthetase[10]. This results in an ambiguous translation of the considered codon. Two ambiguous translation types have been identified: 1) Ambiguity involving sense codons, according to which there is a competition between the wild-type cognate tRNA and non-cognate misreading tRNA or between two different amino-acyl tRNA synthetases charging the cognate tRNA. 2) Ambiguity involving a stop codon. In this case there is a competition between the release factor and the non-sense suppressor tRNA [69,70, 71, 72, 73].

Sengupta S and Higgs P. have introduced a unified gain-loss model of codon reassignment incorporating, as particular cases, the above-mentioned theories and two additional theories, namely, the unassigned codon and compensatory change [71, 72, 73].

[10] It has also been considered the possibility of mutations that affect genes involved in tRNA splicing and posttranscriptional base modifications as well as the genes for translational release factors.

2.3 Methods to quantify the genetic code robustness

The development of methods to quantify the robustness of the genetic codes have been essential to test different scenarios of the origin and evolution of the standard genetic code. These methods have also been useful to compare different natural genetic codes with respect to their robustness. These methods have three main components, namely, the computation of the mean phenotypic change, the procedures to assess the relevance of the mean phenotypic change values and representations of the genetic code. Different variations of such methods have been explored with the main objective of identifying the conditions under which the highest mean phenotypic change values are reached for the considered genetic code.

2.3.1 Exploring different functions of weighted mean phenotypic change

We point out below some of the most interesting results on the amino acid properties and weighting used in the mean phenotypic change. Several functions of mean phenotypic change, or robustness, of the standard genetic code based on different amino acid properties have been computed for comparison purposes. According to the statistical approach, the most relevant values were found for Polarity requirement [25, 29, 35, 62]. For this property, the proportion of codes more robust than the standard code was shown to be 2 in a sample of 10^4 randomly generated codes [25]. Later, by using a mean phenotypic change that include weights biased with respect to the codon position and substitution type, much smaller estimates of this proportion were obtained (only one code more robust than the standard code in a sample 10^6 randomly generated codes) [26]. Since this genetic code model including a more realistic weighting based on translation errors, led to most relevant values of robustness, the authors considered these results as a good evidence for the correctness of the adaptive theory [26]. In other words, the standard genetic code has been structured by selective forces in such a way that the effect of mistranslations is minimized. Furthermore, other authors have shown, by using a population genetic model of code-message coevolution, that the reduction of the effect of mutations could have also driven the evolution of the genetic codes [39]. Even more outstanding estimates of the genetic code robustness (of the order of 2×10^{-9}) were obtained by considering, first, a weighting that combines the substitution type and position with the amino acid frequency and second, a weight matrix whose elements are the effect in terms of folding free energy caused by amino acid substitutions [36].

Other research projects aimed at exploring other aspects of the genetic code optimality, have been considered using different weightings based on the relative frequencies of tRNAs gene copies [74], the codon usage [35, 40, 43, 44] as well as the two known classes of Aminoacyl-tRNA synthetases [75]. All these studies shed light on the relationship between the genetic code ability to minimize the effect of mistranslations and mutations and different factors like the amino acid and codon usage, the tRNA frequencies as well as the identities of the amino acids recognized by both Aminoacyl-tRNA.synthetases.

2.3.2 Evaluating different scenarios of genetic code evolution

The representation based on codon-blocks[11] has been the most frequently used representation. However, in order to validate certain genetic code evolution models, specific representations and randomization methods have been devised. We briefly explain some of these works below.

By using the approach based on optimization, Novozhilov and Koonin, have observed that a representation of the primordial genetic code based on 16 codon blocks shows high values of minimization percentage. These results are consistent with the expected low accuracy of the translation, DNA replication and repair systems at the initial phase of the genetic code evolution. In

[11] Each alternative code is generated by randomly allocating each of the amino acids to the codon blocks observed in the natural genetic code, while the stop codons remain invariant.

this scenario, the most robust codes must have represented a significant evolutionary advantage [76]. By using a similar approach based on optimization, Di Giulio, have validated different representations linked to the stages of genetic code expansion associated to the evolution of amino acid biosynthesis pathways [60, 62].

On the other hand, Massey stated that certain scenarios of the neutral evolution of primordial codes may be at the origin of the robustness of the standard genetic code structure. This study has provided evidence for the hypothesis that the addition of novel amino acids to the evolving codes derived from a process of duplication and divergence of Aminoacyl-tRNA synthetase and/or tRNA genes tends to favour the assignment of similar amino acids to similar codons. For their simulation experiments of code evolution, they used representations of genetic codes with different numbers of amino acids and codon block structures. The robustness was assessed by using the proportion of alternative genetic codes better than the standard code [63, 64].

As for Freeland and Hurst, they have shown that the pattern of biosynthetic connection between the amino acids encoded by codons starting with the same base does not explain the standard genetic code robustness. In this work, they used the known block-based representation of the standard genetic code and the statistical approach to assess the code robustness relevance. The amino acids were classified into four groups with respect to the base identity that their corresponding codons have at the first position. For estimating the proportion of codes more conservative than the natural code, two randomization schemes were applied, one for which the amino acid assignments were randomized without any restriction and the other for which the amino acid assignments were randomized under the restriction that each amino acid can only be reassigned within its corresponding group [14].

Buhrman et al. have incorporated in the method for assessing the robustness, the genetic code patterns described by the coevolution theory [14, 21] and some results of the aptamer experiments. Regarding the aptamer experiments, the assignments of seven amino acids were fixed on the basis of the enrichment observed for their codons in binding sites. The randomization of codon assignments was restricted to three sub-groups of biosynthetically connected amino acids [77].

2.3.3 The optimization algorithms used to assess genetic code robustness

For computing the optimization percentage, it is required to find the code with the minimum value of mean phenotypic change. Several metaheuristics have been applied for this purpose, such as, simulated annealing [28], genetic algorithms [20, 29, 30, 31, 78, 79], Great deluge [74, 81] and record-to-record travel algorithms [80]. The mean phenotypic change used as objective function in these studies, frequently included one or two amino acid properties, such as, polarity and volume. Overall, some authors have preferred to independently compute the robustness for each property of a previously chosen set of amino acid properties and compare the results [35, 43]. This approach based on meta-heuristic mono-objective optimization has been improved in two directions, by using multi-objective optimization algorithms or exact optimization algorithms.

As for the multi-objective optimization strategy, De Oliveira et al. [33] argued that the evolution of the genetic code to its present form can be more accurately described by a simultaneous optimization of two robustness functions, the first for polar requirement and the other, for hydropathy index or molecular volume. By applying a bi-objective genetic algorithm they have obtained codes with high values of optimization percentage. Other authors have applied multi-objective genetic algorithm with eight objective functions. They conclude that the genetic code is moderately optimized to mitigate the effect of errors [32].

Buhrman et al. [27] formulated the load minimization problem as a Quadratic assignment problem and solved it by using an exact branch and bound QAP-solver QAPBB [82]. They stated that the solution obtained by using the record-to-record travel algorithm was actually the global optimum for the block-based representation of the genetic code [80].

Other classical problems have been considered to study the genetic code structure, such as, the graph clustering problem [84] and the Traveling Salesman Problem (TSP) [83]. In particular, the load minimization problem was formulated as a TSP and solved by using a Hopfield neural network [83].

2.4 Robustness at the gene and genome level

The mean phenotypic change weighted with codon usage indicates the extent to which codon usage is optimized according to the genetic code structure. A significant association between codon usage and robustness to errors could indicate that natural selection for buffering the effect of errors could be an important factor in the evolution of coding sequences and genomes.

Zhu et al. have observed that, in *Escherichia coli*, the weighting based on codon position, transition/transversion bias and codon usage at the genome level decrease the error minimization. More exactly, they observed that the proportion of better codes is larger than those observed for Freeland and Hurst [35]. However, the genetic code turned out to be more robust according to codon usage preference in *Saccharomyces cerevisiae* [40]. Najafabadi et al. have demonstrated that introducing, as weights, the tRNA gene copy numbers into the mean phenotypic change used by Zhu et al.[12], increased the error minimization level estimated for the *Escherichia coli* genome [43]. In other work, an index, called error adaptation index, was defined from the mean phenotypic change to estimate the robustness to errors at the gene level. The authors showed that this index has significant

[12] The tRNA gene copy numbers is correlated with the tRNA abundance within the cell.

correlation with the codon adaptation index and mRNA levels [44]. These results suggest that the codon usage is selected in such way that the effect of errors/mutations on protein structure and function is minimized, mainly in highly expressed genes [43, 44]. Marquez et al., have suggested that there is not selection on codon usage for minimizing the effect of errors but rather for specific error minimization levels. They argued that the ability to modify the protein evolution rate by changing the usage of codons with different robustness could represent a selective advantage [45].

On the other hand, Archetti observed that in *Drosophila* and rodents, genes tend to prefer the most robust codons, that robustness is correlated with codon usage bias, and the error minimization is correlated with the rate of non-synonymous substitutions. He concluded that natural selection for minimizing the impact of errors at the protein level is an important factor in the evolution of coding sequences [42]. These results are consistent with several other findings, for example, it was observed that the frequency of codons more robust to translation errors tend to be higher in ligand-binding sites [85]. It was, also, detected that differences in synonymous codon usage between thermophiles and mesophiles are subject to constraints related to robustness to translation errors, indicating that the association between robustness and frequencies of synonymous codon usage is affected by the growth temperature range [50]. Moreover, in [86] the attenuation observed in the infection caused by reengineered poliovirus having several synonymous mutations in its capsid genes suggests, according to the authors, a link between the viral mutational robustness and synonymous codon usage.

Others authors have found by studying base changes in antibiotic resistance gene *TEM-1* β-lactamase and the fitness cost of substitutions in two influenza hemagglutinin inhibitor genes, that the standard genetic code structure is such that the deleterious impact of mutations is minimized, thus, increasing the probability of adaptive mutations [87]. More recently, a large-scale *in silico* mutagenesis experiment in which the changes in folding free energy produced by single amino acid replacements were computed for more than 20,000 protein structures suggested that codon usage is optimized for mitigating the errors at the protein level [88]. It was shown that, even, empirical mutational matrices are optimized to reduce the cost of amino acid replacements in bacterial protein-coding sequences [89]. Moreover, the ability to buffer the impact of errors has not only been considered for the synonymous codon usage but also for the amino acid usage. Hormoz observed that the natural amino acid composition leading to more stable proteins is also tuned for a higher robustness to errors, which is consistent with the association observed between thermostability and mutational robustness in thermophilic proteins [90, 91].

CHAPTER 3

A BRIEF INTRODUCTION ON THE QUADRATIC ASSIGNMENT PROBLEM

The Quadratic Assignment Problem (QAP) was first introduced in 1957 by Koopmans and Beckmann as a model for the facility location problems [92]. Since then, a wide spectrum of applications has been identified for the QAP in a wide variety of different areas such as wiring problems, statistical data analysis [93,94], microarray layout problems [95], scheduling, parallel and distributed computing, and graph alignment among others [96, 97, 98]. In 1976, Shani and Gonzalez have shown that this problem is NP-hard [99].

Consider a set, denoted by S_n, of permutations of the set $\{1,2 \ldots n\}$ and two nxn coefficient matrices $A = (a_{ij})$ and $B = (b_{ij})$. The Koopmans-Beckmann QAP is formulated as the problem of finding the permutation, $\pi_0 \epsilon S_n$, minimizing the following double sum,

$$\min_{\pi \epsilon S_n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{\pi(i)\pi(j)} b_{ij}$$

3.1 Solution methods

A variety of authors have addressed the problem with the objective of developing exact solution methods. The main strategies considered to find the global optimal solution are Dynamic programming, Cutting planes, Branch and Bound, and Branch and Cut algorithms [95, 96] To date, Branch and Bound algorithms have been the most frequently used optimization approaches to solve this problem. All the above exact algorithms are extremely inefficient, even, for relatively small instances ($N = 30$) [98, 100][13]. However, a great variety of real-world problems are formulated as

[13] For our codon-block based representation of the genetic code $N = 20$, because it does not include the termination codons. For the codon-based representation of the standard genetic code, $N = 64$ and without the termination codons, $N = 61$.

QAP instances of size much larger than N. Consequently, numerous heuristic methods have been developed for finding good suboptimal solutions in reasonable running times. Some of the main heuristic search strategies are the construction methods, limited enumeration methods, the improvement methods, the Tabu search algorithms, Simulated annealing, Genetic algorithms, Greedy Randomized Adaptive Search Procedure, Ant systems, Iterated local search, Neural networks and other methods [98, 101]. Some hybrid metaheuristics have also been developed like, for example, the method called ANGEL that combines three different strategies, namely, an ant colony optimization strategy with a genetic algorithm and a Local search method or a method that interleaves descent local search and genetic algorithms [103]. The performance of each of these heuristics depends on the problem characteristics [98].

3.2 Lower bounds

The QAP lower bounds have been extensively studied for two main reasons. First, they are an essential component of the Branch and Bound procedures. The research endeavours related to these procedures mostly focused on developing tight and computationally efficient lower bounds. Second, the lower bounds have been useful to verify the quality of the heuristic solutions. There are five main kinds of lower bounds [98, 101, 102, 103]: Gilmore-Lawler and related lower bounds, eigenvalue related lower bounds, reformulation-based bounds, the lower bounds based on LP relaxations as well as those based on semidefinitive relaxations. The Gilmore-Lawler lower bound, $LB,$ has been the most commonly used lower bound for the QAP [98, 104].

Below, we focus on Gilmore-Lawler lower bounds, but before we will introduce the proposition of Hardy, Littlewood and Polya [98]. This proposition is important for two reasons: 1) It is the basis for the definition of these lower bounds, 2) It is very useful to understand under which conditions the weighted mean phenotypic change reaches extremes values.

Consider two n dimensional vectors, *A* and *B,* which are sorted in increasing or decreasing order (denoted by the superscripts i, and d), and the inner product between them. The scalar product of two vectors, is defined by, $\langle A, B\rangle = \sum_{i=1}^{n} a_i b_i$

Proposition (Hardy, Littlewood, Polya) [98]

Let A and B be two n dimensional vectors. Then, the following inequalities hold for any permutation π ϵ S_n

$$\langle A^d, B^i\rangle \leq \langle A^\pi , B\rangle \leq \langle A^d, B^d\rangle$$

Returning to the definition of the Gilmore-Lawler bound, the entries λ_{ij} of a nxn matrix are calculated by sorting the rows, $a_{i,*}$ and $b_{j,*}$ in increasing and decreasing order, respectively,

$$\lambda_{ij} = \min_{\pi \epsilon S_n ,\pi(j)=i} \sum_{k=1}^{n} a_{i\pi(k)} b_{jk}$$

This implies from the proposition of Hardy, Littlewood and Polya that,

$$LB = \min_{\pi \epsilon S_n} \sum_{i=1}^{n} \lambda_{\pi(i)i} \leq \min_{\pi \epsilon S_n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{\pi(i)\pi(j)} b_{ij}$$

The *LB lower bound* is computed by solving the linear assignment problem shown on the left-hand side of the above inequality. Hence, LB is the lower bound for the optimal solution value of the QAP. *LB* can be calculated in $O(n^3)$ time by using the Hungarian algorithm [98, 104].

3.3 Instances of QAP solvable in polynomial time

There are several versions of the QAP whose coefficient matrices have some specific properties that make them solvable in polynomial time. These matrices can be classified according to these properties as Monge and anti-Monge matrices, Toeplitz and Circulant matrices, sum and product matrices, and graded matrices [98]. We have found that one of the matrices used to compute the genomic robustness is a sum matrix. Below, we will focus on this kind of matrices.

A matrix $A = (a_{ij})$ is called a sum matrix if each of its elements can be computed as follows, $a_{ij} = \alpha_i^r + \alpha_j^c$ for $1 \leq i.j \leq n$, where α_i^c and α_j^r are vectors of real numbers called column and row generating vectors respectively. A matrix is symmetric if $a_{ij} = a_{ji}$ and, skew symmetric if $a_{ij} = -a_{ji}$ for $1 \leq i.j \leq n$.

Theorem 1 (E. Cela): *If the matrix A is a sum matrix and the matrix B is an arbitrary matrix, then the QAP instance is solvable in $O(n^3)$ time, where n is the size of the problem.*

As was proven by Cela E. [98] this QAP instance can be transformed to a linear assignment problem as follows,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{\pi(i)\pi(j)}\, b_{ij} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_{\pi(i)}^r + \alpha_{\pi(j)}^c\right)b_{ij}\\ &= \sum_{i=1}^{n}\left(\alpha_{\pi(i)}^r \sum_{j=1}^{n} b_{ij}\right) + \sum_{j=1}^{n}\left(\alpha_{\pi(j)}^c \sum_{i=1}^{n} b_{ij}\right)\\ &= \sum_{j=1}^{n} \alpha_{\pi(j)}^r\, \beta_j^r + \sum_{j=1}^{n} \alpha_{\pi(j)}^c\, \beta_j^c\\ &= \sum_{i=1}^{n} d_{\pi(i)i} \quad \text{where, } \ d_{ij} = \alpha_i^r \beta_j^r + \alpha_i^c \beta_j^c\end{aligned}$$

Thus, in this case the QAP reduces to linear assignment problem, $\min_{\pi \epsilon S_n} \sum_{i=1}^{n} d_{\pi(i)i}$ which is solvable in $O(n^3)$ time. As shown by theorem 2, the QAP is solvable in $O(n^2)$ time if an additional requirement is fulfilled.

Theorem 2 (E. Cela): *If the matrix A is a sum matrix and at least one of the matrices is symmetric or skew symmetric, then the QAP is solvable in $O(n^2)$ time, where n is the size of the problem.*

As will be shown in section (Methods), the problem of finding the global minimum of the genomic robustness is solvable in $O(n^2)$ time, according to theorem 2.

Concerning the formulation of the QAP in terms of graphs, other polynomial solvable cases have been identified. If the matrices, *A* and *B*, represent the weight matrices of two isomorphic undirected graphs, the set over which the minimum is computed is the set of isomorphisms between both graphs and is a subset of S_n. Christofides and Gerrard have shown that if both graphs are isomorphic chains, cycles, wheels or trees, these QAP instances are polynomially solvable [98, 105, 106].

CHAPTER 4

METHODS

4.1 Research objectives and methods

In this section, we briefly present some terms that we will develop further in the chapter. We then set out our two main objectives and the methods applied to achieve them. We also give a general idea of some procedures that we propose to speed up the calculations.

Robustness is the ability to mitigate the effect of errors and is computed as the opposite of the mean phenotypic change. The robustness of the genetic code was already addressed in Chapter 1 and derives from the fact that similar amino acids are assigned to similar codons. The robustness of a codon depends on its corresponding amino acid and the amino acids encoded by the codons that differ from this codon in one single base. Thus, the robustness of a coding sequence is calculated directly from the relative frequency and robustness values of each of its codons. To assess the robustness of genetic codes and genomes, methods based on randomly generated permutations and metaheuristics have been employed. We propose accurate and efficient methods to compute the relevance values based on Castelli's upper bound (or scores) and optimization percentages.

Our first main objective is to test whether the robustness to errors is associated with the codon reassignments giving rise to the alternative genetic codes. We computed the significance scores and Cantelli's upper bounds for the robustness values of 23 genetic codes under 235 different amino acid indices. It is well known that hydrophobicity is behind the error-mitigation property of the standard genetic code, but the relationship of many other amino acid indices with this code property is unknown. We chose a wide variety of amino acid properties, some of them also closely linked to protein stability.

Since each amino acid index is the result of an approximate measurement or modelling, we select several indices per amino acid property. The Cantelli's upper bounds indicating the significance of code robustness values computed under these amino acid properties and for a given genetic code will be sorted in increasing order. The presence of clustering patterns clearly distinguishable of groups of indices representing the same property in these lists would reveal a more general picture of the natural genetic code robustness to single-base changes at the protein level.

It has been observed that there is no link between the increase in robustness to errors through codon reassignments and the evolution of several alternative genetic codes when the whole genetic structure is considered. Since the translation error rates differ by substitution position and type, it is therefore possible a partial strengthening of the ability of alternative genetic codes to buffer the effect of errors with respect to the codon positions or type of single base changes (transition or transversion).

The neighborhood structures of genetic codes will be also explored. The robustness, or the mean phenotypic change, is a function that assigns real values to a given codon (or codon blocks[14]) computed from the information on their neighboring codons[15] (or neighboring codon blocks[16]) and the corresponding amino acids. We define the neighborhood structure of the genetic code as a representation based on two subsets of groups of synonymous codons, one containing codons with equal robustness values and the other containing codons and sub-blocks[17] with different robustness values. This representation based on the code neighborhood structure will be employed with two purposes: reducing the number of codons and codon blocks to process for robustness computations and exploring amino acid-to-codon assignment patterns that could be biologically meaningful.

[14] The codon block is the set of synonymous codons.

[15] The codon differing by one single base substitution from a given codon is called neighboring or adjacent codon.

[16] The codon blocks differing by at least one single base substitution from a given codon block is called neighboring or adjacent codon block.

[17] The codon sub-block is a sub-set of synonymous codons with the same robustness values.

Our second main objective is to test whether the robustness to errors is associated with the synonymous codon usage in thermophilic and non-thermophilic prokaryotic genomes. We computed the optimization percentages and scores on 324 thermophilic and 418 non-thermophilic prokaryotes for 84 amino acid indices. With this application, we will be able to answer the following questions, *Is the robustness to single-base changes important for the evolution of synonymous codon usage in thermophilic and non-thermophilic prokaryotes? Does this ability is stronger for properties linked to protein stability and mistranslation-based weighting?* It has been shown that in certain genes, conserved protein sites and viral genomes, the most robust codons tend to occur more frequently. However, in prokaryotic genomes the results have been equivocal. Several factors have been shown to play an important role in the evolution of codon usage bias, the robustness and growth temperature range are two of them. On the other hand, significant differences have been observed at the level of amino acid and codon usage between thermophiles and mesophiles. However, only one study has been carried out so far to test the relationship among growth temperature range, synonymous codon usage and robustness. It would be interesting to explore this relationship using a much larger sample size and, besides, at the codon block level. As we are interested in testing the association between the synonymous codon usage and robustness, we will only consider the codon blocks containing codons with different robustness values, called here heterogeneous codon blocks.

In this chapter, we describe the statistical and optimization-based methods used to compute the relevance of robustness values for genetic codes and genomes. We also describe the three different genetic code representations applied in the computations, one based on codon blocks, and the other two, on sense codons and on the whole set of codons. We detailly explain the two methods to assign numerical values to stop codons[18] as well as both weightings applied, one based on the mistranslation rates, called here biased weighting, and the other based on equal weights and used for comparative purposes.

We perform the statistical method to solve the following two practical problems corresponding to the two above-mentioned objectives:

1) *Given p amino acid properties and three representations of q genetic codes, compute the robustness parameters for these genetic codes according to these properties, the two weightings and two methods to assign numerical values to stop codons.*
2) *Given p amino acid properties, the synonymous codon usages of d genomes, the three representations of a given genetic code, compute the robustness parameters for these genomes according to these properties, the two weightings and two methods to assign numerical values to stop codons.*

The second practical problem is also considered in the context of the optimization-based approach. The robustness parameters in the above two problems include, for the statistical approach, the mean phenotypic change for a given genome or genetic code, and the first two moments of the distribution of its values for all possible amino acid-to-codon (or codon block) assignments. These parameters are used to compute the Cantelli's upper bound and scores. For the optimization-based approach, the robustness parameters in the second problem comprises the mean phenotypic change weighted with the synonymous codon usage of a given genome, the first moment of the previously mentioned distribution as well as the minimum value of the mean phenotypic change. In this chapter, we propose efficient methods and algorithms to compute these parameters. These algorithms have the following

[18] We use three methods to process the stop codons, two of them assign numerical values to these codons according to different criteria and the other, entails ignoring the stop codons and the single-base changes from them to the other codons. This method based on neglecting the stop codons is equivalent to the representation based on sense codons.

general characteristics:1) They are based on partitioning the graph representing the genetic codes into two subgraphs, one represents an invariant component and the other, the component for which the values of the above parameters vary depending on the genetic code (first problem) or genome (second problem). We apply four main partitioning criteria, the first criterion is based on the code neighborhood structure (second problem), the second criterion is based on pairwise comparisons (first problem) of genetic codes and the other two criteria are based on the distinctions between the sense and stop codons, synonymous and non-synonymous single-base changes (both problems). 2) We propose equations for more efficient computations of the first two moments of the distribution of robustness values corresponding to all possible amino-acid-to-codon (or codon block) assignments. 3) Since the matrices for computing the mean phenotypic change weighted with the relative frequency of synonymous codons have the characteristics specific to a known instance of the Quadratic Assignment Problem with polynomial solution, an exact algorithm will be used to find the amino acid-to-codon assignment corresponding to the minimum value of this function.

4.2 Two methods for assessing the relevance of the mean phenotypic change

There are two related approaches to tackle the problem of assessing the relevance of the mean phenotypic change, namely, the approach based on optimization and the statistical approach. According to the first one, this problem is formulated as an optimization problem which will be called here as load minimization problem (LMP). The difference between both approaches, consists in the way we define the relevance or significance of the values taken by the function of weighted mean phenotypic change for a given genetic code and amino acid property. The mean phenotypic change allows us to quantify the ability of the genetic code to mitigate the effect of translation errors or mutations. It is not enough to compute the values of this function, we must have also a measure of how far these values are found from what is expected, assuming that the hypothesis of random assignment of amino acids to codons, or codon blocks, is true. That is, we need a measure of significance of these values. According to LMP, the optimization percentage is considered as a measure of the robustness relevance. As for the statistical approach, the proportion of random genetic codes more robust than a given genetic code is used to assess the significance of the robustness values computed for this genetic code. The statistical approach is adopted in all practical applications explored in this work. The LMP is only considered for computing the relevance of the genomic robustness.

4.2.1 Load minimization problem (LMP)

A given genetic code is represented in terms of two undirected graphs, G^g and G^p, where G^p represents the relationship between amino acids or translation-stop signals, while G^g represents the genetic relationship between codons or blocks of synonymous codons for the corresponding phenotypes. In other words, each vertex of G^p denoting a phenotype p will have a single corresponding vertex in G^g which represents one block or one of all of the codons for p. Conversely, each vertex of G^g has a single corresponding vertex in G^p. Hence, both graphs have the same number of vertices. More details follow.

Graph G^p:

The graph $G^p = (V^p, E^p, \omega)$ is defined as a weighted complete graph representing the distances between amino acids and translation-stop signals. More precisely, each vertex is assigned to an amino-acid or translation-stop signal (called phenotype), whereas a given phenotype may be assigned to more than one vertex. We denote by θ this labeling function $\theta: V^p \rightarrow \Sigma$, $V^p = \{1 \ldots n\}$ is a set of vertices representing amino acids and translation-stop signals, where Σ denotes the set of such elements.

The weight function, $\omega: E^p \to \mathbb{R}_{\geq 0}$, is the squared Euclidean distance between the amino acids or translation-stop signals, $\omega(i,j) = (p(i) - p(j))^2$, where $p: V^p \to \mathbb{R}$

Graph G^g:

$G^g = (V^g, E^g, \gamma)$ represents the adjacency structure of codons or blocks of synonymous codons in a given genetic code. More precisely, G^g is an undirected weighted graph representing the relation between codons or synonymous codon blocks. $V^g = \{1 \dots n\}$ is the set of vertices, each representing a codon or a block of synonymous codons. An edge is defined between two vertices, u and v, if and only if we can transform a codon of u into a codon of v by making a single base change. We also define a weight function, $\gamma: E^g \to \mathbb{R}_{\geq 0}$ representing the weight of a single-base change corresponding to an edge of E^g.

We denote by A^g and A^p, the weight matrices of the graphs G^g and G^p, respectively. More precisely, for each vertex pair, u, v of $\mathrm{V^p}$, $A^p(u,v) = \omega(u,v)$, and for each vertex pair, u, v of V^g, $A^g(u,v) = \gamma(u,v)$, if the edge (u,v) is defined, otherwise, $A^g(u,v) = 0$. The total number of single-base changes between the codons represented in the graph G^g is denoted by *N.*

Let π be a permutation, i.e. $\pi \in S_n$ where S_n is the set of all permutations of size n, n standing for the number of vertices of each graph (remember that both graphs have the same number of vertices). The matrix, $A_\pi^p = \omega(\pi(u), \pi(v))$, is obtained by permuting the rows and columns of A^p according to the permutation π. This permutation represents a mapping of phenotypes to codons or codon blocks defined from a given genetic code (figure 4.1). The function of mean phenotypic change according to a given permutation is defined in terms of the Frobenius inner product as follows,

$$F_\pi = \frac{1}{N}\langle A_\pi^p, A^g \rangle_F = \frac{1}{N}\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma(u,v)\omega\big(\pi(u),\pi(v)\big) \qquad (1)$$

$$F_\pi = \frac{1}{N}\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma(u,v)\big(p(\pi(u)) - p(\pi(v))\big)^2 \qquad (1a)$$

The term inside the double sum can be interpreted as the cost of simultaneously assigning the amino acid $\pi(u)$ to the codon, or block, u and the amino acid $\pi(v)$, to the codon, or block, v.

The load minimization problem can be stated as a Quadratic assignment problem, as follows:

$$\min_{\pi \in S_n} F_\pi \tag{1b}$$

Thus, this problem consists of finding a permutation, π_0, that minimizes the weighted mean phenotypic change, F_{π_0}. The robustness corresponding to the permutation π that represents a given genetic code is defined as,

$$R_\pi = -(F_\pi) \tag{2}$$

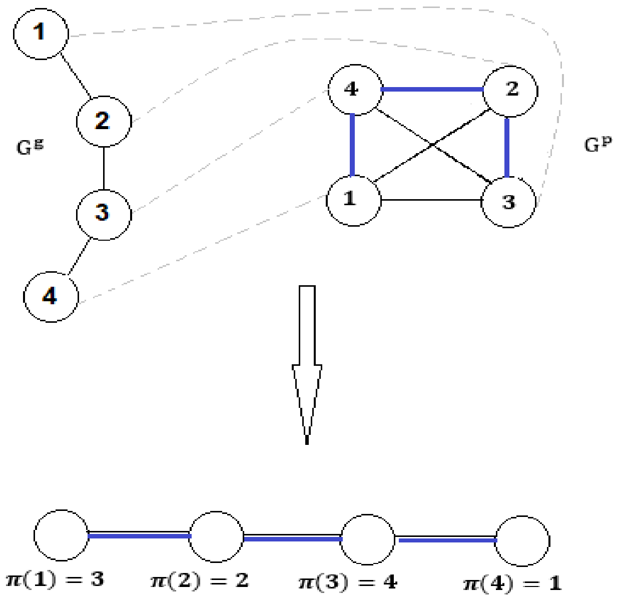
Since one term is the exact opposite of the other, both are equivalent. Smaller estimates of mean phenotypic change imply a greater robustness. Hence, if the robustness $R_\pi$is used instead of F_π, the LMP problem can be formulated as a maximization problem.

They are often used interchangeably in this work. Under this approach, the optimization percentage, *OP*[19], is considered a measure of the relevance of the value of mean phenotypic change for a given genetic code, F_{π_z}. It is defined in terms of the mean of the statistical distribution of F_π for all π in S_n, μ, and the value of an optimal assignment F_{π_0}, as follows,

$$OP = \left(\frac{\mu - F_{\pi_z}}{\mu - F_{\pi_0}}\right) 100 \qquad \text{if } F_{\pi_z} \leq \mu \tag{3}$$

$$OP = 0 \qquad \text{otherwise.}$$

[19] Other definition used in the literature: $OP = \left(\frac{F_{\pi_z} - F_{\pi_0}}{F_{\pi_m} - F_{\pi_0}}\right) 100$, where F_{π_m} is the maximum value of the mean phenotypic change and the other terms have the same definition explained in the text [31, 32].



$$F_\pi = \frac{1}{N}\left(\gamma_{12}\omega_{\pi(2)\pi(4)} + \gamma_{23}\omega_{\pi(1)\pi(2)} + \gamma_{34}\omega_{\pi(1)\pi(3)}\right)$$

Figure 4.1 Computing the mean phenotypic change F_π according to a given permutation π. Two graphs of order 4, $\mathrm{G^g} = (\mathrm{V^g}, \mathrm{E^g}, \gamma)$ and $\mathrm{G^p} = (\mathrm{V^p}, \mathrm{E^p}, \omega)$, that represent a hypothetical code of four codons specifying four amino acids. Dashed grey line: Bijective mapping between $\mathrm{V^p}$ and $\mathrm{V^g}$. Blue line: Edges of the complete graph $\mathrm{G^p}$ whose weights contribute to F_π after the bijective mapping. Bottom: The Mean Phenotypic Change according to the permutation π.

In previous studies, the parameterμ has been computed by randomly generating samples of genetic codes. This approach has two drawbacks. The first is its inaccuracy, as different samples may lead to different mean values. The second problem is its inefficiency, as improving accuracy requires increasing the size of the considered sample, which is time consuming. In this work, μ is accurately computed using an analytical equation for this parameter. As for F_{π_0}, rather than using suboptimal local search approaches as in former works, we will use an exact algorithm by taking advantage of some characteristics of the weight matrices specific to the genomic robustness context (explained later in the text). Such characteristics are related to known instances of the quadratic assignment problem that are solvable in polynomial time.

4.2.2 The statistical approach

While the mean phenotypic change is defined in the same way as before, a different approach is used for computing its significance. This approach consists of estimating the probability $P\left(R_\pi \geq R_{\pi_z}\right)$ of randomly generating a code more robust than a given genetic code, assuming the hypothesis of random arrangement. This probability can also be defined as the probability of obtaining a random code with mean phenotypic change value lower than that of a given genetic code. In other words:

$$P(F_{\pi} \leq F_{\pi_z}) = P(R_{\pi} \geq R_{\pi_z})$$

Since the shape of the distribution of robustness values corresponding to the whole population of genetic codes is unknown, this proportion has been estimated by using empirical sampling distributions. Rather, we use the known moments of the population distribution of the assignment costs for the quadratic assignment problem. On this basis, we propose a method that essentially relies on the knowledge of the equations for the mean, μ, and variance, σ^2, of the distribution of F_{π}, for all possible amino acid-to-codon assignments. Both moments can be accurately computed using an algorithm with a running time quadratic in number of codons or synonymous codon blocks. Even if the shape of the density function for this distribution is unknown, it is possible to estimate how far the robustness statistics is from the population mean, given a knowledge of its moments. In this sense, P, the probability of obtaining codes more conservative than a given genetic code was replaced by two other measures of significance of R_{π_z} or F_{π_z}*:* the Cantelli's bound (*UB*) which is just an upper bound on $P,$ defined as,

$$P(F_{\pi} \leq F_{\pi_z}) \leq UB$$

and the score S, defined as follows,

$$S = (F_{\pi_z} - \mathrm{\mu})/\sigma$$

This score indicates how many standard deviations a given robustness value is from the population mean. Thanks to these improvements, we will not need to use the random sampling of the huge population of all possible codes. In that way, we get rid of the most expensive part of the method based on the statistical approach. Even though the Cantelli's bound doesn't allow us to estimate, in exact terms, the significance of the robustness values, it is very useful to efficiently compute and compare the robustness relevance of several genomes or codes according to numerous amino acid properties. The analysis of the neighborhood structure of the genetic codes, aimed at detecting the codons and synonymous codon blocks that share a common amino acid neighborhood, has been effective to reduce the number of vertices and edges to process. Likewise, the set of vertices with the same amino acid assignment in genetic code representations based on codons, has been very useful to decrease the running time. These improvements among others have been included in the algorithms implemented for computing the genetic code and genome robustness.

4.3 Representations of the genetic codes

In this work, we use 3 representations $(G^g, G^p, \pi, \gamma, \omega)$ of the genetic code, two based on codons and one based on synonymous codon blocks. For each representation, we will consider either the whole genetic code representation, as its name implies, containing the whole edge set, E^g, or the partial genetic code representations formed by specific subsets of E^g defined by the type and codon position of the considered substitution. Recall that there is one-to-one correspondence between E^g and E^p defined by π. Every partition of E^g implies, therefore, a unique partition of E^p (see eq 1, 2).

The block-based representation has been used to compute the mean and variance for the set of all possible codes with the same degeneracy structure as the natural genetic code. The bock structure of the genetic code is a result of the interaction between the cognate tRNA and the codon in mRNA. The anticodon interacts through Watson-Crick base-pairing with the first two bases of the codon and can form wobble base pairs at the third codon position. The wobble base pairs at the third position allow a single cognate tRNA to read multiple codons, thereby, determining the degeneracy of the genetic code [13, 107]. The structure of synonymous codon blocks has the effect of greatly increasing the robustness. Thus, we can affirm that the genetic code robustness has two components, one well-known component caused by the codon-anticodon interaction and other component whose origin is unknown. For evaluating the predictions from different theories proposed to explain the origin of the unknown component, the block-based representation has been used to hold the degeneracy structure constant when exploring the space of all possible codes. In so doing, the estimates of the robustness relevance will be adjusted for the synonymous codon block structure of the genetic code.

In contrast, the codon-based representations allow us to obtain estimates of the robustness relevance unadjusted for the codon block structure of the genetic code. More exactly, these estimates depend on the two characteristics of the known natural genetic codes: the codon block structure and the characteristic according to which the similar amino acids tend to be encoded by codons differing by one single base change. The codon-based representation is used to compute the mean and variance for the set of all possible codes with the same number of codons for each amino acid as the natural genetic code. This constraint is much weaker than that based on the block-structure. Hence, the set of all possible codes generated under the codon-based model is much larger than that based on block-based model.

4.3.1 The codon-based representations

We first present the model based on the whole set of codons including, not only the sense codons but also the stop codons. For that reason, we use, in this case, the term phenotypes which encompasses the amino acids specified by the sense codons as well as the translation-stop signals. As defined above, the genetic code representation involves two graphs, the graph G^g representing the neighborhood structure of a given genetic code and G^p, the distances between the phenotypes. Since each vertex in the graph G^g, represents only one codon, in the other graph G^p, one amino acid can be represented by more than one vertex. The number of vertices representing the same amino acid depends on the number of codons for this amino acid in the considered genetic code.

According to this representation, both graphs have 64 vertices. The graph G^g has 288 edges that connect vertices whose codons differ by a single-base change. The graph G^p, which is a complete graph (K_{64}), has 2016 edges. Some of them, represent zero-valued distances either because of amino acids with the same property values or because of vertices denoting the same amino acid. It is worth noting that if two amino acids are assigned to codons that differ by more than one nucleotide, their distance is multiplied by zero in the equation for the mean phenotypic change.

It is worth noting that the graph G^g is the same regardless of the genetic code. In contrast, the graph G^p varies according to the number of codons representing each amino acid, which is specific to each genetic code. Likewise, the mapping of V^p to V^g represented by π is specific to each genetic code.

Another codon-based representation including only the sense codons is considered in this work. The standard code for example, which has 3 stop codons, is represented by a graph G^p of 61 vertices and 263 edges. The considered codon-based representation will be specified in the text, whenever used.

4.3.2 The representations based on synonymous codon blocks

In this model, the vertices of the graph G^g represent the synonymous codon blocks formed by sense codons. Hence, synonymous substitutions are ignored. If two codons from two synonymous codon blocks differ by one single-base change, then there will be a single edge between the vertices representing these two codon blocks. Notice that this single edge in block-based representation will be weighted by the sum of weights of all edges linking codons from both blocks (figure 4.2) This yields 20 vertices and 77 edges for the graph $G^{g\cdot}$ representing the standard genetic code. This genetic code representation only incorporates the missense single-base changes between the blocks formed by sense codons. In contrast to the codon-based models, in block-based models each amino acid amino acid is associated with a single vertex of V^p.

For the genetic codes containing ambiguous assignment rules that involve stop and sense codons, the property value of the corresponding amino acid is considered instead of the stop codons. This method to suppress this kind of ambiguity, results in two codes (Karyorelict nuclear code and Condylostoma nuclear code) with similar amino acid assignments.

4.3.3 Partial genetic code representations

In order to evaluate the contribution to the genetic code or genome robustness of each codon position and substitution type, different partial representations of the genetic codes will be considered.The edge set, E^g, of the graph G^g will be partitioned according to two criteria: the codon position affected by a substitution or the type of substitution (transition or transversion)[20]. These criteria give rise to 5 different partial representations.

The block- or codon-based models that include separately the single-base changes between one of the three codon positions, are actually disconnected graphs. The same is true for the graphs representing the transitions. The graph of the block-based model for the third codon position has 16 components and those for the other two positions have 4 components. The graphs of the codon-based models for each codon position have each one 16 components. The graphs, whose edges represent only the transitions, have 4 and 8 components in the models based on codon blocks and in the models based on the whole set of codons, respectively.

4.3.4 Block and codon neighborhood representations

For quantifying the independent contribution of each synonymous codon block, and codon, to the general genetic code or genome robustness, the representation of vertex neighborhoods based on

[20] A transversion is a substitution between a pyrimidine (U,C) and a purine (A,G) and a transition is a substitution between two pyrimidines or two purines.

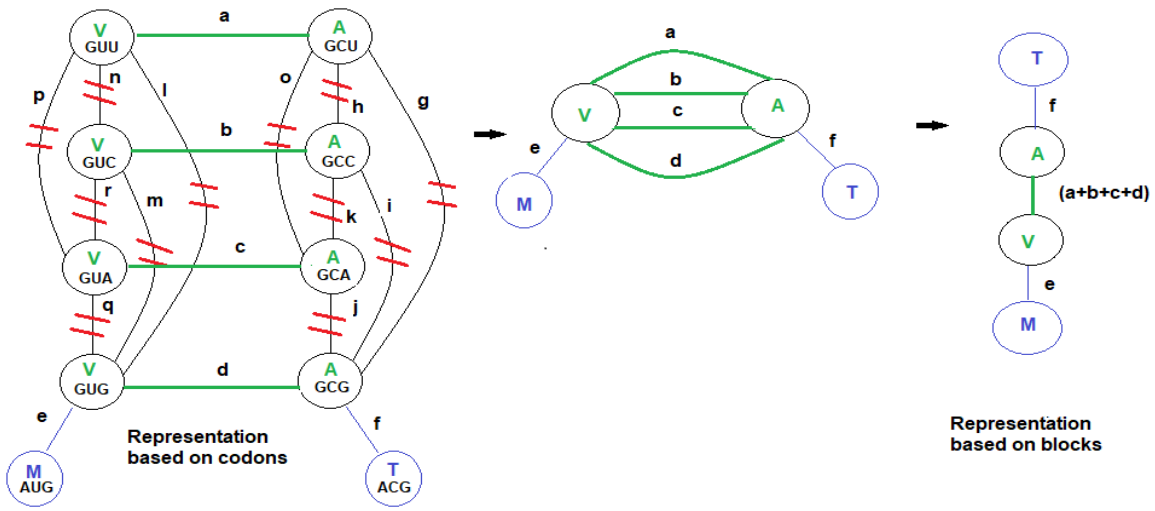


Figure 4.2 Relationship between the codon-based and the synonymous codon block representations. Red lines: Contraction of the edges between codons that specify the same amino acids. Green lines: edges connecting multiple synonymous codons of two adjacent blocks. All these edges except one, are deleted. The only edge that remains is weighted with the sum of parallel edge weights, (a+b+c+d). Blue lines: non-deleted edges whose weights stay the same.

disconnected graphs is also considered. Each disconnected graph has n connected components, where n is the number of vertices of the corresponding connected-graph representation. The connected components are star graphs. The star graphs represent the neighborhood of codons or synonymous codon blocks. According to the codon-based representation, in this disconnected graph, each codon,u, is represented by $c_u + 1$ vertices, where, c_u denotes the number of nodes adjacent to u. It follows that the disconnected graph has $((\sum_u^n c_u) + n)$ vertices and $2n$ edges (figure 4.3).

The contribution of every connected component representing the codons of the block ϱ and their neighboring codons,v, to the total mean phenotypic change,F_π, is denoted by F_π^ϱ. This contribution involves a given number of single-base changes, N_ϱ and is related to F_π as follows,

$$R_\pi = -F_\pi = -\left(\frac{1}{N}\sum_{\varrho=1}^{n} N_\varrho F_\pi^\varrho\right) \qquad (4)$$

According to the codon-based representation, F_π^ϱ is defined as a double sum,

$$F_\pi^\varrho = \frac{1}{\mathrm{N}_\varrho}\sum_{u=1}^{t}\sum_{v=1}^{c_u}\gamma_{uv}\,(p(\pi(u)) - p(\pi(v)))^2 \qquad (4a)$$

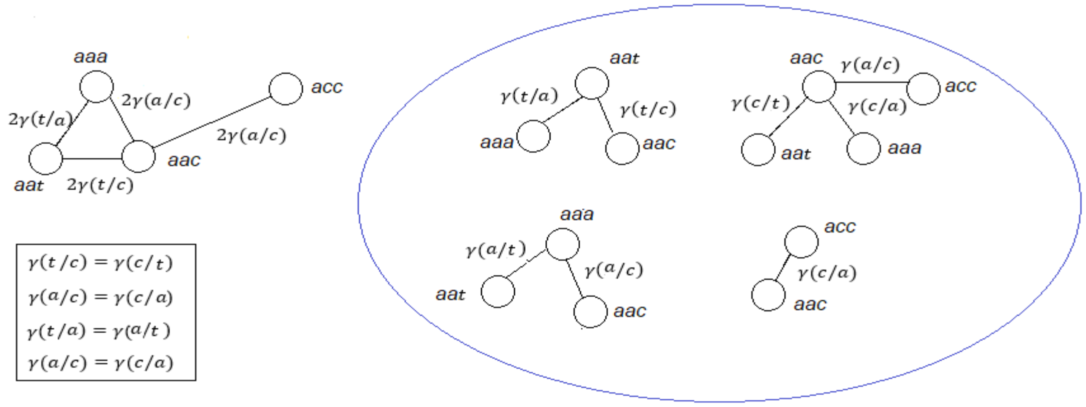


Figure 4.3 Set of 4-star graphs that represent the neighborhood of 4 codons. Only a subset of edges i represented. On each edge, the weight of the single-base change between the corresponding codons is shown

These equations are derived from the equations (1a) and (2). According to the block-based representation, F_{π}^{ϱ} is defined as a simple sum over the neighboring blocks of ϱ. Similarly, the independent contributions of codons in codon-based representations are defined as simple sums over their neighboring codons.

4.4 Weight functions

Four weights functions are considered in this work, two for computing the mean phenotypic change of genetic codes and the other two for the mean phenotypic change of the genomes. In this work, the term weight function refers to the function, $\gamma: E^{g} \rightarrow [0,1]$. Each of these functions is defined for two models, one based on codons and the other based on synonymous codon blocks.

4.4.1 Weights for genetic code robustness

Under the unbiased weighting in the representation based on codons, all edges have the same weight value equal to 1. Thus, considering that all these codons are grouped in the blocks represented by vertices in block-based models, this yield weights equal to the number of codons connecting two blocks, $n_{uv} \in \mathbb{N}_{>0}$. The definition of these weights assumes unbiased error frequency between the neighboring vertices u and v. (table 4.1)

We also considered a known mistranslation-based weighting, for which the weights are biased with respect to the type and position of the substitution[21] (table 4.2) [20, 26, 30, 35, 36, 37, 40, 43]. For the codon representations, h_{uv}, stands for the weight assigned to the edge linking codons u and v. For the block-based representation, h_i denotes the weight of the single-base change *i* of a total of n_{uv} single-base changes connecting synonymous codon blocks through the edge (u, v). (table 4.1) The weights, h_{uv} or h_i , take the set of values shown in table 4.2

Table 4.1 Weights for the mean phenotypic change of genetic codes according to codon and block-based models F_π^c , F_π^b

Weighting type	**Genetic code representations**	
	Codon-based model F_π^c	Block-based model F_π^b
Unbiased weights	$\gamma_{uv} = 1$	$\gamma_{uv} = n_{uv}$
Biased weights	$\gamma_{uv} = h_{uv}$	$\gamma_{uv} = \sum_{\mathrm{i}=1}^{\mathrm{n_{uv}}} h_i$

The vertices *u, v* represent either codons in codon based models or synonymous codon blocks in block based models.
Biased weights: Weights based on substitution type and codon position.
Unbiased weights: Binary weights.
γ_{uv} : Weight on the edge between the vertices *u* and *v*.
n_{uv} : Number of adjacent codons of the blocks *u* and *v*.
h_{uv}*:* Weights on the edges (*u*,*v*) based on substitution type and position, values shown in table 4.2
h_i : Weight of the single-base change i between codons belonging to adjacent blocks, u and v.

Table 4.2 Biased weighting values based on mistranslation rates (h_i, h_{uv})

Substitution types	**Codon positions**		
	Codon position 1	Codon position 2	Codon position 3
Transition	1	0.5	1
Transversion	0.5	0.1	1

[21] The biased weights reflect, mainly, the fact that the codon positions differ in the frequency of transitional versus transversional translation errors. Hence, the mean phenotypic change with this biased weighting is a measure of the optimization level of the genetic code, or genome, with respect to translation errors [20, 26, 30, 35, 36, 37, 40, 43].

4.4.2 Weights for genomic robustness

These functions are used to compare the groups of thermophile and non-thermophile genomes with respect to the robustness or mean phenotypic change at codon usage level. As for the codon-based model, γ_{uv} is defined from the genomic frequency f_u of codon u and the genomic frequency fs_u of the synonymous codon block s_u containing u, where,$\sum_u^t \frac{f_u}{fs_u} = 1$, t representing the number of synonymous codons belonging to codon block u. (table 4.3)

Table 4.3 Weights for the mean phenotypic change of genomes according to codon and block-based models F_π^c , F_π^b.

I

Weighting type	Genetic code representations	
	Codon-based Model F_π^c	Block-based model F_π^b
Unbiased weights	$\gamma_{uv} = \frac{f_u}{fs_u}$	$\gamma_{uv} = \frac{\sum_{i=1}^{n_{uv}} f_i}{fs_u}$
Biased weights	$\gamma_{uv} = \frac{h_{uv} f_u}{fs_u}$	$\gamma_{uv} = \frac{\sum_{i=1}^{n_{uv}} f_i\, h_i}{fs_u}$

The vertices *u, v* represent either codons in codon based models or synonymous codon blocks in block based models.
Biased weights: Double weights based on the synonymous codon frequency as well as the substitution types and position.
Unbiased weights: Weights based on synonymous codon frequency.
γ_{uv} : Weight on the edge between the vertices *u* and *v*.
n_{uv} : Number of adjacent codons of the blocks *u* and *v*.
h_{uv}: Weights on the edges (*u, v*) based on substitution type and position, values shown in table 4.2.
h_i : Weight of the single-base change i between codons belonging to adjacent blocks, u and v.
f_i, f_u: Frequency of codons *i* and *u*.
fs_u: Frequency of the synonymous codon block s_u containing *u*.

In the context of the representation based on the codon blocks, the weight f_i corresponds to the frequency of codon *i* from the codon block s_u. This codon *i* differs by one single-base change from a given codon that belongs to the synonymous codon block v. The weight of this base change is h_i. We should have,$0 \le f_u, \sum_i^{n_{uv}} f_i \le 1$, n_{uv} denoting the number of codons of the block u that are adjacent to codons of the block v (see table 4.3). For a synonymous codon block u formed by only one codon, the weight on the edge between this block and its neighbors, γ_{uv}, is equal to zero.

The reason of this requirement is to eliminate the maximal weight on blocks formed by only one codon. If the aim is to evaluate the effect on the genomic robustness of the synonymous codon usage bias, it does not make sense to include blocks of one codon (for example, in the standard code, AUG(MET) and UGG(TRP)). If the amino acids are mapped to codons, or synonymous codon blocks, in such a way that codons or blocks with the greatest mean phenotypic change, or the smallest robustness, are less frequent at the genome level and, on the other hand, those codons or blocks with the smallest mean phenotypic change, or the greatest robustness, are the most frequent ones, then the genomic robustness is maximized. (see proposition of Hardy, Littlewood and Polya proposition in Chapter 3 and Appendix I, figure I.1). In view of that, the genomic robustness can be considered as a measure of association between the robustness, or mean change in a given amino acid property, and the usage frequency of the synonymous codons at the genome level[22].

4.5 Methods to process stop codons

The capability of the genetic code and genomes to mitigate the effect of single-base changes is naturally measured for the sense codons by using the amino acid properties in graph G^p. But how to assess the effect of the stop codons?, Ignoring their effect is the simplest approach. This implies for the standard code the exclusion of 27 edges of E^g and E^p. These edges are part of the neighborhoods of vertices representing not only stop codons but also amino acids. As a consequence, the effect of

[22] If the relationship between both the synonymous codon robustness and usage frequency is monotonically increasing, the genomic robustness reaches a maximum value. Conversely, if this relationship is monotonically decreasing, the genomic robustness reaches a minimum value.

the corresponding single-base changes is overlooked and thereby, the contribution of some amino acids such as, tyrosine, arginine, serine, tryptophan among others, are strongly affected.

Another possibility is to arbitrarily assign large values to vertices or edges involving stop codons. The rationale behind this approach is the highly disruptive consequences of mutations resulting in premature stop codons. However, these over-valued weights for the translation-stop signals reduce the relative contribution of the other amino acids to the robustness relevance. On the other hand, this approach does not consider natural mechanisms to counter the negative effects provoked by premature stop codons.

Yet another method has been proposed which is based on the property value of the most probable amino acids inserted by nonsense suppressor tRNAs (NST) [27]. The NST resulting from naturally occurring mutations is able to read the premature stop codons and compete with the release factors for decoding them. Thus, these tRNAs introduce specific amino acids, thereby allowing the translation into proteins which, otherwise, would be truncated. This suggests that the robustness to nonsense errors or mutations is rather guaranteed by external factors like these tRNAs. Besides, it is known that the stop codons are reassigned to amino acids in some variant genetic codes. The method takes advantage of these facts to quantify the contribution of the stop codons to genetic code robustness.

We propose a method based on the “mean” suppressor tRNAs. Our method entails assigning to each stop codon, the mean of the property values corresponding to amino acids coded by all codons adjacent to it according to a given genetic code (Appendix I, table I.1). We consider also a method which consists of assigning to stop codons the mean of the amino acid property scale. Therefore, the edges of E^p representing the nonsense base changes will tend to be weighted with small values as happened with the mean suppressor tRNA method, but unlike the latter, this weighting does not depend on the amino acid encoded by the codons adjacent to stop codons.

Three methods to incorporate the contribution of the stop codons are considered in this work: One method just consists of ignoring the stop codons, the other two as explained above, are based either on the “mean” suppressor tRNAs or on the property-scale mean.

4.6 The neighborhood structure of the genetic codes

In this section we will explore some aspects of the neighborhood structure of the genetic codes. This will be very useful for identifying some biologically meaningful patterns and for reducing the mean phenotypic change computation running time.

The function of mean phenotypic change, or robustness, is essentially defined from the codon or block neighborhoods represented by star graphs (see section 4.3.4). For each star graph, the peripheral vertices represent the codons or blocks connected by single base changes to the considered codon, or block, which is in turn represented by the central vertex. Each star graph corresponds to a row in the weight matrix of the graph G^g. The robustness, or the mean phenotypic change, is a function that assigns a numerical value to a given codon, or codon block, represented by the central vertex of the star graph. Thus, the neighborhood structure of a genetic code is a representation based on a set of codon blocks formed by sub-blocks or codons with different values of robustness. There are two types of codon blocks, the codon blocks containing codons with equal robustness values and those containing codons with different robustness values. Below we will focus on some interesting regularities of the neighborhood structure in the context of the codon-based representation.

A set of codons is defined as a *homogeneous sub-block* if it complies with three requirements: 1) The codons of the set are synonymous, 2) The sets of amino acids encoded by the neighboring codons are equal, 3) The considered codons are connected to the adjacent codons specifying the same amino acid, through edges with equal weights.

The synonymous codon blocks whose all codons belong to the same *homogeneous sub-block*, are called *homogenous blocks*. The *homogenous blocks* are, hence, a particular type of *homogeneous sub-block*. In other terms, the homogeneous blocks are the blocks whose all codons have equal robustness. The synonymous codon blocks for which at least one subset of codons is a *homogeneous sub-block* are called heterogeneous synonymous codon blocks or *heterogeneous blocks*. In other terms, the heterogeneous blocks are the codon blocks containing at least two codons with different robustness. There are *heterogeneous blocks* that only contain different *homogeneous sub-blocks*, such as, those that code for Val, Ala in the standard code. There are other *heterogeneous blocks* formed by *homogeneous sub-blocks* as well as codons, like those that correspond, in the standard code, to leucine, arginine and serine.

The organization of amino acid assignments in the standard code (even in the variant genetic codes studied in this work) is such that the maximum size of the *homogenous sub-block or block* is 2. These two codons are always adjacent and differ by one transition in the third position. For the standard code and the other 24 alternative genetic codes studied in this work, the codons adjacent through the edges involving the first and second codon position always belong to neighboring *homogeneous sub-blocks* or *blocks* with C or U in the third position. This is a consequence of a general pattern according to which the codons with pyrimidine in the third position are organized in doublets that code for the same amino acid. Since all these *homogeneous sub-blocks* are doublets, only the reassignments that involves one of the codons of each doublet, breaks the structure of neighboring *homogeneous sub-blocks*.[23]

This neighborhood structure according to which all codons ending in pyrimidines and most of those ending in purines are grouped in neighboring *homogeneous sub-blocks* is seen in all these natural

[23] It is noteworthy that only the codons ending in purines are involved in the reassignments that give rise to the 24 natural genetic codes from the standard code, except for the Yeast mitochondrial code (ttable:3, NCBI).

genetic codes. Even in the Yeast mitochondrial code, its *homogeneous sub-blocks* (CUU, CUC) is found also in the standard code except that it specifies another amino acid (Thr). Thus, we could claim that almost all codons adjacent to codons of *homogeneous sub-blocks* belong to other *homogeneous sub-blocks*.

The codons belonging to a given *homogeneous sub-block* always have the same value of weighted mean phenotypic change or robustness. Conversely, codons with equal robustness values no necessarily belong to the same *homogeneous sub-block*. The robustness values do not only depend on edge weights of the star sub-graphs representing the codon neighborhoods and on how amino acids is assigned to these vertices but also on the values attached to them according to a given property. For example, if the methionine and isoleucine have the same property values according to a given hypothetical scale, the codons AUA and AUG will have the same robustness or mean phenotypic change (see an example in Appendix I, Figure I.2).

This is because these codons fail to fulfil the requirement of synonymy for forming a *homogeneous sub-block*. Then, if the property values of these two amino acids are equal, both could be considered as the same amino acid. Even the codons CUA and CUG that code for Leucine would seem to form a *homogeneous sub-block*, in this case, because both seemingly meet the second criterion of equal sets of neighboring amino acids. None of these codons actually forms a *homogeneous sub-block*. For that reason, a labelling function based on amino acid identity instead of the amino acid property is used to validate for each codon the fulfilment of the three requirements to form a *homogeneous sub-block* (Appendix I, pseudocode I.1).

We have developed an algorithm to determine the codons belonging to *homogeneous sub-blocks* (L) or *homogenous* blocks (B). Recall that A^g denotes the weight matrix of the graph G^g representing the codon neighborhood structure of the considered genetic code. *Θ* is an array of size 64 that represents the assignment of 20 amino acids to 64 codons (that is, the labelling function).

This algorithm allows us to verify whether each codon represented in A^g fulfils the three requirements for being part of a *homogeneous sub-block or block* (pseudocode 1, lines 4 and 9,10). If all codons specifying the same amino acid belong to a unique *homogeneous sub-block*, then we have *a homogeneous block* (Appendix I, pseudocode I.1, line 29).

The characterization of the neighborhood structure of the genetic codes allows us, as will be shown in the next section, to reduce the number of blocks or codons to process. For example, the standard genetic code can be represented, considering only the sense codons, as a set containing 10 homogeneous blocks and 10 heterogeneous blocks for both weighting used and regardless de amino acid property.

4.7 Methods to compute the mean phenotypic change

The three strategies described in this section aim to reduce the running time of mean phenotypic change computation. These methods rely on splitting the G^g vertices into two groups according to three criteria: 1) The types of codons represented by these vertices, such as, the sense and stop codons, 2) The types of blocks defined according to their neighborhood structures for a given genetic code, such as, the *homogeneous* and *heterogeneous blocks*, 3) The two groups of codons determined by pairwise comparisons between the standard code and any other natural genetic code with respect to their phenotypic assignments, such as, the group of codons assigned to the same amino acid in both genetic codes and the other group of codons with different amino acid assignments in both codes. All these methods are based on the partitioning of the graph representing the genetic codes into two subgraphs according to the above criteria. More precisely, the part of the graphs for which the mean phenotypic change is invariant for all different genetic codes, models or genomes, is separated from the part with a variable mean phenotypic change. The subgraph with a constant mean phenotypic change will be processed only once for each amino acid property, thus reducing the total running time. This improvement is possible, due to the additivity of the robustness or mean phenotypic change function. In other words, the robustness value of a graph representing a given genetic code is equal to the sum of robustness values of its sub-graphs. Moreover, the mean phenotypic change of the whole genetic code representation is obtained by adding the mean phenotypic change of the partial representations for each type and codon position of the single-base changes.

4.7.1 Computing the mean phenotypic change for different genetic code representations

It is possible to take advantage of the similarity among the genetic code representations with the aim of reducing the number of vertices and edges used for computing the values of the mean phenotypic change.

We denote by F_{π}^{c}, the mean phenotypic change under the codon-based representation of the genetic code based on the 64 codons, F_{π}^{stp}, that of the stop codons and F_{π}^{se}, that of the sense codons, where π is a given mapping between G^{g} and G^{p} vertices. The equations shown below are obtained by dividing into two groups the terms of the sum: one corresponding to the single-base changes between the stop codons and their adjacent codons (F_{π}^{stp}), whereas the other group includes only those terms that correspond to the single-base changes involving sense codons, (F_{π}^{se}). In terms of graphs, the underlying graph G^{g} is partitioned into two subgraphs, one for the stop codons and their adjacent codons and the other for the sense codons and their adjacent codons. The mean phenotypic change is computed separately for both. Let s represents a stop codon, v any of the n codons adjacent to it that could be a sense codon or another stop codon, ns the number of stop codons of a given genetic code and p the value of a given property assigned to the stop codon s or to its neighboring amino acid. Then, considering the additivity of the mean phenotypic change, we have,

$$F_\pi^c = F_\pi^{se} + F_\pi^{stp} \qquad (5)$$
$$= F_\pi^{se} + \frac{1}{N}\left(\sum_{s=1}^{ns} \sum_{v=1}^{9} \gamma_{sv} \left(p(\pi(s)) - p(\pi(v))\right)^2\right)$$

For the genetic codes studied in this work, $0 \leq ns \leq 4$.

Since the stop codons and their neighboring codons represent less than 10 percent of the genetic codes, computing F_π^c from F_π^{se} by adding the contribution of the vertices representing these codons, F_π^{stp}, rather than computing it from scratch, greatly reduce the number of vertices and edges to process (equation 5).

As for the mean phenotypic change defined under the block-based representation, F_π^b is equal to F_π^{se}. The contraction of the edges representing synonymous substitutions in the block-based representation (see figure 4.2) and the fact that several codons specify the same amino acid in a codon-based model, leads to the elimination of the influence of synonymous substitutions on the robustness values. It is worth noting that even though the robustness values are equal, the null population mean and variance computed under both representations for the same genetic code, will not necessarily be equal.

4.7.2 Considering the codon neighborhood structure for computing the genomic robustness

For computing the robustness of two set of genomes (thermophiles and non-thermophiles) according to the standard code and a set of previously chosen amino acid properties, the mean phenotypic change weighted with the genomic codon and block usage frequencies will be used. Before performing the inner product of two weight matrices, namely, the matrix formed by weights based on synonymous codon usage and the distance matrix, we perform the partition of the graph G^g into two subgraphs, one formed by the vertices representing the codons of the *homogeneous* blocks and the other, formed by codons belonging to *heterogeneous* blocks. This allows us to take advantage of the fact that the blocks formed by codons with the same robustness, have the same contribution to the total genomic robustness among different genomes. This can be proved as follows, for a given codon-based representation, the mean phenotypic change (F_π^ϑ) for each *homogeneous* block ϑ containing the codons q is defined by the equation, $F_\pi^\vartheta = \frac{1}{N_\vartheta}\sum_{q=1}^{y}\sum_{v=1}^{c_q}\frac{h_{qv}f_q}{f_{\vartheta q}}(p(\pi(q)) - p(\pi(v)))^2$. This equation is derived from the equation 4a by redefining it in terms of ϑ and substituting γ_{qv} with the expressions for the biased weightings (table 4.3). We denote by y the number of synonymous codons of ϑ , c_q, the number of vertices adjacent to q and N_ϑ, the number of single-base changes involving the codons of ϑ. As the terms that correspond to the contributions of all codons belonging to the considered *homogeneous* block are equal, they can be grouped together, as follows,

$$F_\pi^\vartheta = \frac{1}{N_\vartheta}\sum_{q=1}^{y}\frac{f_q}{f_{\vartheta q}}\sum_{v=1}^{c_q} h_{qv}\,(p(\pi(q)) - p(\pi(v)))^2 \qquad (6)$$

$$= \frac{1}{N_\vartheta}\left(\sum_{v=1}^{c_q} h_{qv}\,(p(\pi(q)) - p(\pi(v)))^2\right)\left(\sum_{q=1}^{y}\frac{f_q}{f_{\vartheta q}}\right),\ \text{Since by definition,}\left(\sum_{q=1}^{y}\frac{f_q}{f_{\vartheta q}}\right) = 1,$$

$$= \frac{1}{N_\vartheta}\left(\sum_{v=1}^{c_q} h_{qv}\left(p(\pi(q)) - p(\pi(v))\right)^2\right)\text{, where } q \text{ is any codon of } \vartheta \text{ (6a)}$$

Consequently, the mean phenotypic change for these blocks in particular does not depend on the synonymous codon usage. Thus, the contribution of the codons of all *homogeneous* blocks for the considered genetic code is previously computed for each amino acid property k, $S_\pi^q(k)$, thereby, excluding it from the computations of the mean phenotypic change for each genome (7). Then, inside the loop over the genomes, it will be only necessary to process the x codons from the heterogeneous blocks and their n neighboring codons for computing the mean phenotypic change per genome g and property, k, for the sense codons $F_\pi^{se}(k, g)$. (pseudocode 2)

$$S_{\pi}^{q}(k) = \sum_{q=1}^{z} \sum_{v=1}^{n} h_{qv} \left(p(\pi(q)) - p(\pi(v))\right)^{2} \quad (7a)$$

$$F_{\pi}^{se}(k,g) = \frac{1}{N}\left(S_{\pi}^{q}(k) + \left(\sum_{e=1}^{x} m(e,g) \sum_{v=1}^{n} h_{ev} \left(p(\pi(e)) - p(\pi(v))\right)^{2}\right)\right) \quad (7b)$$

Where $m(e,g)$ denotes the frequency of codon e of the heterogeneous block s_e in the genome g. Then, $(e,g) = \frac{f_e}{fs_e}$, where f_e is the frequency of the codon e and fs_e, the frequency of s_e. The weights on the edges between e (or q) and v, are denoted by h_{ev} (or h_{qv}).

Thus, this method is applied to compute the mean phenotypic change or robustness for several genomes and properties. Since the homogenous blocks represent roughly half of each genetic code, the decrease in the number of edges to process is considerable. Due to the fact that this algorithm (Appendix I, pseudocode I.2) is applied to one genetic code, the number of stop codons, as well as, the number of the *homogeneous* and *heterogeneous* blocks are constant. Therefore, the running time of this algorithm, as measured in terms of the number of genomes G and amino acid properties K, is $O(GK)$.

4.7.3 Computing the mean phenotypic change for different genetic codes

For computing the mean phenotypic change of a set of n genetic codes according to k amino acid properties, we previously make pairwise comparisons between two amino acid to codons assignments, one for the standard code represented by the permutation π and the other for a given alternative genetic code represented by the permutation β, with the aim of determining the sub-set of t codons assigned to different amino acids for each pair of these genetic codes,$\pi(t) \neq \beta(t)$. After this preprocessing step, for each property, k, the mean phenotypic change under the codon-based model is computed, $F_{\pi}^{co}(k)$. From the subgraph of G^g whose vertices represent the subset of codons t, the mean phenotypic change is computed for two subsets of amino acids assigned to these codons, one is defined from the standard code, $F_{\pi}^{t}(k,c)$ and the other, from the alternative genetic code, $F_{\beta}^{t}(k,c)$. Then, $F_{\beta}^{co}(k,c)$ is computed for the whole set of vertices and edges corresponding to β and each amino acid property k, as follows:

$$F_{\pi}^{t}(k,c) = \frac{1}{N}\left(\sum_{t=1}^{r}\sum_{v=1}^{n}\gamma_{tv}\left(p^{k,c}(\pi(t)) - p^{k,c}(\pi(v))\right)^{2}\right),\ t\colon \pi(t) \neq \beta(t) \qquad (8)$$

$$F_{\beta}^{t}(k,c) = \frac{1}{N}\left(\sum_{t=1}^{r}\sum_{v=1}^{n}\gamma_{tv}\left(p^{k,c}(\beta(t)) - p^{k,c}(\beta(v))\right)^{2}\right)$$

$$F_{\beta}^{co}(k,c) = F_{\pi}^{co}(k) - F_{\pi}^{t}(k,c) + F_{\beta}^{t}(k,c) \qquad (8a)$$

Notice that the number and identities of these codons and amino acids depend on the alternative genetic code c. The number of codons with different amino acid assignments, r, is between 1 and 6, for the genetic codes included in this work. On the other hand, the function that assigns numerical values to phenotypes, $p^{k,c}$, depends not only on the amino acid property k but also on the genetic code c, because of the mean suppressor method to quantify the contribution of the stop codons. This method allows us to greatly reduce the number of processed vertices per genetic code due to the small number of codon reassignments and stop codons in the genetic codes (line 8, Appendix I, pseudocode I.3). Considering that the number of codon reassignments and stop codons is constant, this algorithm has quadratic complexity in terms of the number of amino acid properties and genetic codes (Appendix I, pseudocode I.3).

4.8 Methods to assess the relevance of genome and genetic code robustness

The statistical method to assess the significance of the weighted mean phenotypic change had been previously based on the empirical sampling distribution of this measure. The weighted mean change, or robustness, of a given genetic code indicates the efficiency of this code for minimizing the effect of errors or mutations. The statistical approach is distinguished from that based on optimization by the way the strength or the relevance of the robustness values is measured. This approach is based on estimates of the proportion of random codes more efficient for minimizing the effect of errors than the considered genetic code. Empirical sampling distributions of the weighted mean phenotypic change are used to assess the relevance of these estimates, assuming that the hypothesis of random assignment is true. Since the shape of the null distribution is unknown, we propose to evaluate the significance by using two other criteria, namely, the Cantelli's upper bound and the scores. These parameters depend on the first two moments of the population distribution. Equations for these parameters are known in the context of the statistical approach to the Quadratic Assignment Problem[24] [93, 108]. The most important advantage of this method over the previous one, based on random sampling of codes, is its performance in terms of running time. In this section, we propose methods to compute these parameters for different genetic code representations, genetic codes and genomes.

4.8.1 Mean of the null population distribution

The mean of the population distribution of the mean phenotypic change is a parameter used in both approaches, one based on probability values estimated from the empirical sampling distribution of mean phenotypic change and the other, on the estimates of the optimization percentage. Previous works have mainly used random samples of codes to estimate the population mean of the mean phenotypic change. Since the Load minimization problem and the Quadratic assignment problem are essentially the same problem and the mean phenotypic change used in both approaches are similar to objective functions used in the context of the quadratic assignment problem, the equation for the population mean can be used with a few modifications. Below, the general equation used to compute the population mean of the mean phenotypic change, μ, depends on weight matrices of both graphs, G^g and G^p. The sum of weights, γ_{uv}, corresponding to single-base changes between the blocks, or

[24] We used essentially the same equations for the first two moments by virtue of the equivalence between the Quadratic assignment and Load minimization problems.

codons, u and v, (9a) is computed independently of the sum of weights of G^p (9b), where $p(i)$ represents the property values assigned to the phenotypes, i, j and N, the total number of single-base changes.

$$\mu = \frac{1}{n(n-1)N}\left(\sum_{u=1}^{n}\sum_{v=1}^{c_u}\gamma_{uv}\right)\left(\sum_{i}^{n}\sum_{j}^{n}(p(i)-p(j))^2\right) \tag{9}$$

$$T = \left(\sum_{u=1}^{n}\sum_{v=1}^{c_u}\gamma_{uv}\right) \tag{9a}$$

$$P = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(p(i)-p(j))^2\right) \tag{9b}$$

The mapping of the amino acids to codons represented by the permutation, π (where, $\pi \epsilon S_n$) is not required for the computation of μ. The equation for the mean in the context of the LMP is proved as follows,

$$\mu = \frac{1}{(N)n!}\sum_{\pi\epsilon S_n}^{n!}\left(\sum_{u=1}^{n}\left(\sum_{v=1}^{n}\gamma_{uv}(p(\pi(u))-p(\pi(v)\,))^2\right)\right) \tag{9c}$$

$$\mu = \frac{1}{(N)\mathrm{n}!}\ \sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv}\left(\sum_{\pi\epsilon S_n}^{n!}(p(\pi(u))-p(\pi(v)\,)^2\right) \tag{9d}$$

$$\mu = \frac{(n-2)!}{(N)\mathrm{n}!}\ \left(\sum_{u=1}^{n}\sum_{v=1}^{n}y_{uv}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(p(i)-p(j))^2\right). \tag{9e}$$

In the set of all possible mappings of n amino acids to n blocks or codons, each pair of vertices u and v, representing codons or blocks, will be obviously assigned to all possible permutations of two amino acids (equation 9d). Hence, each pair of u and v will be allocated to the same pair of amino acids in $(n-2)!$ permutations of S_n. In other words, every permutation of two amino acids assigned to u and v, will be repeated $(n-2)!$ times in the set, S_n. (equation 9e). Two strategies will be adopted to reduce the number of operations required to compute the null population mean under the codon-based models:

1) The block-based representations are built under the constraints imposed by the synonymous block structure. In contrast, the codon-based representations do not consider this block structure in G^g but preserves in G^p the number of each amino acid assigned to codons

instead. We take advantage of this characteristic by grouping in vectors n the right-hand terms of equation (9) that correspond to the same amino acids in graph G^p, thereby obtaining the equation (10a). More precisely, $n_{(c,i)}$, $n_{(c,j)}$ are the number of amino acids, i or j , encoded by the genetic code c and p^k, the values of the property k assigned to i and j.

$$P^{se}_{(c,k)} = \sum_{i=1}^{64-ns(c)} \sum_{j=1}^{64-ns(c)} \left(p^k(i) - p^k(j)\right)^2 \tag{10}$$

$$P^{se}_{(c,k)} = \sum_{i=1}^{20} \sum_{j=1}^{20} n_{(c,i)}\, n_{(c,j)} \left(p^k(i) - p^k(j)\right)^2 \tag{10a}$$

This grouping based on amino acid identity, reduces up to fourfold the number of operations performed on G^p weight matrices for the codon-based models.

2) The partitioning of both graphs, G^p and G^g, into two sub-graphs is performed, one sub-graph formed by the vertices representing sense codons (G^g) (or amino acids in G^p) and the other, formed by the stop codons (G^g) (or translation-stop signals in G^p). Firstly, the summation of G^p weights is performed for the model based on sense codons (eq 10a) then, the only vertices used to compute the right-handed term for the models based on the whole set of codons, $P^{co}_{(c,k)}$, are those representing the stop codons (eq 11a). This decreases more than tenfold the number of vertices and edges used to compute the null population mean for each genetic code or genome. The same idea is applied for the left-handed terms, $T^{se}_{(c)}$, of the equation of the null population mean for the representation based on sense codons (eq 11b).

$$P^{co}_{(c,k)} = \sum_{i=1}^{64} \sum_{j=1}^{64} \left(p^{k,c}(i) - p^{k,c}(j)\right)^2 \tag{11}$$

$$P^{co}_{(c,k)} = P^{se}_{(c,k)} + \sum_{s=1}^{ns(c)} \sum_{j=1}^{9} \left(p^{k,c}(s) - p^{k,c}(j)\right)^2 \tag{11a}$$

$$T^{se}_{(c)} = \left(\sum_{u=1}^{64-ns(c)} \sum_{v=1}^{64-ns(c)} \gamma_{uv}\right) = T^{co} - \sum_{s=1}^{ns(c)} \sum_{v=1}^{9} \gamma_{sv} \tag{11b}$$

Where, ns is the number of stop codons of the genetic code c and $p^{k.c}(s)$, the property values assigned to the stop codons, s, by using the mean-suppressor or scale-mean method according the genetic code c and amino acid property k.

As for the models based on the whole set of codons, the double sum of γ_{uv} can be formulated by grouping the weights with respect to the substitution types and positions as follows,

$$T^{co} = \sum_{u=1}^{64} \sum_{v=1}^{64} \gamma_{uv} \ , \ \ 192 \leq n_{pt}, n_l \leq 576, \ \ 0 \leq r_{pt}, r_l \leq 1 \qquad (12)$$

$$T^{co} = \sum_{p=1}^{3} \sum_{t=1}^{2} n_{pt}\, r_{pt} = \sum_{l=1}^{6} n_l\, r_l \text{ , grouping } p \text{ and } t \text{ as } l.$$

Where n_{pt} is the number of single-base changes for the codon position p and substitution type t. The weight of single-base changes according p and t is r_{pt}. The sum of weights for the models based on sense codons is not constant among the genetic codes, because the number and nature of sense codons (n_{sc}) is specific to each genetic code ($n_{sc} = 64 - ns(c)$).

The term, $T_{(c)}^{b}$, is defined as the sum of graph G^g weights for the codon-block model. Owing to the fact that these weights depend on the synonymous codon block structure, $T_{(c)}^{b}$ differs among genetic codes. Since the weights of the model based on codon blocks represent the contribution of missense single-base changes ($\theta(u) \neq \theta(v)$) between synonymous codon blocks, $T_{(c)}^{b}$ can be computed from those G^g weights corresponding to the missense substitutions in the representation based on sense codons, as follows:

$$T_{(c)}^{b} = \sum_{u=1}^{64-ns} \sum_{v=1\,,(\theta(u) \neq \theta(v))}^{64-ns} \gamma_{uv} = T_{(c)}^{se} - \sum_{u=1}^{64-ns(c)} \sum_{v=1,(\theta(u)=\theta(v))}^{64-ns(c)} \gamma_{uv}. \qquad (13)$$

To compute $T_{(c)}^{b}$ we therefore consider another partitioning criterion based on the distinction between synonymous and missense single-base changes. These methods can be applied to reduce the number of operations performed to solve two practical problems: 1) compute the null population means for a previously chosen set of genetic codes and amino acid properties (pseudocode 4), 2) compute the null population means for a previously chosen set of genomes and amino acid properties[25].

These algorithms to compute the null population mean of the genetic code (or genome) robustness values for a set of genetic codes (or genomes) and several amino acid properties have quadratic time complexity in terms of the number of genetic codes (or genomes) and amino acid properties (Appendix I, pseudocode I.4). The empirical sampling distribution of code robustness values had been previously used to compute estimates of this parameter. This method is clearly more expensive in terms of running time, on account of an additional loop to iterate over the random codes for calculating their robustness values.

4.8.2 Variance of the null population distribution

The variance is required to compute the Cantelli's upper bound and scores. Both parameters are considered as measures of significance or strength of the genetic code and genome robustness values. Regarding the statistical approach to the quadratic assignment problem, it is known the first two moments of the distribution of the cost of all possible assignments. The equation for the variance of this distribution was used with a few modifications inasmuch as the cost function has essentially the same form as the mean phenotypic change. Additionally, taking into account that the weight matrices of G^g and G^p are symmetrical this equation is expressed as follows,

$$\sigma^2 = \frac{1}{n(n-1)(N)^2}\left(\frac{D_4}{(n-2)(n-3)} + \frac{4D_3}{(n-2)} + 2D_2\right) - \mu^2 \qquad (14)$$

n: the number of vertices, N:Total number of single base changes.

[25] The pseudocode for this practical problem is very similar to the first one except for three differences : The weighting based on synonymous codon usage is used, the control-variables of the loops are replaced by the genome array index and $T_{(z)}^{co}$ is inside the loop that iterates over the genome array.

Each of the above terms,D_4, D_3 and D_2 are equal to the product of two terms which are independently computed from the weight matrices of the graphs G^g and G^p. Below, the equations for these terms are presented. Besides, the relationships between them are briefly explored in order to show some shortcuts for reducing the number of operations. The term D_2 represents the contribution of the assignments of each pair of amino acids to every pair of adjacent codons or blocks (eq 15-15b). There are $(n-2)!$ permutations with the same pair of amino acids assigned to each pair of codons or blocks, from which it follows that every specific mapping of two amino acids to two codons is contained in a permutation of size n (where, $n > 2$); these permutations represent a proportion of $\frac{1}{n(n-1)}$ of all permutations of size n (eq 14). (for more details see appendix VI)

$$D_2 = (T_2)(P_2) \tag{15}$$

$$T_2 = (\sum_{u=1}^{n} \sum_{v=1}^{n} \gamma_{uv}^2) \tag{15a}$$

$$P_2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} (p(i) - p(j))^4\right) \tag{15b}$$

The term D_3 represents the contributions to variance of the assignments of three amino acids to three codons or blocks represented by vertices of two adjacent edges of G^g (eq 16). There are $(n-3)!$ assignments of each three amino acids to every pair of adjacent edges. These assignments are represented by permutations of size 3 inside other permutations of size n (where, $n > 3$). Then, the permutations containing each of these permutations of size 3 represents a proportion of $\frac{1}{n(n-1)(n-2)}$ of the set of permutations of size n (eq. 14). (for more details see appendix VI)

$$D_3 = (T_3)(P_3) \tag{16}$$

$$T_3 = \left(\sum_{u=1}^{n}\left((\sum_{v=1}^{n} \gamma_{uv})^2 - \sum_{v=1}^{c_u} \gamma_{uv}^2\right)\right), \tag{16a}$$

$$P_3 = \left(\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{n}(p(i) - p(j))^2\right)^2 - \sum_{j=1}^{n}(p(i) - p(j))^4\right)\right) \tag{16b}$$

Both terms, T_3 and P_3, can be defined from T_2 and P_2, respectively, as follows,

$$T_3 = \sum_{u=1}^{n}(\sum_{v=1}^{n}\gamma_{uv})^2 - \sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv}^2$$

$$T_3 = \sum_{u=1}^{n}(\sum_{v=1}^{n}\gamma_{uv})^2 - T_2 \qquad (17)$$

And for the weight matrix of G^p:

$$P_3 = \sum_{u=1}^{n}\left(\sum_{v=1}^{n}\left(p(i)-p(j)\right)^2\right)^2 - \sum_{u=1}^{n}\sum_{v=1}^{n}\left(p(i)-p(j)\right)^4$$

$$P_3 = \sum_{u=1}^{n}\left(\sum_{v=1}^{n}\left(p(i)-p(j)\right)^2\right)^2 - P_2 \qquad (17a)$$

The term D_4 represents the contribution to variance of the assignments of four amino acids to four codons, or blocks, represented by vertices of two non-adjacent edges of G^g (eq 1). There are $(n-4)!$ assignments of each quartet of amino acids to every pair of non-adjacent edges. Hence, the permutations containing these patterns represents a proportion of $\frac{1}{n(n-1)(n-2)(n-3)}$ of the set of permutations of size n. This term is defined as follows,

$$\mathrm{D}_4 = (T_4)(P_4) \qquad (18)$$

$$\mathrm{T}_4 = ((\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv})^2 - 4\sum_{u=1}^{n}(\sum_{v=1}^{n}\gamma_{uv})^2 + 2\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv}^2) \qquad (18a)$$

$$\mathrm{P}_4 = \left(\left(\sum_{u=1}^{n}\sum_{v=1}^{n}\left(p(i)-p(j)\right)^2\right)^2 - 4\sum_{u=1}^{n}\left(\sum_{v=1}^{n}\left(p(i)-p(j)\right)^2\right)^2 + 2\sum_{u=1}^{n}\sum_{v=1}^{n}(p(i)-p(j))^4\right)$$

Both terms, T_4 and P_4, are defined by exclusion from T, T_3, T_2 and P, P_3, P_2 , respectively.

$$T_4 = (\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv})^2 - 4(T_3) + 2(T_2) \qquad (19)$$

$$P_4 = (\sum_{u=1}^{n}\sum_{v=1}^{n}(p(i)-p(j))^2)^2 - 4(P_3) + 2(P_2) \qquad (19a)$$

As proven below, for T_4,

$$T_4 = \left((\sum_{u=1}^{n}\sum_{v=1}^{\mathrm{n}}\gamma_{uv})^2 - 4\left[\sum_{u=1}^{n}\left((\sum_{v=1}^{n}\gamma_{uv})^2 - (\sum_{v=1}^{n}\gamma_{uv}^2)\right)\right] + 2\sum_{u=1}^{n}\sum_{v=1}^{\mathrm{n}}\gamma_{uv}^2\right)$$

$$T_4 = ((\sum_{u=1}^{n}\sum_{v=1}^{\mathrm{n}}\gamma_{uv})^2 - 4[(\sum_{u=1}^{n}(\sum_{v=1}^{n}\gamma_{uv})^2) - (\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv}^2)] + 2\sum_{u=1}^{n}\sum_{v=1}^{\mathrm{n}}\gamma_{uv}^2)$$

$$T_4 = ((\sum_{u=1}^{n}\sum_{v=1}^{\mathrm{n}}\gamma_{uv})^2 - 4\sum_{u=1}^{n}(\sum_{v=1}^{\mathrm{n}}\gamma_{uv})^2 + 2\sum_{u=1}^{n}\sum_{v=1}^{\mathrm{n}}\gamma_{uv}^2)$$

$$T_4 = (\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv})^2 - 4(T_3) + 2(T_2) \qquad \text{substituting with (15a) and (16a)}$$

$$T_4 = (T)^2 - 4(T_3) + 2(T_2) \qquad \text{substituting with (9a)}$$

The equation (19a) has a similar proof.

Three methods to reduce the number of operations will be adopted to compute the null population variance for several genetic codes and amino acid properties[26] :

1) Grouping the terms on the basis of the amino acid identity. The codon-based representation doesn't incorporate the synonymous codon block structure but only the number of each amino acid mapped to different codons according to a given genetic code. As in the procedure for computing the mean, we take advantage of this information to reduce roughly fourfold the number of operations per iteration by using an array $n_{(c,i)}$ which contains the number of the amino acid i (Appendix I, pseudocode I.5, lines 10-17) assigned to codons according to genetic code c.

2) As in the methods to compute the population mean and the mean phenotypic change, it will be performed the partitioning of the codon-based graphs into two sub-graphs, representing the stop codons and sense codons, as well as the synonymous and missense single-base changes (Appendix I, pseudocode I.5 lines, 13-16, 20-25). Thus, for computing the variance under the representation based on the whole set of codons, it would be only necessary to process the vertices and edges corresponding to the stop codons and their neighborhoods, which roughly represent much less than 5% of most genetic codes.

3) Using equations that we propose for computing the variance (equations 17, 17a, 19,19a, and the pseudocode 5 lines: 3, 4, 7, 8, 11, 12, 15, 17, 23, 25, 27 and 28.). It allows us to save operations by avoiding processing several times the same sub-graphs.

[26] The method for computing the variance of the null distribution of the mean phenotypic change corresponding to G genomes and K properties, is very similar to the algorithm described in the pseudocode 5. Three changes must be introduced: 1) Input: weights based on codon usage bias, 2) Loop over genomes instead of genetic codes, 3) The terms,T_1^{co} , T_2^{co} , T_3^{co} , T_4^{co} must be inside the loop over genomes.

This algorithm has time complexity of $O(CK|E^p|)$. Assuming that the number of vertices and edges are constant, time complexity would be, $O(CK)$, where *C* denotes the number of genetic codes and *K*, the number of amino acid properties. Since the number of edges processed is roughly constant among different genetic codes, this assumption seems reasonable (Appendix I, pseudocode I.5).

4.8.3 The Cantelli's upper bound and scores

The Cantelli's upper bound [109, 110] and scores are used to assess the strength or the relevance of mean phenotypic change estimates. For comparing genetic codes or genomes, these measures of relevance will be used. The greater the relevance, the farther these estimates are from expected values according to the considered genetic code representation and amino acid property.

In previous works, the method to assess the significance of mean phenotypic change values has been based on estimates of the proportion of more robust codes than a given genetic code. Since it is unknown the shape of the distribution of the mean phenotypic change under the assumption that the hypothesis of random assignment is true, an empirical null distribution is estimated by using random generation of code samples. Thus, with the aim of reducing the standard error of the estimates, the size of this random sample of codes has been increased as much as possible (for example, 10^9 codes [36]). Therefore, this method is inaccurate and too expensive in terms of running time. With the purpose of avoiding these issues, we have used the Cantelli's upper bound and the score as measures of relevance of robustness values. Both measures are useful to quantify how far the robustness values are from the null population mean. For previous methods, it was necessary to compute the mean phenotypic change values of the codes of a random sample to assess the relevance of this parameter estimates for a given genome or genetic code.

The Cantelli's upper bound and score, are defined from the mean and variance of the null population distribution. The Cantelli's upper bound is defined in the context of the one-sided Chebyshev inequalities, as follows,

Let be the mean phenotypic change, F_π, a random variable (Mean phenotypic change) with mean μ and variance σ^2. Then for any $a > 0$,

$$(F_\pi \geq \mu + a) \leq UB_1 = \frac{\sigma^2}{\sigma^2+a^2}$$

$$(F_\pi \leq \mu - a) \leq UB_2 = \frac{\sigma^2}{\sigma^2+a^2}$$

UB_1 denotes the upper bound on the probabilities of right-sided deviations from the mean,μ**.**

UB_2 denotes the upper bound on the probabilities of left-sided deviations from the mean, μ.

Below, the equation to compute the upper bound:

$$UB_1 = UB_2 = \frac{\sigma^2}{\sigma^2+(\mu-F_\pi)^2} \quad \text{, where } a = \mu - F_\pi \qquad (20)$$

The score is defined as,

$$Score = \frac{F_\pi-\mu}{\sigma} \qquad (21)$$

When it is important to know the direction of the deviations of F_π values from the mean, μ, the scores are used instead of the Cantelli's upper bounds.

4.8.4 Approach based on optimization

The optimization percentage is other known measure used to assess relevance of the mean phenotypic change values of amino acid-to-codon assignments. We will apply this measure in the context of the mean phenotypic change at the genome level. Two sets of genomes from thermophiles and non-thermophiles will be previously chosen to compare them according to the optimization percentage. In order to compute this measure, three parameters are required such as, the null population mean, μ, the mean phenotypic change for a given genetic code, F_{π_z} and the minimum value of the mean phenotypic change, F_{π_0} (eq. 3). This section deals with the algorithm to compute the latter parameter. Two weight functions will be used to compute the mean phenotypic change or robustness of a given genome (table 4.3). One weight function depends uniquely on the synonymous codon usage at the genome level and the other function, combines this parameter with a weighting scheme based on mistranslation rates.

The Load Minimization Problem in the context of its application to the genomic robustness, is a polynomially solvable version of the Quadratic Assignment Problem. This is a consequence of the fact that the weight matrix of G^g, whose elements are the frequency of synonymous codons, is a SUM matrix (Chapter 3, section 3.3), as will be shown below. The frequency of the synonymous codon,u, belonging to the block, s_u, is defined as, $m_u = \frac{f_u}{f s_u}$. A right-handed term, b_u , is defined by including the phenotypic distances represented in G^p, as well as, the weights of G^g , h_{uv}, based on the type and position of single-base changes. In this case, m_u is actually the row generating vector of the corresponding nxn matrix, where n stands for the number of vertices representing codons. More specifically, the entries of each row, u, of this matrix are constant and equal to m_u. Hence, we have a Sum matrix with a row generating vector and a column generating vector of zeros. Thus, the QAP instance that corresponds to the problem of minimizing the genomic robustness can be formulated as follows,

$$\min_{\pi \epsilon S_n} F_{\pi_0} = \min_{\pi \epsilon S_n} \frac{1}{N} \sum_{u=1}^{n} m_{\pi(u)} \sum_{v=1}^{n} h_{uv} \left(p^k(u) - p^k(v)\right)^2 \qquad (22a)$$

$$\min_{\pi \epsilon S_n} F_{\pi_0} = \min_{\pi \epsilon S_n} \frac{1}{N} \sum_{u=1}^{n} m_{\pi(u)} b_u \qquad (22b)$$

Additionally, since right-hand term in equation 22a is the Hadamard product of two square and symmetric matrices, the resulting nxn matrix will be also symmetric. Consequently, from the theorem

2 (see Chapter 3), we know that a QAP instance that can be reformulated as the inner product of a SUM matrix and a symmetric matrix is solvable in $O(n^2)$ time. As a result, the minimization problem (eq. 22b) is solved by sorting in opposite order the vectors $m_{\pi(u)}$ and b_u. Computing b_u, takes time $O(n^2)$ and sorting,$O(nlogn)$, consequently the algorithm runs in $O(n^2)$ time.

This algorithm is applied to the practical problem of computing the optimization percentage (eq. 3), used as measure of relevance of robustness values, for a set of previously chosen archaeal and bacterial genomes classified as thermophiles and non-thermophiles.

4.9 Information entropy and robustness of the synonymous codon blocks

The information entropy [111] and the standardized information entropy will be used as measures of synonymous codon usage bias. The information entropy, *IE*, is defined in terms of the genomic proportion of the codon,c of the block b, $p(b,c)$, the number of synonymous codon blocks, nb, and the number of codons of each block, t, as follows,

$$IE = \sum_{b=1}^{nb}\left(-\sum_{c=1}^{t} p(b,c)\log_2\left(p(b.c)\right)\right)$$

The standardized information entropy is defined as the information entropy divided by the maximum entropy per synonymous codon block ($\log_2 t$).

4.10 Statistical analysis and data

4.10.1 Three-level logistic mixed models

In order to select the amino acid indices that better discriminate between thermophiles and non-thermophiles with respect to the genomic robustness, several binomial random mixed models will be built, one for each amino acid property. For every genome and amino acid property, the scores corresponding to the mean phenotypic change will be computed and included as fixed effects in multilevel generalized mixed models with logit as link function, and the thermic status as binary response variable. More specifically, we will build three-level random intercept models [112, 113], using the three taxonomic ranks, such as, phylum, class and genus as grouping factors in a three-level nested design. The coefficients for fixed effects of these models are estimated by the maximum likelihood method based on Laplace approximation. The statistical significance of the fixed effects is determined by using the chi-square distributed likelihood ratio test statistic. For this test statistic, the same nested model will be used for all amino acid properties. The *p* values for the fixed effects will be used to choose the best amino acid properties with respect to different representations of genetic code.

4.10.2 Natural genetic codes and amino acid properties

It will be used 23 variant genetic codes from the web site: https://www.ncbi.nlm.nih.gov/Taxonomy /Utils/wprintgc.cgi. Most amino acid attributes will be obtained from the original sources. The references for the amino acid attributes are found in the appendix IX. A set of 235 amino acid property scales will be used for the application on the genetic code robustness (table IX.1 in appendix IX). We classify the amino acid property scales as general or local properties. We consider as local those property scales linked, for example, to specific secondary structures or transmembrane proteins. On the other hand, other property scales, such as, polarity, hydrophobicity, molecular weight, among others are classified as general properties. The general properties are intrinsic to the amino acids regardless of the protein context in which they are located. Thus, a set of 84 general amino acid properties was chosen for computing the genomic robustness (table IX.1 in appendix IX). We perform the standardization of all amino acid property scales by centering and scaling them by their respective mean and standard deviation.

4.10.3 Genomes

Two samples are chosen, one sample has 324 thermophilic and hyper-thermophilic bacteria and archaea (Growth temperature greater than 40-45^{0}C) and the other, 418 non-thermophilic bacteria and archaea (psychrophiles + mesophiles) (Table IX.3 in appendix IX). The taxonomic composition of the first sample is 23 phyla, 41 classes and 60 orders and that of the second sample is 17 phyla, 39 classes and 80 orders. The genomic codon usage will be downloaded from (Refseq) hive.biochemistry.gwu.edu/review/codon. To determine the growth temperature range group, we consider the following criteria, for hyperthermophiles: optimal growth temperatures (OGT) higher than 80^{o}C; for thermophiles: OGT between 45^{o}C and 80^{o}C; for mesophiles: OGT between 20^{o}C and 45^{o}C, and for psychrophiles: OGT lower than 20^{o}C. The thermophiles and hyperthermophiles are included in the group called thermophiles. Whereas the mesophiles and psychrophiles are grouped as the non-thermophiles.

4.10.4 Code and statistical software

All computations will be made with C++11 programs developed by us and compiled with g++. (https://github.com/Sautie/RGenomeGcode). The generalized linear mixed models are built and fitted by using the Lme4 package [114]. For principal component analyses, it will be used the packages, FactoMineR and Factoextra. For the two-sided Wilcoxon ranksum test and Benjamini–Hochberg procedure to control the false discovery rate, the R base packages will be used [115]. (https://www.R-project.org/).

CHAPTER 5

RESULTS

5.1 Overview on main results

In this section, we outline some of the most remarkable results that will be described in this Chapter and further analyzed in Chapter 6. We computed the mean phenotypic change of 23 natural genetic codes with respect to 235 amino acid indices. The Cantelli's upper bounds and scores were used as measures of relevance of the values of mean phenotypic change or robustness. We used two weighting schemes, three code representations and two methods to assign numerical values to stop codons in order to test under which of these conditions the robustness values are relevant. The standard genetic code showed the most relevant robustness values for hydrophobicity/polarity, the solvent accessible surface area, average long-range contacts, flexibility, Transmembrane helix and Small-linker propensities. We found that the 23 natural genetic codes tend to be more robust at the first and third codon positions for properties linked to protein stability. The standard genetic code was the most robust code for most of the above conditions. Moreover, some nuclear codes resulted to be more robust than the standard genetic code at the first codon position and most mitochondrial genetic codes showed to be more robust than the standard code at third codon position. These results suggest that increasing the robustness with respect to one of the above codon positions is an important factor in the codon reassignments that give rise to some alternative genetic codes.

We computed the robustness of 324 thermophilic and 418 non-thermophilic prokaryotes with respect to 84 amino acid indices. The Optimization percentages and scores were used as measures of relevance of the values of synonymous codon usage robustness. We observed significant values of synonymous codon usage robustness in prokaryotic genomes, indicating that the most robust codons tend to be more frequent at the expense of the least robust codons in these genomes. The synonymous codon usages of prokaryotic genomes tend to be much more robust for hydrophobicity and other properties linked to protein stability, specially, with respect to translational errors. We found that thermophilic prokaryotes are more robust than non-thermophilic prokaryotes, mainly, at the level of the first codon position and codon blocks corresponding to some of the most frequent amino acids in thermophilic and hyperthermophilic proteins, such as, R, K, P, V and L. It is known that these amino acids tend to play an important role in protein thermostability. We could consider these results as evidences of selection on synonymous codon usage for maximizing the robustness to errors, mainly, in prokaryotes living in high temperature environments. However, other selective factors and mutational bias might also be involved in the emergence of these general codon-choice patterns.

5.2 The robustness of natural genetic codes

5.2.1 Cantelli's bound and empirical estimates of robustness relevance

For comparing the weighted mean phenotypic change of genetic codes according to our 235 previously chosen properties, the Cantelli's upper bound was used instead of the more classical method based on generating random codes. As explained in the Methodology Section, the computation of the Cantelli's upper bound is much more efficient. Our objective is to verify that the Cantelli's upper bound produce, to a great extent, the same rankings for the 235 amino acid properties as the method based on empirical estimates of probability. For this purpose, we chose the codon-block based representation of the genetic code, as it is the most frequently used in the literature. Empirical null distributions of the weighted mean phenotypic change were computed from a random sample of codes. The mean phenotypic change was computed for each random code with the same codon-block structure as the standard genetic code. Then, for each amino acid property and codon position, we computed the proportion of codes with unbiased-weighted mean phenotypic change smaller than that of the natural genetic code (fig 5.1 and fig II.1-II.2, table IX.2 in appendices II and IX). For the subset of n amino acid properties with empirical estimates of the proportion smaller than 0.5 and including all codon positions, the spearman correlation coefficient between these empirical estimates and the Cantelli's bound was 0.95 (n=139). For the codon positions 1, 2 and 3, the Spearman correlation coefficients are 0.99(n=142), 0.91(n=84) and 0.87(n=199), respectively (table II.3 in appendix II). The scatter plot (fig II.1) clearly shows that for the third-codon-position model, there is a monotonically-increasing functional relationship between both logarithmically transformed parameters. This relationship is also visible above a given threshold in models including all codon positions or only the first position. Both analyses indicate that the rankings obtained by using both methods are essentially the same for the subset of the amino acid properties with estimates of the proportion of random codes smaller than 0.5. Thus, the association between both rankings is strongest among the amino acid properties for which the natural genetic

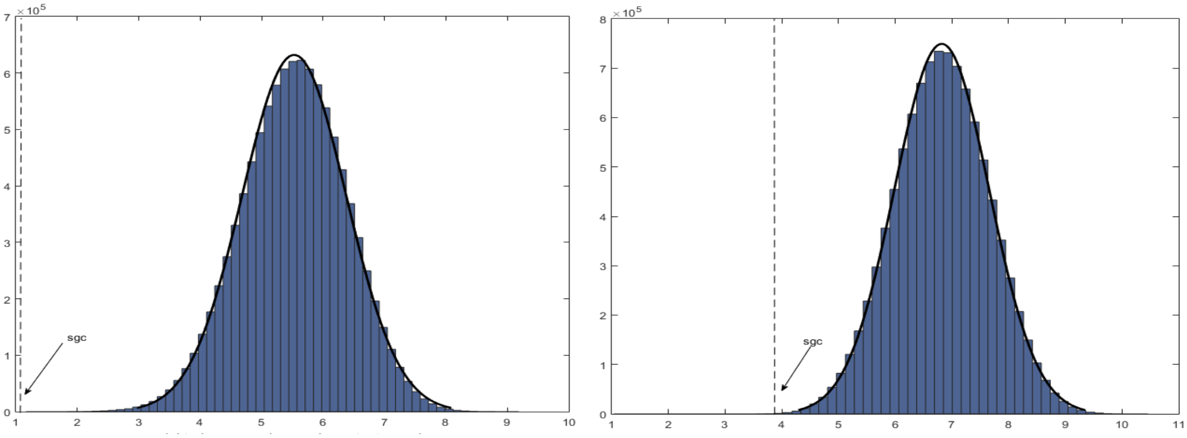


Figure. 5.1 Biased-substitution (left) and Unbiased-substitution (right) weighted mean phenotypic change defined in terms of the hydrophobicity (Miyazawa's contact energies, p132). The block-based model and random samples of 10050000 codes were used. The dot-dashed lines and arrows indicate the values corresponding to the standard genetic code (SGC). The solid black lines represent Normal distribution fittings.

code is more robust (proportions of randomly generated codes for several amino acid indices in Table IX.2, Appendix IX).

5.2.2 The best-preserved amino acid properties by the natural genetic codes

The robustness is defined as the ability of the genetic codes to mitigate the effect of errors or mutations. The smaller the Cantelli's bound, the greater the relevance of the genetic code robustness according to a given amino acid property. Thus, the amino acid property for which the Cantelli's bound reaches the smallest value is the one that best reflects the robustness of the considered genetic code. The genetic code should be more conservative for the most biologically important amino acid properties.

In previous works, several amino acid properties have been explored to find which of them are better preserved by the standard code [16, 25, 29, 35, 116]. It is known that the polarity/hydrophobicity properties are among the best-preserved properties by this genetic code. For that reason, several papers only use one or some of them to measure the ability of the standard code for error minimization. In this regard, very few properties different from the polarity/hydrophobicity have been studied. On the other hand, there are different ways of modelling, or measuring, the same amino acid property. Consequently, different numerical values are assigned to the same amino acid and same property according to different amino acid property scales. Hence, to improve our understanding of the extent of the genetic code's ability for error minimization, the spectrum of properties was broadened by including not only different amino acid properties but also several measurements or scales for the same property. This strategy could reveal new properties and scales preserved by the standard code, not necessarily related to polarity/hydrophobicity. The place that a given scale or group of scales tends to have in the ranked list of the 235 amino acid properties reflects its biological significance. In this way, the standard genetic code itself might be used to discern the relative importance of each amino acid property.

Here we will focus on the first 10 of these properties with the smallest Cantelli's bound value. Overall, for the three graph representations of the standard code (block-based, sense-codon based, and codon based) as well as for the biased and unbiased weightings, 7 of the 37 hydrophobicity/polarity scales appear among the top ten properties in the ranking of the 235 amino acid attributes according to the Cantelli's bounds (see details in tables 5.1-5.2, tables II.4-II.7 in appendix II).

The Polar Requirement has been the most used scale to measure the capacity of the standard genetic code for error minimization. Our findings suggest that Polar Requirement is the best attribute to quantify the efficiency of the genetic code for error minimization, but only when using the codon block-based representation for the standard genetic code. (table 5.1, table II.1 in appendix II).

This finding corroborates previous results based on this scale and the same representation of the standard code but using other methods. Nevertheless, using the codon-based representations, the Miyazawa's contact energies turned out to be the amino acid hydrophobicity scale best reflecting the capacity of the standard genetic code for error minimization, regardless of weights used for the graph G^g. Moreover, the Miyazawa Contact energies is, on average, the scale for which the 23 natural genetic codes showed to be the most effective for error minimization (fig 5.2, table 5.2, tables II.2, II.6, II.7, in appendix II). In general, some of the hydrophobicity scales ranked, as expected, among the best-preserved scales by the genetic codes (fig 5.2, fig II.3 in appendix II), corroborating that the genetic codes are structured in such a way that amino acids similar in hydrophobicity/polarity are encoded by similar nucleotide codons [16, 25, 116]. This arrangement of the genetic codes has the effect of minimizing changes in hydrophobicity/polarity caused by single-base substitutions, thus, being a significant buffering mechanism at the level of proteins, given the crucial role that this amino acid property plays in protein folding and stability [117, 118]. Apart from hydrophobicity/polarity scales,

Table 5.1 Biased-weighted mean phenotypic change under the block-based model (rob). The first 10 aa properties (from a total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than that of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. Pr sim: Probability estimated by numerical simulation. Pr norm: Probability estimated by normal approximation. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

Amino acid properties	rob	Score	CB	Pr(AC)	Pr,sim	Pr norm
Polar requirement (p149)	0.2779	-3.2203	0.0879	0.3548	2.0000E-06	0.0006
Hydrophobicity (Wimley, p148)	0.3489	-2.1718	0.1749	0.5000	3.2338E-05	0.0149
Hydrophobicity (Meek, p 130)	0.4005	-2.1192	0.1821	0.4444	5.2100E-05	0.0170
Long-range contacts (p164)	0.2959	-2.1154	0.1826	0.3685	6.4000E-06	0.0172
Flexibility (2FN, MS, p209)	0.3068	-2.0717	0.1890	0.3468	1.5100E-05	0.0191
Flexibility (2FN, ML, p184)	0.3343	-2.0639	0.1901	0.4169	1.6100E-05	0.0195
Solvent accesible surface (p44)	0.3055	-2.0269	0.1958	0.3266	1.5700E-05	0.0213
Hydrophobicity (Cowan,p 117)	0.3308	-2.0108	0.1983	0.3347	3.8900E-05	0.0222
Flexibility (MS,p212)	0.3217	-1.9289	0.2118	0.3581	6.2687E-06	0.0269
Hydrophobicity (Miyazawa, p132)	0.2858	-1.9108	0.2150	0.4266	7.8000E-06	0.0280

2FN: Two flexible neighbors, MS: Mean scale parameter, ML: Mean location parameter

Table 5.2 Biased-weighted mean phenotypic change (rob) under the model based on sense codons (rob). The first 10 aa properties (from a total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than that of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**
Hydrophobicity(Miyazawa, p132)	0.2858	-11.2272	7.8709E-03	0.1605
Hydrophobicity(Kyte, p125)	0.3442	-11.1653	7.9578E-03	0.2113
Hydrophobicity(Cowan, p117)	0.3308	-10.9024	8.3429E-03	0.1774
Transmembrane alpha-helix(p35)	0.3434	-10.8956	8.3533E-03	0.1806
Long-range contacts (p164)	0.2959	-10.8850	8.3693E-03	0.1347
Solvent accesible surface (p44)	0.3055	-10.8621	8.4044E-03	0.1581
Transmembrane alpha-helix(p28)	0.4039	-10.8394	8.4394E-03	0.2177
Hydrophobicity(Parker, p135)	0.3314	-10.7593	8.5644E-03	0.1718
Polar requirement (p149)	0.2779	-10.6454	8.7470E-03	0.1726
Hydrophobicity(Cornette,p115)	0.3669	-10.6039	8.8151E-03	0.1823

we observe 5 groups of scales clustering among the top positions for the standard genetic code regardless of the weightings and genetic code representations (tables 5.1, 5.2, tables II.4, II.5, fig II.3 in appendix II). These scales are grouped under the following categories, the average long-range contacts, the hydrophilic accessible surface, the conformational flexibility of amino acids in proteins, as well as, the amino acid propensities for small linkers and for transmembrane alpha-helices (fig 5.2, fig II.3 in appendix II). One can wonder why these properties rank, on average, much better than most hydrophobicity/polarity scales.

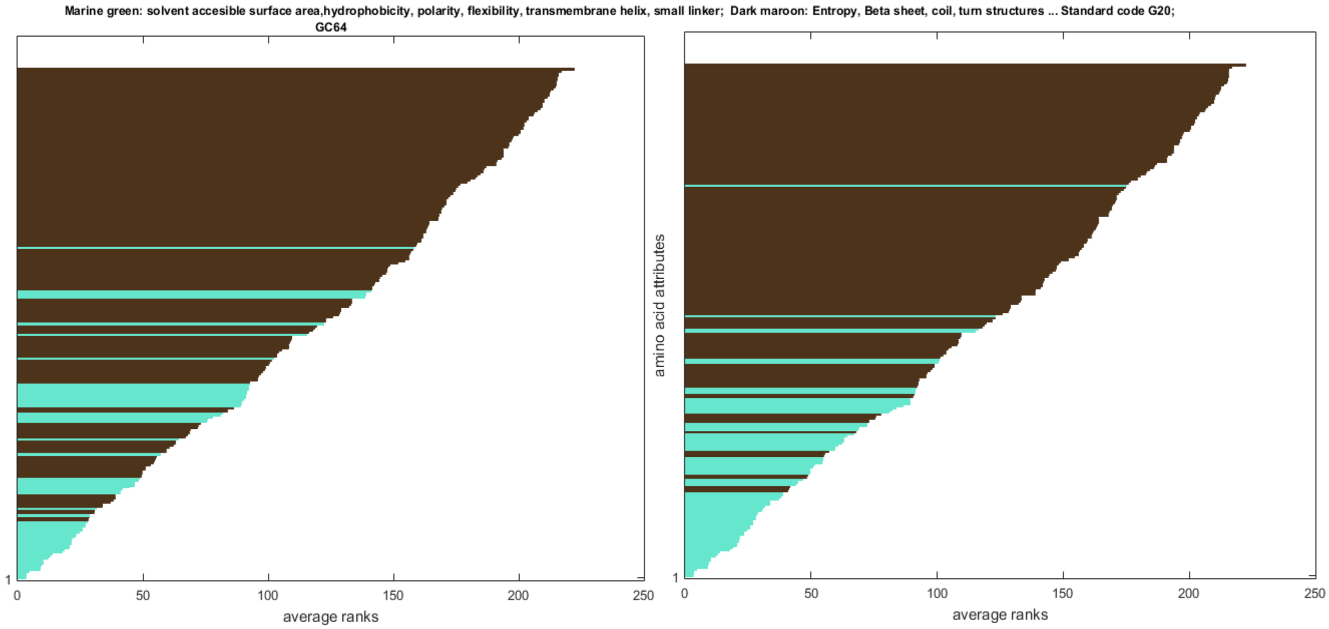


Figure 5.2 Average ranks calculated from the Cantelli's upper bounds for the biased weighted mean phenotypic change for 23 genetic codes and 235 amino acid attributes. Block-based model (the figure on the right side) and Codon-based Model, stop codon=Mean suppressor (the figure on the left side). Marine green: Long-range contacts, Solvent Accessible surface area, Hydrophobicity/polarity, flexibility, transmembrane helix and small linker propensities; Dark maroon: The other amino acid properties.

We can raise three reasons, the first reason is that the genetic code is only optimized for hydrophobicity, and the high ranking of the 5 other groups of properties is only due to the close relation between them and the hydrophobicity property. The second reason is that the genetic code is simultaneously optimized for many properties. The third reason is that the fact that some of the hydrophobicity scales are not highly ranked is only due to the approximate way of estimating them using experimental or computational methods. In general, for all amino acid properties and weighting schemes the measures of relevance for codon-based representations have shown the highest values. This is due to the following factors: the weighing schemes, the methods to assign numerical values to stop codons as well as the constraints imposed by different genetic code representations for computing the moments of the distribution. Notice that the block-based representations have the following two constraints: 1) Each vertex represents one block of synonymous codons, thereby, excluding the synonymous single-base changes. 2) The stop codons are excluded from the model and thus also, all the single-base changes towards or from these codons. In contrast, the only constraint imposed by the model based on the whole set of codons is to keep the number of codons assigned to each amino acid. This constraint is much weaker than the first constraint of block-based models because it does not imply the exclusion of any single-base change. Moreover, the results obtained by biased-weighting result in score and Cantelli's bound values smaller than those computed under the unbiased weighting. This difference clearly visible when comparing the codon-based models with both weightings (tables 5.2, tables II.2, II.4-II.7 in appendix II) is consistent with previous reports on the genetic code [26]. Finally, we observed that artificial codes obtained by changing the assignment of a single codon is sufficient to obtain genetic codes more robust than the standard genetic code. This observation is interesting as it validates previous results obtained by different analysis methods stating that the standard genetic code is neither a local minimum, nor a global minimum [30, 37, 119]. This observation is even more evident for the representations based on codon blocks (Table 5.1 and 5.2, II.1-II.2, II.4-II.7 in appendix II, Pr(AC)).

5.2.3 Comparing natural genetic codes according to robustness

Most of the known species use the standard genetic code. The difference between this code and the known as Bacterial, archaeal and plant plastid code is restricted to the identity of the start codons. They have, therefore, the same mean phenotypic change value. Moreover, in presence of ambiguous coding rules involving stop codons, only the sense codons were considered. In this sense, the Karyorelict and the Condylostoma nuclear codes turned out to be also equivalent.

In this section, we use the Miyazawa's contact energies hydrophobicity scale, as it is the scale showing the most relevant value of genetic code robustness (see previous section). As observed in tables 5.3 and 5.5 (Pr(AC), see also tables III.1, III.2, III.4 in appendix III), the standard genetic code is among the three most robust natural codes, which is consistent with previous results [120]. Overall, the number of natural genetic codes more robust than the standard genetic code is very small compared to the proportion of artificial genetic codes (obtained by only one codon reassignment compared to the standard genetic code) more conservative than the standard genetic code (tables 5.3, 5.5 tables III.1, III.2, III.4, appendix III). This is even more evident for the block-based models. Moreover, none of the nuclear codes containing one reassigned codon with respect to the standard code is more robust than it. It is known that the variant genetic codes arise from the standard code by codon reassignments. Therefore, these results suggest that the codon reassignments leading to more robust codes do not play a crucial role in the evolution of variant genetic codes from the standard genetic code at least for some natural genetic codes.

5.2.4 The robustness of the standard genetic code for each codon position

For the top 10 amino acid properties in codon-based models, the robustness values of the standard code at the third codon position are more relevant than those at the first position. This trend was also observed in the block-based models except for one hydrophobicity scale (Meek). For all amino acid properties and genetic code representations, the third and first codon positions are more robust than the second codon position. (Tables 5.4, 5.6, III.3, III.5 in appendix III) These findings corroborate previous results according to which the robustness values are biased with respect to codon position [26, 79]. Important evolutionary constraints imposed to the arrangement of the genetic codes by the translation errors could explain why the codon positions significantly differ in robustness. These differences in the degree of load minimization reflects mainly differences in the relative frequencies

Table 5.3 Biased-weighted mean phenotypic change (rob) under the block-based model. Scores for 23 genetic codes sorted in increasing order of their Cantelli's bounds (CB), The phenotype is expressed in terms of hydrophobicity (Miyazawa's contact energies), Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers in parentheses (In the footnotes and in first column of the table) indicate the NCBI translation table.

Genetic codes	rob	Score	CB	Pr(AC)
Traustochytrium mitochondrial Code (23)	0.28408	-1.91423	0.21440	0.44773
The standard genetic Code(1)*	0.28582	-1.91081	0.21500	0.42661
The Ciliate, Dasycladacean and Hexamita Nuclear Code (6)	0.29305	-1.88353	0.21989	0.45164
Peritrich Nuclear Code (30)	0.29696	-1.87946	0.22063	0.45000
Mesodinium Nuclear Code (29)	0.29035	-1.87530	0.22140	0.44672
The ascidian Mitochondrial Code (14)	0.29831	-1.87261	0.22189	0.48158
The Mold, Protozoan, Coelenterate Mitochondrial Code (4)**	0.29836	-1.86993	0.22239	0.42063
The Vertebrate Mitochondrial Code (2)	0.30275	-1.85809	0.22459	0.42656
The Euplotid Nuclear Code (10)	0.30169	-1.84668	0.22675	0.43145
Candidate Division SR1 and Gracilibacteria Code (25)	0.30766	-1.84411	0.22723	0.44677
Pachysolen tannophilus Nuclear Code (26)	0.31648	-1.84298	0.22745	0.42984
The Invertebrate Mitochondrial Code (5)	0.29998	-1.84139	0.22775	0.43281
Trematode Mitochondrial Code (21)	0.29737	-1.83934	0.22815	0.43571
The Echinoderm and Flatworm Mitochondrial Code (9)	0.29636	-1.83112	0.22973	0.44194
Karyorelict Nuclear Code (27)***	0.30937	-1.83038	0.22987	0.44365
Blastocrithidia Nuclear Code (31)	0.31385	-1.82457	0.23100	0.44286
Pterobranchia Mitochondrial Code (24)	0.30946	-1.82253	0.23140	0.43175
The Alternative Flatworm Mitochondrial Code (14)	0.29736	-1.82132	0.23163	0.44344
Cephalodiscidae Mitochondrial UAA-Tyr Code(33)	0.31170	-1.80746	0.23436	0.43952
Scenedesmus obliquus Mitochondrial Code (22)	0.33100	-1.77118	0.24172	0.43468
Chlorophycean Mitochondrial Code(16)	0.33136	-1.75852	0.24436	0.43790
The alternative yeast nuclear Code (12)	0.35567	-1.70472	0.25601	0.43790
The Yeast Mitochondrial Code(3)	0.36198	-1.67550	0.26265	0.42969

* The Bacterial, archaeal and plant plastid Code (11) has the same parameter values as the standard code,
** Full name: The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (4)
***The Condylostoma nuclear Code (28) has the same parameter values as the Karyorelict nuclear code,

of mistranslation errors, because these errors are much more frequent than transcription errors and mutations [1,3].

5.2.5 Comparing natural genetic codes according to substitution positions and types

This study extends previous results on the standard code by showing that the robustness of the natural genetic codes differs according to codon position. The Cantelli's bound values shown on both axes indicate that the third codon position is the most robust followed by the second and first codon positions in that order (fig 5.3).

Table 5.4 Biased-weighted mean phenotypic change (rob) under the partial block-based models for the standard code. The 10 amino acid properties correspond to those of table. p1: first codon position. p2: second codon position. p3: third codon position. rob: standard code robustness. cb: Cantelli's upper bound. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

	p1			**p2**			**p3**		
Amino acid properties	**rob**	**score**	**cb**	**rob**	**score**	**cb**	**rob**	**score**	**cb**
Polar requirement (p149)	0.5107	-2.8134	0.1122	0.3024	-1.6981	0.2575	0.0233	-4.5171	0.0467
Hydrophobicity (Wimley,p148)	0.6141	-1.9094	0.2152	0.2604	-1.5997	0.2810	0.1751	-2.5457	0.1337
Hydrophobicity (Meek, p130)	0.5847	-2.1950	0.1719	0.3067	-1.4258	0.3297	0.3122	-1.9376	0.2103
Long-range contacts (p164)	0.3621	-2.2992	0.1591	0.3958	-0.6200	0.7223	0.1307	-2.7856	0.1142
Flexibility (2FN, MS, p209)	0.5263	-1.9139	0.2144	0.2754	-1.3575	0.3518	0.1214	-2.8754	0.1079
Flexibility (2FN, ML, p184)	0.4718	-2.1413	0.1790	0.3184	-1.1427	0.4337	0.2143	-2.3433	0.1541
Solvent accesible Surface (p44)	0.4474	-2.0530	0.1918	0.3686	-0.7703	0.6276	0.1021	-2.9447	0.1034
Hydrophobicity (Cowan, p117)	0.5191	-1.9700	0.2049	0.4246	-0.4550	0.8285	0.0509	-3.3819	0.0804
Flexibility (MS,p212)	0.5247	-1.8434	0.2274	0.3537	-0.8459	0.5829	0.0891	-2.9829	0.1010
Hydrophobicity (Miyazawa, p132)	0.3142	-2.1465	0.1783	0.5178	0.1096	0.9881	0.0258	-3.1182	0.0933

Table 5.5 Biased-weighted mean phenotypic change (rob) under the codon-based model with codon stop=scale mean and scores for 23 genetic codes sorted in increasing order of their Cantelli's bounds (Cbound). The phenotype is expressed in terms of hydrophobicity (Miyazawa's contact energies), rob: robustness, CB: Cantelli's bound, Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers in parentheses (In the footnotes and first column of the table) indicate the NCBI translation table.

Genetic codes	**rob**	**score**	**CB**	**Pr(AC)**
The Ciliate, Dasycladacean and Hexamita Nuclear Code (6)	0.30050	-11.54775	0.007443	0.14344
The standard genetic Code (1)*	0.29419	-11.54154	0.007451	0.13468
The Invertebrate Mitochondrial Code (5)	0.30172	-11.52646	0.007471	0.11875
The Mold, Protozoan, and Coelenterate Mitochondrial Code (4)**	0.30020	-11.52475	0.007473	0.12143
Trematode Mitochondrial Code (21)	0.29783	-11.52116	0.007477	0.11984
The Echinoderm and Flatworm Mitochondrial Code (9)	0.29688	-11.52022	0.007479	0.12258
Peritrich Nuclear Code (30)	0.30462	-11.51620	0.007484	0.14918
The ascidian Mitochondrial Code (14)	0.30015	-11.51391	0.007487	0.12813
Mesodinium Nuclear Code (29)	0.29658	-11.50963	0.007492	0.13770
The Alternative Flatworm Mitochondrial Code (14)	0.29783	-11.50771	0.007495	0.12377
Karyorelict Nuclear Code (27)***	0.30937	-11.49534	0.007511	0.13651
The Euplotid Nuclear Code (10)	0.30378	-11.49447	0.007512	0.12258
Pterobranchia Mitochondrial Code (24)	0.31064	-11.46938	0.007545	0.13968
Blastocrithidia Nuclear Code (31)	0.31385	-11.46045	0.007556	0.14286
Cephalodiscidae Mitochondrial UAA-Tyr Code (33)	0.31171	-11.45581	0.007562	0.14032
Candidate Division SR1 and Gracilibacteria Code (25)	0.30797	-11.36612	0.007681	0.15161
The Vertebrate Mitochondrial Code (2)	0.30974	-11.33189	0.007727	0.15781
Chlorophycean Mitochondrial Code (16)	0.34145	-11.02311	0.008163	0.15000
Traustochytrium mitochondrial Code (23)	0.32120	-10.99462	0.008205	0.13952
Pachysolen tannophilus Nuclear Code (26)	0.32218	-10.98076	0.008225	0.12823
Scenedesmus obliquus Mitochondrial Code (22)	0.34511	-10.92780	0.008304	0.16694
The alternative yeast nuclear Code (12)	0.35797	-10.49485	0.008998	0.14597
The Yeast Mitochondrial Code(3)	0.36006	-9.97786	0.009945	0.12344

* The Bacterial, archaeal and plant plastid Code (11) has the same parameter values as the standard code,

** Full name: The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (4)

***The Condylostoma nuclear Code (28) has the same parameter values as the Karyorelict nuclear code,

Table 5.6 Biased-weighted mean phenotypic change (rob) under the partial models based on sense codons of the standard code. The 10 amino acid properties correspond to those of table. p1: first codon position. p2: second codon position. p3: third codon position. rob: standard code robustness. Cb: Cantelli's upper bound. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

	p1			**p2**			**p3**		
Amino acid properties	**rob**	**score**	**cb**	**rob**	**score**	**cb**	**rob**	**score**	**cb**
Hydrophobicity (Miyazawa, p132)	0.3142	-6.6535	0.0221	0.5178	1.4752	0.3148	0.0258	-9.4326	0.0111
Hydrophobicity (Kyte, p125)	0.4326	-6.2770	0.0248	0.4628	-0.7053	0.6678	0.1383	-8.9600	0.0123
Hydrophobicity (Cowan, p117)	0.5191	-5.3754	0.0335	0.4246	-0.5688	0.7555	0.0509	-9.3153	0.0114
Transmembrane alpha-helix (p35)	0.4790	-5.7662	0.0292	0.4426	-0.5178	0.7886	0.1103	-9.0482	0.0121
Long-range contacts (p164)	0.3621	-6.1674	0.0256	0.3958	-0.3247	0.9046	0.1307	-8.8082	0.0127
Solvent accesible Surface (p44)	0.4474	-5.5833	0.0311	0.3686	-0.9240	0.5394	0.1021	-9.0119	0.0122
Transmembrane alpha-helix (p28)	0.4694	-6.2648	0.0248	0.5789	0.3525	0.8895	0.1643	-8.8927	0.0125
Hydrophobicity (Parker, p135)	0.3784	-6.2651	0.0248	0.4938	0.7836	0.6196	0.1227	-8.9236	0.0124
Polar requirement (p149)	0.5107	-4.5651	0.0458	0.3024	-1.3870	0.3420	0.0233	-9.3166	0.0114
Hydrophobicity (Cornette, p115)	0.3066	-6.9096	0.0205	0.6361	2.5487	0.1334	0.1572	-8.8204	0.0127

The biased and unbiased weighted mean change in Miyazawa's contact energies were computed for each codon position. Among the 23 genetic codes, 3 codes are more efficient than the standard genetic code at the first codon position, 11 perform better than the standard code at the second position and 12, at the third codon position. For transversions and transitions, only one genetic code for each of these substitution types is more robust than the standard code. (fig 5.3) Thus, the standard genetic code is more robust than most genetic codes at the first position and with respect to both substitution types.

Concerning the substitution type, no nuclear code is more efficient than the standard code (fig 5.3). Furthermore, only two nuclear codes are more robust than the standard code in the third position. There are three nuclear genetic codes more robust than the standard genetic code at the first and second codon positions, namely, the Peritrich Nuclear Code, the Candidate Division SR1 and Gracilibacteria Code and the Ciliate, Dasycladacean and Hexamita Nuclear code (fig 5.3, top left). At the second and third codon positions, the Mesodinium nuclear code is more robust than the standard genetic code. Even, the Ciliate, Dasycladacean and Hexamita Nuclear code is the only genetic code more robust than the standard code according to the codon-based model (fig 5.3). Thus, some variant nuclear genetic codes could have evolved from the standard genetic code by codon reassignments that do not increase the robustness of the whole code but only the robustness at one or two codon positions.

As for the mitochondrial genetic codes, we observed that 5 and 10 codes are more robust than the standard genetic code at the codon positions 2 and 3, respectively, of a total of 13 mitochondrial codes. Even though the standard genetic code is more robust than most mitochondrial genetic codes according to several models representing the whole set of codons or blocks, we obtained that almost all mitochondrial genetic codes are more robust than the standard code at the third codon position. This result suggests that the increase of the robustness at the third codon position seems to play a crucial role for the evolution of the mitochondrial codes. It is interesting to note that, unlike the other mitochondrial codes, there are 5 mitochondrial codes more robust than the standard code at two codon positions. More specifically, the Ascidian Mitochondrial Code, the Trematode Mitochondrial Code, the Echinoderm and Flatworm Mitochondrial Code and the Alternative Flatworm Mitochondrial Code are more robust than the standard code at the codon positions 2 and 3. Moreover, the vertebrate mitochondrial code is more robust than the standard code for the transversions, as well as, for the codon positions 2 and 3. (fig 5.3). In general, the codon reassignments lead to more robust codes at the third codon position for the mitochondrial genetic codes and, at the first codon position, for the nuclear genetic codes. This finding is consistent with the idea that some alternative genetic codes have evolved from the canonical genetic code by codon reassignments that result in a partial optimization with respect to codon position.

5.2.6 Neighborhood structure of natural genetic codes

In this work, the classification of the codon blocks as homogeneous, or heterogeneous, according to their amino acid neighborhoods has been useful for reducing the number of vertices and edges to.

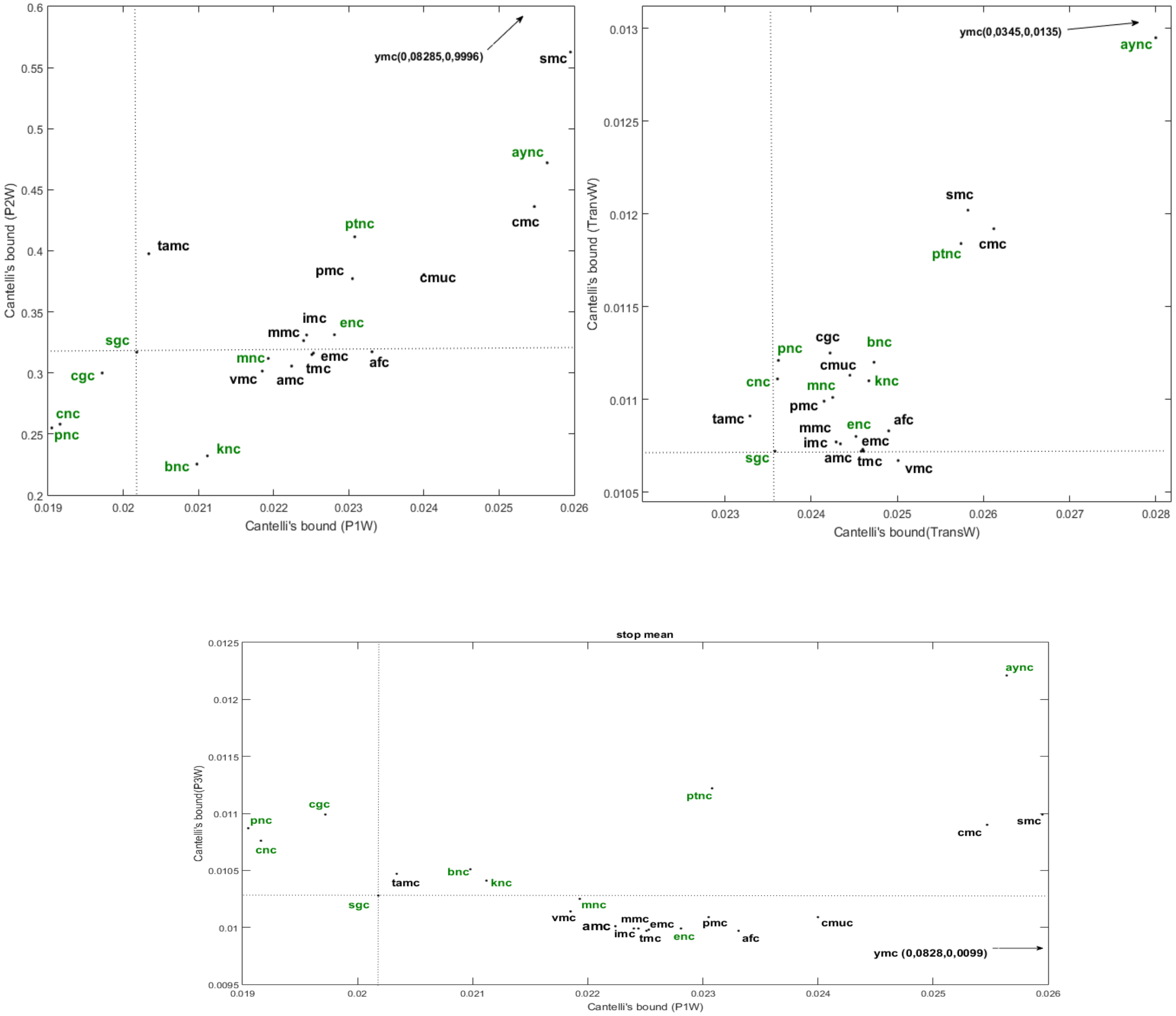


Figure 5.3 Cantelli's bounds corresponding to the Biased weighted mean change in Miyazawa's contact energies under the codon-based model with stop codon=mean suppressor. Codon positions and the substitution type (transition/transversion) for 23 genetic code variants. Top left: Second codon position Cantelli bound versus first codon position Cantelli's bound, Top right: Transversion Cantelli bound versus Transition Cantelli's bound, Bottom: Third codon position Cantelli bound versus first codon position Cantelli's bound,. Letters in green: Nuclear genetic codes, Letters in black: Mitochondria and plastid genetic codes, sgc: standard code, vmc: The Vertebrate Mitochondrial Code, ymc: The Yeast Mitochondrial Code, mmc: The Mold, Protozoan, and Coelenterate Mitochondrial Code, ivmc: The Invertebrate Mitochondrial Code, cnc: The Ciliate, Dasycladacean and Hexamita Nuclear Code, emc: The Echinoderm and Flatworm Mitochondrial Code, enc: The Euplotid Nuclear Code, amc: The Ascidian Mitochondrial Code, aync: The Alternative yeast nuclear code, afc: The Alternative Flatworm Mitochondrial Code, cmc: Chlorophycean Mitochondrial Code, tmc: Trematode Mitochondrial Code, smc: Scenedesmus obliquus Mitochondrial Code, pmc: Pterobranchia Mitochondrial Code, cgc: Candidate Division SR1 and Gracilibacteria Code, ptnc: Pachysolen tannophilus Nuclear Code, knc: Karyorelict Nuclear Code, mnc: Mesodinium Nuclear Code, pnc: Peritrich Nuclear Code, bnc: Blastocrithidia Nuclear Code, cmtc: Cephalodiscidae Mitochondrial UAA-Tyr Code, tamc: Traustochytrium mitochondrial code.

compute the genomic robustness. As explained in chapter 4, this is possible by excluding the homogeneous blocks, because only the contribution of heterogeneous blocks to the mean phenotypic change weighted with the synonymous codon usage vary among different genomes. In this section we explore some regularities in the neighborhood structure of the 23 genetic codes.

Considering only the sense codons, the standard genetic code has a set of 10 synonymous codon blocks containing at least two codons with different robustness values which represents the half of the code. These codon blocks specify the following amino acids, L, S, P, R, I, T, K, V, A and G (table III.6). Two of these amino acids, K and I, are encoded by the codons of blocks with homogeneous neighborhood in some genetic code variants. In contrast, some other codon blocks which are homogeneous in the standard genetic code, turn out to be heterogeneous in other genetic codes, such as, those specifying the amino acids, Y, Q, W, N and C (table III.6, Figure VIII.1). We describe two kinds of codon blocks: 1) the general heterogeneous blocks, defined as synonymous codon blocks with heterogeneous amino acid neighborhood in the 23 genetic codes, 2) the general homogeneous blocks, defined as the synonymous codon blocks with homogenous amino acid neighborhood in the 23 genetic codes.

1) L, S, P, R, T, V, A and G
2) F, H, M and D

Although these codon blocks vary according to the number of synonymous codons and the composition of their amino acid neighborhoods among different genetic codes, the neighborhoods of the group 1 remain heterogeneous and those of the group 2, homogenous, in all the variant genetic codes considered in this work. It is noteworthy that all the amino acids of the group 1, except R, are considered primitive amino acids because they have been detected in experiments that simulate the conditions of the early earth, in the Murchison meteorite and in the estimates of ancestral sequence composition [121].

We can wonder whether the heterogeneous structure of the neighborhoods of type 1 would confer evolutionary advantage or not, at least before the standard genetic code fixation. The existence of codons specifying the same amino acid but with different robustness values allows to adaptively adjust the robustness of a given genome, by modifying its codon usage bias. We can consider that the synonymous substitutions leading to more robust genomes increase the fitness of the organism. But there are other known factors that may affect in opposite direction the codon usage bias by augmenting the frequency of less robust synonymous codons, for example, the genetic drift or the positive selection to enhance the RNA stability or the translation efficiency of highly expressed genes [122]. Hence, a greater number of heterogeneous blocks or a greater number of homogeneous sub-blocks and isolated codons per heterogeneous block (ρ), could improve the ability for adaptive evolution of microorganisms exposed, for example, to environmental factors increasing the error rate, because there would be more possibilities to maximise the robustness under the constraints imposed by the other factors.

The standard code has the fourth greatest value of ρ among all genetic codes included in this work. (table 5.7). Moreover, all the nuclear codes, except two, have values of ρ greater than 3 and all the mitochondrial codes except one have values of ρ below 3 (table 5.7). This difference between the nuclear and mitochondrial codes is a consequence of the difference in the number of codon reassignments between both sub-groups of genetic codes. More precisely, almost all the mitochondrial codes have between 4 and 5 reassigned codons and all the nuclear codes have between 1 and 3 reassigned codons with respect to the standard code (table 5.7).

Table 5.7 Neighborhood structure of 22 natural genetic codes. HB: the number of homogeneous blocks. SIC: number of homogeneous sub-blocks and isolated codons. ρ: the number of homogeneous sub-blocks and codons per heterogeneous block, in this case, neither the homogeneous blocks nor the stop codons were included in the sets of homogeneous sub-blocks and isolated codons, respectively. CR: the number of codon reassignments with amino acid encoded by adjacent (n) and non-adjacent (c) codons. Totc: Total number of codons and their adjacent codons affected by codon reassignments taking the standard code as reference. The numbers in parentheses (In the footnotes and in first column of the table) indicate the NCBI translation table. Green: nuclear codes.

Genetic codes	**HB**	**sic**	**ρ**	**CR (n,c)**	**tot c**
The Standard Code (1) *	10	41	3.1000	0	0
The Vertebrate Mitochondrial Code (2)	12	30	2.2500	4 n	24
The Yeast Mitochondrial Code (3)	12	31	2.3750	2 n+4 c	37
The Mold, Protozoan, and Coelenterate Mitochondrial Code (4)**	10	37	2.7000	1 n	9
The Invertebrate Mitochondrial Code(5)	12	31	2.3750	4 n	24
The Ciliate, Dasycladacean and Hexamita Nuclear Code (6)	10	43	3.3000	2 n	15
The Echinoderm and Flatworm Mitochondrial Code (9)	8	40	2.6667	4 n	24
The Euplotid Nuclear Code (10)	9	42	3.0000	1 n	9
The Ascidian Mitochondrial Code (13)	13	31	2.5714	4 n	24
The Alternative yeast nuclear Code (12)	10	42	3.2000	1 c	9
The Alternative Flatworm Mitochondrial Code (14)	5	44	2.6000	5 n	27
Chlorophycean Mitochondrial Code (16)	8	44	3.0000	1 n	9
Trematode Mitochondrial Code (21)	9	37	2.5455	5 n	28
Scenedesmus obliquus Mitochondrial Code (22)	7	43	2.7692	2 n	16
Pterobranchia Mitochondrial Code (24)	9	41	2.9091	3 n	20
Candidate Division SR1 and Gracilibacteria Code (25)	9	42	3.0000	1 n	9
Pachysolen tannophilus Nuclear Code (26)	8	41	2.7500	1 c	10
Karyorelict Nuclear Code (27)***	10	40	3.0000	3 n	21
Mesodinium Nuclear Code (29)	9	43	3.0909	2 n	15
Peritrich Nuclear Code (30)	10	43	3.3000	2 n	15
Blastocrithidia Nuclear Code (31)	10	38	2.8000	3 n	21
Cephalodiscidae Mitochondrial UAA-Tyr Code (33)	7	44	2.8462	4 n	25

*Bacterial, archaeal, plant plastid Code (11)

** The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (4)

***Condylostoma Nuclear Code (27)

5.3 Genomic robustness

The genomic robustness is referred to the ability of genomes to minimize the effect of errors and mutations. A genome is more robust if the more robust codons are more frequently used than the less robust codons (see section Methods). The proteins of thermophilic bacteria and archaea have numerous adaptations that make them stable and functional at high temperature. Significant differences have been reported between thermophiles and non-thermophiles at the amino acid and codon usage level. However, little has been made to explore the link between the codon robustness, codon usage and thermophily..

There are two approaches to explain the evolution of codon usage bias, one is based on the idea of the non-random distribution of mutational pressure at the genome level, and the other approach uses the concept of natural selection to understand how some codons are favoured over the other synonymous codons. It was shown, mainly for the highly expressed proteins, that the amino acids encoded by the most frequent codons tend to be more efficiently and accurately incorporated in proteins than those encoded by the less frequent codons [122, 123]. Thus, different contributions of codons to the efficiency and accuracy of protein expression leads to a biased distribution of synonymous codon frequency. In addition, several other factors have been shown to drive the evolution of codon usage bias, such as, protein hydrophobicity, RNA stability, optimal growth temperature and codon robustness, among others [48,122, 123]. In this section, we focus on the last two factors.

We identified the amino acid properties for which the genomic robustness reaches the greatest values according to a given genetic code model. The genomic robustness could be interpreted as a measure of correlation between the genomic synonymous codon bias and the codon robustness according to a given amino acid property (see section Methods). If the value of the score corresponding to genomic robustness for a given property is significantly smaller than that obtained by using other amino acid properties, this amino acid property will be considered as a relatively more important factor in the evolution of the genomic synonymous codon bias than the other properties. We also determined the amino acid properties for which the codon robustness is more strongly correlated with the synonymous codon usage in thermophiles than in non-thermophiles as well as the codon position or substitution type that contributes more to genomic robustness in thermophiles than in non-thermophiles. To compute the robustness of thermophilic (324) and non-thermophilic prokaryotes (418) under 84 different amino acid indices. Two measures of relevance of robustness were considered, namely, the scores and optimization percentage.

5.3.1 The amino acid properties that maximize the genomic robustness

According to all codon-based models, the two scales for which the genomic robustness reach the most relevant values are the Miyazawa's contact energies for the optimization percentage, and the Kyte's hydropathy index for the scores (tables 5.8 and 5.9). More precisely, the Kyte scale is linked to the most significant genomic robustness values for scores, and to the second most optimized robustness values according to the optimization percentage, reaching a value of 6.72 standard deviations below the mean which is equivalent to an optimization percentage of 90.43% in thermophiles. We observed that the top 5 to 10 indices linked to the most significant values of genomic robustness, correspond to the hydrophobicity property regardless of the methods used to weight single-base changes, to process the stop codons or to estimate the genomic robustness. In general, 20 of the 31 chosen hydrophobicity scales was found among the first 30 positions of the 84 amino acid indices sorted in decreasing order of their corresponding values of optimization percentage or scores (figure 5.4, figures IV.1, IV.2 and IV.4 in appendix IV). These findings suggest that the genetic code structure and protein stability are strongly associated with general trends in codon usage bias. The Polar Requirement is linked to values of optimization percentages and scores less significant than most of those corresponding to hydrophobicity/polarity and accessible surface area indices. This is interesting because the Polar Requirement has been used to estimate the degree of error minimization in coding sequences [35,45]. In general, the thermophiles showed estimates of robustness more relevant according to both relevance measures than those observed for non-thermophiles. Likewise, it was observed for both groups of genomes that the robustness values computed under the weighting based on mistranslation rates (also called, biased weighting, table 5.9) are more relevant than those observed for the

Table 5.8: Medians corresponding to the Polar Requirement and the 3 amino acid properties linked to the largest medians of the Optimization percentage (MP) for thermophilic and non-thermophilic prokaryotes for each genetic code representation. MSW: The standard code model based on sense codons and biased weighting. MS: The standard code model based on sense codons and unbiased weighting. M0W: The codon-based model of the standard code with biased weighting and scale mean values assigned to stop codons. M0: The codon-based model of the standard code with unbiased weighting and scale mean values assigned to stop codons. MMW: The codon-based model of the standard code with biased weighting and the values assigned to stop codons according to the "mean suppressor" method. MM: The codon-based model of the standard code with unbiased weighting and the values assigned to stop codons according to the "mean suppressor" method. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324).The First (1Q) and third (3Q) quartiles are shown in parenthesis.

Models	Thermophiles		Non-thermophiles	
	Amino acid properties	Medians(1Q, 3Q)	Amino acid properties	Medians (1Q, 3Q)
MSW	Hydrophobicity(Miyazawa,40)	91.28(90.54, 92.02)	Hydrophobicity(Miyazawa,40)	90.43(89.18, 91.18)
	Hydrophobicity(Kyte,33)	90.50(89.65, 91.24)	Hydrophobicity(Kyte,33)	90.18(89.06, 90.93)
	Hydrophobicity(Cowan, 27)	90.01(89.14, 90.85)	Hydrophobicity(Cowan, 27)	89.68(88.53, 90.49)
	Polar Requirement (Woese)	72.73(71.90,73.77)	Polar Requirement (Woese)	71.96(70.56,73.16)
MS	Hydrophobicity(Miyazawa,40)	79.27(77.76, 80.46)	Hydrophobicity(Miyazawa,40)	78.98(76.16, 80.59)
	Hydrophobicity(Kyte,33)	77.78(75.88, 79.70)	Hydrophobicity(Kyte,33)	77.84(75.05, 79.50)
	Hydrophilicity(Parker,42)	76.71(74.87, 78.37)	Hydrophilicity(Parker,42)	75.95(72.93, 77.53)
	Polar Requirement (Woese)	46.99(45.66,47.91)	Polar Requirement (Woese)	46.68(44.65,47.83)
M0W	Hydrophobicity(Miyazawa,40)	91.62(90.84, 92.23)	Hydrophobicity(Miyazawa,40)	91.02(89.32, 91.77)
	Hydrophobicity(Kyte,33)	90.43(89.56, 91.24)	Hydrophobicity(Kyte,33)	90.33(88.89, 91.19)
	Hydrophobicity(Cowan, 27)	89.98(88.93, 90.66)	Hydrophobicity(Cowan, 27)	89.73(88,36, 90.50)
	Polar Requirement (Woese)	70.51(69.58,71.73)	Polar Requirement (Woese)	69.54(68.21,70.79)
M0	Hydrophobicity(Miyazawa,40)	79.88(78.42, 81.23)	Hydrophobicity(Miyazawa,40)	79.65(76.80. 81.14)
	Hydrophobicity(Kyte,33)	78.05(76.22, 80.09)	Hydrophobicity(Kyte,33)	77.97(75.40. 79.69)
	Hydrophilicity(Parker,42)	77.00(75.07, 78.69)	Hydrophilicity(Parker,42)	76.23(73.20. 77.64)
	Polar Requirement (Woese)	45.66(44.39,46.90)	Polar Requirement (Woese)	44.93(43.32,46.08)
MMW	Hydrophobicity(Miyazawa,40)	90.94(90.18, 91.55)	Hydrophobicity(Miyazawa,40)	90.28(88.64, 91.00)
	Hydrophobicity(Kyte,33)	90.22(89.39, 90.93)	Hydrophobicity(Kyte,33)	90.04(88.81, 90.79)
	Hydrophobicity(Cowan, 27)	89.79(88.74, 90.48)	Hydrophobicity(Cowan, 27)	89.36(88.20. 90.21)
	Polar Requirement (Woese)	71.06(70.14,72,25)	Polar Requirement (Woese)	70.11(68.75,71.28)
MM	Hydrophobicity(Miyazawa,40)	79.09(77.66, 80.50)	Hydrophobicity(Miyazawa,40)	78.78(76.05, 80.27)
	Hydrophobicity(Kyte,33)	77.35(75.54, 79.43)	Hydrophobicity(Kyte,33)	77.28(74.70. 78.98)
	Hydrophilicity(Parker,42)	76.95(74.96, 78.61)	Hydrophilicity(Parker,42)	76.05(73.10. 77.52)
	Polar Requirement (Woese)	46.21(44.98,47.55)	Polar Requirement (Woese)	45.63(43.97,46.77)

unbiased weighting regardless of the temperature range group, the method used to process the stop codons, and both relevance measures used (table 5.8). Similar general patterns were seen for the optimization percentages and scores corresponding to genomic robustness using code representations containing all single-base changes and those that include the single-base changes involving the first or third codon positions (figure IV.4 in Appendix IV). The Spearman rank correlation was used to determine the strength of the association between two rankings of the 84 amino acid properties, one based on the median scores and the other, on the only one of the codon positions or substitution types, strong correlation was found between both relevance

Table 5.9: Medians corresponding to the Polar Requirement and the 3 amino acid properties linked to the largest medians of the scores for thermophilic and non-thermophilic prokaryotes for each genetic code representation. MSW: The standard code representation based on sense codons and biased weighting. MS: The standard code representation based on sense codons and unbiased weighting. M0W: The codon-based model of the standard code with biased weighting and scale mean values assigned to stop codons. M0: The codon-based model of the standard code with unbiased weighting and scale mean values assigned to stop codons. MMW: The codon-based model of the standard code with biased weighting and the values assigned to stop codons according to the "mean suppressor" method. MM: The codon-based model of the standard code with unbiased weighting and the values assigned to stop codons according to the "mean suppressor" method. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324).The First (1Q) and third (3Q) quartiles are shown in parenthesis.

Models	**Thermophiles**		**Non-thermophiles**	
	Amino acid properties	Medians (1Q,3Q)	Amino acid properties	Medians (1Q,3Q)
MSW	Hydrophobicity(Kyte,33)	-6,4168(-6,9369, -5,8882)	Hydrophobicity(Kyte,33)	-6,1237(-6,5696, -5,4601)
	Hydrophobicity(Miyazawa,40)	-6,2499(-6,7558, -5,6859)	Hydrophobicity(Miyazawa,40)	-5,9352(-6,3592, -5,2385)
	Hydrophobicity(Cowan, 27)	-5,9177(-6,4212, -5,4171)	Hydrophobicity(Cowan, 27)	-5,6468(-6,0842, -5,0275)
	Polar Requirement(Woese)	-3.4722(-3.8651,-3.1053)	Polar Requirement(Woese)	-3.2585(-3.5630,-2.8209)
MS	Hydrophobicity(Miyazawa,40)	-3,8613(-4,1619, -3,4997)	Hydrophobicity(Miyazawa,40)	-3,6949(-3,9898, -3,2249)
	Hydrophobicity(Kyte,33)	-3,7660(-4,0952, -3,4296)	Hydrophobicity(Kyte,33)	-3,6491(-3,9175, -3,2072)
	Hydrophobicity(Fauchere, 29)	-3,4713(-3,7745, -3,1660)	Hydrophobicity(Fauchere, 29)	-3,3366(-3,6110. -2,9350)
	Polar Requirement(Woese)	-1.1780(-1.3139,-1.0525)	Polar Requirement(Woese)	-1.1224(-1.2279,-0.9639)
M0W	Hydrophobicity(Kyte,33)	-6,7169(-7,1973, -6,1287)	Hydrophobicity(Kyte,33)	-6,3468(-6,8318, -5,7125)
	Hydrophobicity(Miyazawa,40)	-6,5632(-7,0448, -5,9766)	Hydrophobicity(Miyazawa,40)	-6,1647(-6,6346, -5,5283)
	Hydrophobicity(Cowan, 27)	-6,2659(-6,7469, -5,7111)	Hydrophobicity(Cowan, 27)	-5,9165(-6,3944, -5,3215)
	Polar Requirement(Woese)	-3,7079(-4,0616, -3,3056)	Polar Requirement(Woese)	-3,4260(-3,7536, -3,0266)
M0	Hydrophobicity(Kyte,33)	-4,7292(-5,1201, -4,2988)	Hydrophobicity(Kyte,33)	-4,4761(-4,8568, -3,9994)
	Hydrophobicity(Miyazawa,40)	-4,7261(-5,0978, -4,2920)	Hydrophobicity(Miyazawa,40)	-4,4685(-4,8378, -3,9490)
	Hydrophobicity(Cornette, 25)	-4,3256(-4,6801, -3,9217)	Hydrophobicity(Fauchere, 29)	-4,0871(-4,4171, -3,5930)
	Polar Requirement(Woese)	-1,9361(-2,1298, -1,7292)	Polar Requirement(Woese)	-1,7924(-1,9667, -1,5770)
MMW	Hydrophobicity(Kyte,33)	-6,8869(-7,3773, -6,2847)	Hydrophobicity(Kyte,33)	-6,5091(-7,0061, -5,8586)
	Hydrophobicity(Miyazawa,40)	-6,6900(-7,1843, -6,0892)	Hydrophobicity(Miyazawa,40)	-6,2864(-6,7692, -5,6363)
	Hydrophobicity(Cowan, 27)	-6,3950(-6,8817, -5,8277)	Hydrophobicity(Cowan, 27)	-6,0425(-6,5269, -5,4240)
	Polar Requirement(Woese)	-3,8070(-4,1722, -3,3955)	Polar Requirement(Woese)	-3,5230(-3,8603, -3,1069)
MM	Hydrophobicity(Kyte,33)	-4,9358(-5,3445, -4,4877)	Hydrophobicity(Kyte,33)	-4,6750(-5,0698, -4,1846)
	Hydrophobicity(Miyazawa,40)	-4,9143(-5,3010. -4,4562)	Hydrophobicity(Miyazawa,40)	-4,6487(-5,0398, -4,1074)
	Hydrophobicity(Cornette, 25)	-4,5464(-4,8980. -4,1225)	Hydrophobicity(Cornette, 25)	-4,2817(-4,6377, -3,8211)
	Polar Requirement(Woese)	-2,0593(-2,2733, -1,8411)	Polar Requirement(Woese)	-1,9190(-2,1106, -1,6869)

measures (estimates between -0.7781 and -0.9795) (Figure IV.3, Table IV.1 in Appendix IV). median minimization percentages. This statistic showed a high consistency between both rankings for different definitions of genomic robustness.

More specifically, significant negative rank correlation coefficients were observed between both relevance measures, (between -0.9174 and -0.9607), for genomic robustness values that include all single-base changes involving the three codon positions and both substitution types (transitions and transversions).

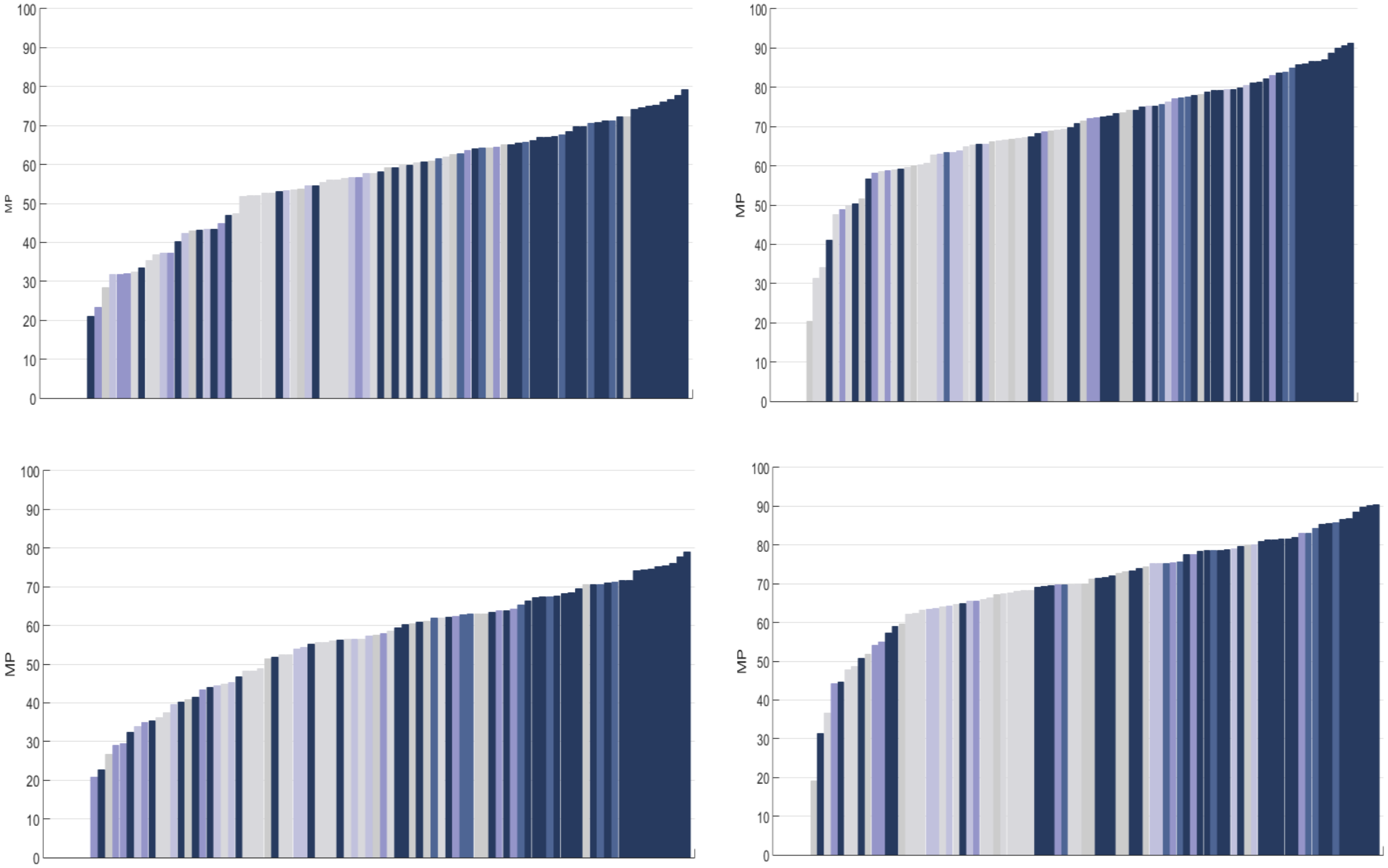


Figure 5.4 The 84 amino acid property scales sorted in order of increasing values of the median Minimization percentages for thermophilic (N=324) genomes and non-thermophilic (N=418) genomes. Each color corresponds to a given type of amino acid property. Each figure corresponds to the Minimization percentage medians computed under the standard code representation based on sense codons. Top left: For thermophilic prokaryotes and unbiased weightings, Top right: For thermophilic prokaryotes and biased weightings, Bottom left: For non-thermophilic prokaryotes and unbiased weightings, Bottom right: For non-thermophilic prokaryotes and biased weightings. (The reference of each amino acid index, in Table IX.1, Appendix IX).

Whereas, for genomic robustness values that include single-base changes involving only one of the codon positions or substitution types, strong correlation was found between both relevance measures (estimates between -0.7781 and -0.9795) (Figure IV.3, Table IV.1 in Appendix IV).

5.3.2 Comparing the thermophilic and non-thermophilic prokaryotes

We aimed at identifying the amino acid properties for which the association between the binary thermal categorisation of genomes and the genomic robustness scores is strongest. Thereby, we could determine those factors linked to thermal stability of proteins that have an important influence on the evolution of codon usage bias. The presence of a strong association between the genomic codon usage bias and codon robustness in thermophiles would suggest a mechanism of thermal adaptation involving codon usage bias. The Three-level logistic mixed models have been used to estimate the association between the relevance measures included as fixed effects and thermic status. Both relevance measures were computed by using three genetic code representations, two weightings and two methods to assign numerical values to stop codons. The Principal component analysis was also applied to the scores and optimization percentages computed under these conditions. To determine which of the first two principal components discriminate better between both temperature-range groups, three-level logistic mixed models were, also, fitted using the principal components as fixed effects.

We observed the most significant differences between both groups in terms of scores and optimization percentage for the genomic robustness computed under the genetic code representations based on sense codons and unbiased weightings. Both groups are clearly separated along the second principal component, showing a significant change in log odds of -2.84 for one-unit increase in the second principal component coordinates (P=1.836e-05) (figures 5.5, table 5.10, tables V.1-V.6, figures V.1-V.3 in Appendix V). Although the average long-range contacts tend to be linked to much smaller robustness values than those linked to hydrophobicity property, the most relevant difference between both temperature range groups were observed for this property (tables 5.10 and 5.11, figure 5.6). In other terms, the association between the robustness and the genomic synonymous codon frequency was stronger in thermophiles than in non-thermophiles when the average long-range contacts were considered. The optimization percentage and scores computed for the representation of the single base changes involving the first codon position, unbiased weightings and scale means assigned to stop codons, showed the most significant expected changes in log odds for each one-unit increase in the second principal component (for optimization percentage values, p value:< 2.2e-16),

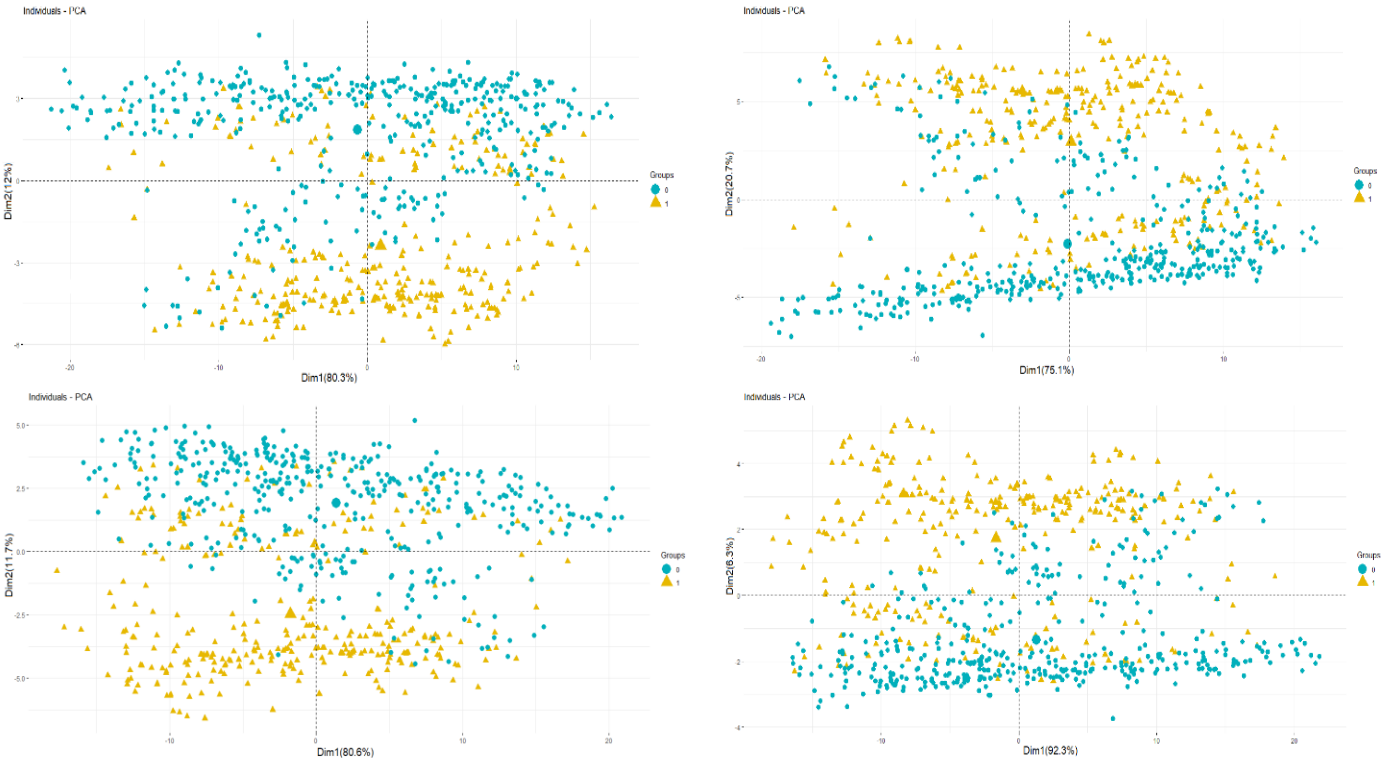


Figure 5.5: Principal component analysis of the Optimization percentages and scores for 84 amino acid properties and the model based on sense codons. Top left: The first two principal components for the Minimization percentages computed under genetic code models based on sense codons and unbiased weighting. Top right: The first two principal components for the Minimization percentages computed under genetic code models based on sense and biased weighting, Bottom left: The first two principal components for the scores computed under genetic code models based on sense and unbiased weighting. Bottom right: The first two principal components for the scores computed under genetic code models based on sense and biased weighting. Blue: Non-thermophilic prokaryotes (N=418), Orange: Thermophilic prokaryotes (N=324).

Table 5.10 The top 5 amino acid properties corresponding to the coefficients (coeff) with the smallest p values for the scores computed under the biased (MSW) and unbiased (MS) weighted mean phenotypic changes and genetic code representation based on sense codons. These coefficients, sorted in order of increasing p values, belong to Three-level logistic mixed models whose dependent variables has two levels: thermophiles and non-thermophiles. The third and fourth columns contain the medians as well as the first and third quartiles (shown in parentheses) for the scores. se: standard error, AIC: Akaike Information criteria. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The reference of each amino acid index, in Table IX.1, Appendix IX).

	Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff(se)**	**pvalue**	**AIC**
MS	Long-range contacts (41-50) [p163]	-0.3835(-0.4420.-0.3277)	-0.2643(-0.3123, -0.2344)	-255.32(10.78)	2,20E-16	257.26
	Long-range contacts (21-30) [p161]	-0.2402(-0.2888, -0.1950)	-0.1343(-0.1814, -0.0959)	-186.4(40.81)	2,20E-16	267.58
	Long-range contacts (11-20) [p160]	-0.5963(-0.6665, -0.5414)	-0.4723(-0.5260, -0.4326)	-84.26(20.64)	4,74E-12	296.67
	Long-range contacts (>50) [p164]	-1.1371(-1.2489, -1.0059)	-1.0125(-1.1121, -0.9047)	-33.11(11.53)	5,51E-08	314.93
	Thermodynamic stability [p177]	-0.2588(-0.2972, -0.2135)	-0.2365(-0.2655, -0.1849)	-137.66(40.76)	1,99E-07	317.42
MSW	Long-range contacts (31-40) [p162]	-1.2300(-1.3808, -1.0963)	-0.9895(-1.1125, -0.8781)	-25.53(8.85)	8,16E-08	315.69
	Long-range contacts (21-30) [p161]	-2.0717(-2.3010, -1.8509)	-1.7714(-1.9566, -1.5589)	-12.91(4.96)	1,46E-06	321.26
	Long-range contacts (41-50) [p163]	-2.1552(-2.3889, -1.9299)	-1.8456(-2.0289, -1.6252)	-12.06(4.74)	2,81E-06	322.52
	Long-range contacts (11-20) [p160]	-2.8946(-3.2306, -2.6047)	-2.6063(-2.8400, -2.2680)	-6.93(2.53)	3,37E-05	327.26
	Long-range contacts (>50) [p164]	-3.5130(-3.9060, -3.1783)	-3.2257(-3.4944, -2.8109)	-5.9(2.12)	3,72E-05	327.45

Table 5.11 The top 5 amino acid properties corresponding to the coefficients (coeff) with the smallest p values for the optimization percentages computed under the biased (MSW) and unbiased (MS) weighted mean phenotypic changes and genetic code representation based on sense codons. These coefficients, sorted in order of increasing p values, belong to Three-level logistic mixed models whose dependent variables has two levels: thermophiles and non-thermophiles. The third and fourth columns contain the medians as well as the first and third quartiles (shown in parentheses) for the scores. se: standard error, AIC: Akaike Information criteria. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The reference of each amino acid index, in Table IX.1, Appendix IX).

	Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff(se)**	**pvalue**	**aic**
MS	Long-range contacts (41-50) [p163]	31.66(30.03, 33.03)	29.00(27.86, 30.21)	306.1(61.1)	1,79E-11	299.27
	Long-range contacts (21-30) [p161]	31.91(30.48, 33.37)	29.48(28.40, 30.57)	362.27(58.38)	3,07E-11	300.33
	Long-range contacts (11-20) [p160]	37.31(0.361, 0.386)	35.01(33.58, 36.31)	188.47(47.98)	1,04E-06	320.59
	Long-range contacts [p165]	44.72(43.08, 46.12)	43.28(41.52,44.69)	157.72(51.89)	4,94E-06	323.60
	Conformational Entropy [p100]	28.33(27.07, 29.82)	26.60(24.87, 27.86)	124.91(46.91)	5,11E-05	328.05
MSW	Long-range contacts (31-40) [p162]	48.86(47.40, 49.85)	44.20(42.59, 46.65)	175.44(49.29)	4,27E-07	318.89
	Long-range contacts (21-30) [p161]	58.06(56.90, 59.11)	54.11(52.61, 56.07)	177.28(52.45)	1,61E-06	321.45
	Long-range contacts (41-50) [p163]	58.75(57.25, 60.22)	54.95(53.06, 56.86)	131.17(47.12)	9,01E-06	324.75
	Long-range contacts [p165]	72.03(70.89, 73.08)	69.58(67.94, 70.84)	145.65(51.57)	7,05E-05	328.66
	Long-range contacts (11-20) [p160]	68.66(67.25, 69.57)	65.44(63.49, 67.15)	111.88(40.05)	0,0002826	331.27

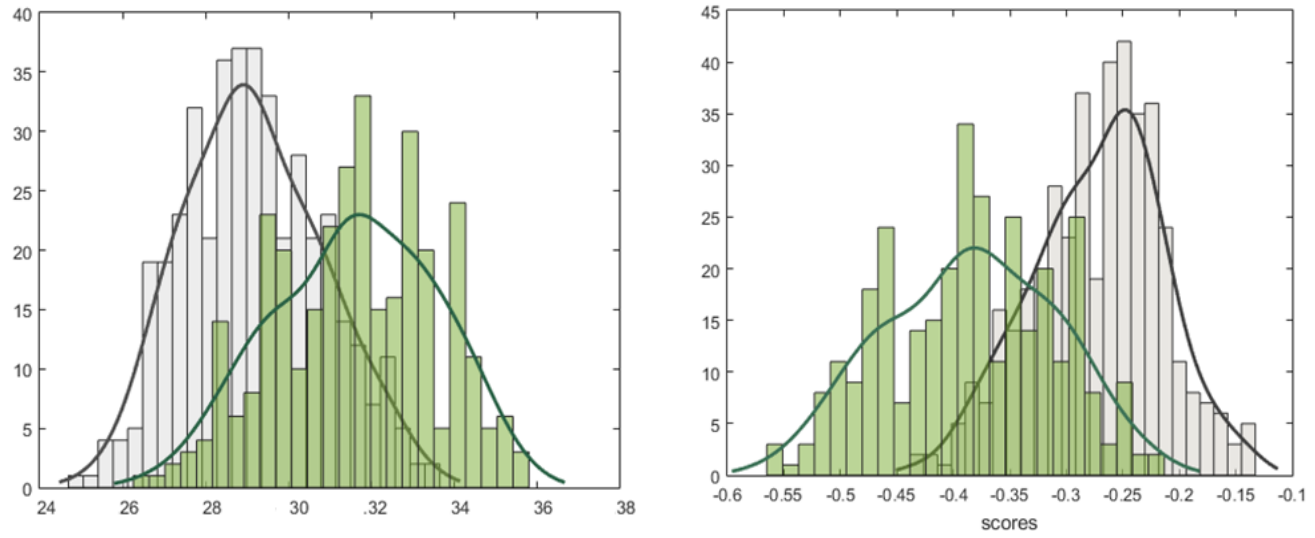


Figure 5.6 Histograms of the scores and Optimization percentages computed under different standard code models and amino acid properties. Green: Thermophilic prokaryotes (N=324), Grey: Non-thermophilic prokaryotes (N=418). A: The standard code models based on sense codons with unbiased weighting and long-range contacts (p163). C: The standard code models based on sense codons with unbiased weighting and Long-range contacts (p161).

among all representations explored for codon positions and substitution types (Transitions and transversions). The first two principal components computed for both relevance measures according to the previously mentioned first codon position representation, showed the clearest distinction between both thermal categories (figure 5.7, tables V.7-V.10 in Appendix V). The second most significant expected change in log odds were observed, also, for the first codon position

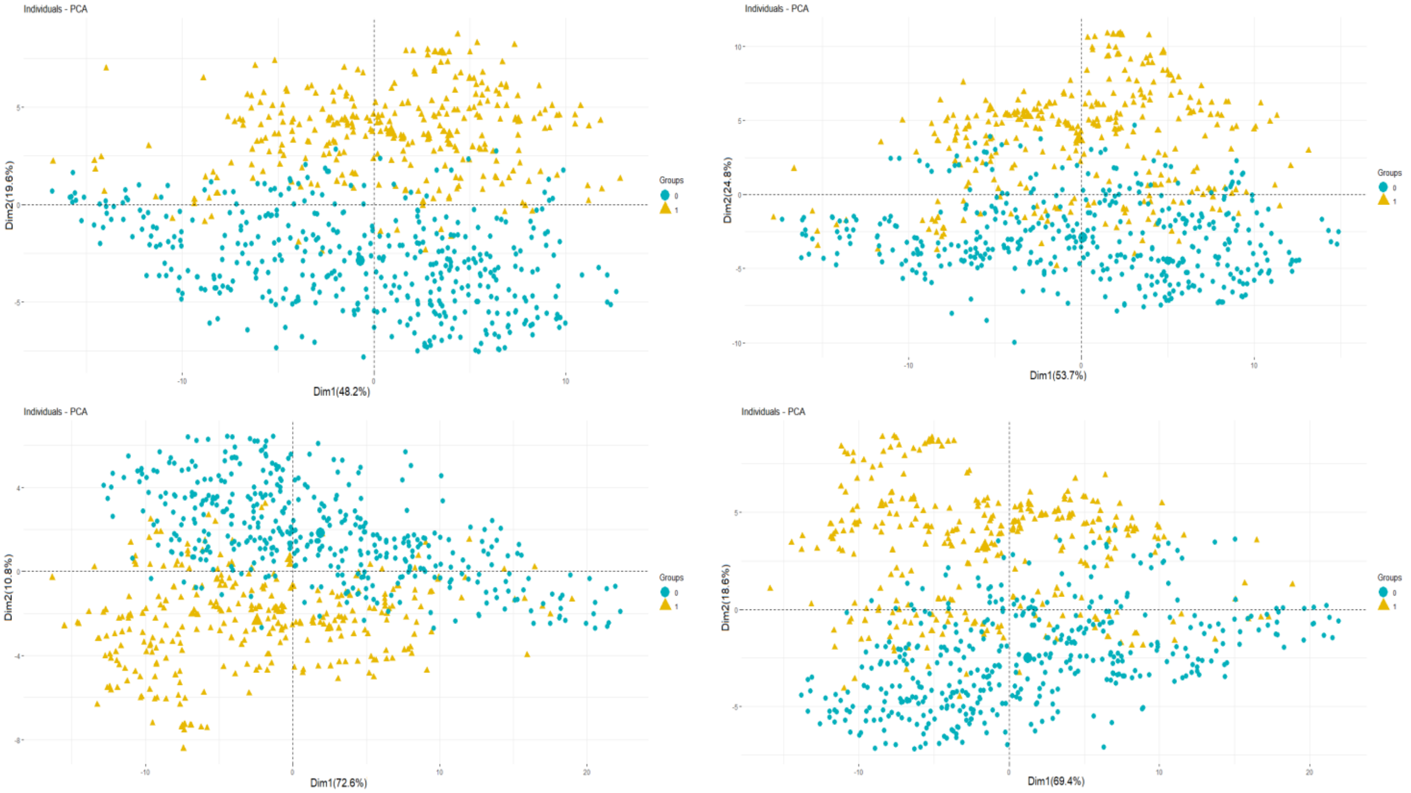


Figure 5.7: Principal component analysis of the Optimization percentages and scores for 84 amino acid properties and four partial standard code models. Top left: The first two principal components for the Optimization percentages computed under the first codon position models based on the whole set of codons with scale mean values assigned to stop codons and unbiased weighting. Top right: The first two principal components for Optimization percentages computed under models using biased weighting and scale mean values assigned to stop codons, Bottom left: The first two principal components for the scores computed under the first codon position models with scale mean values assigned to stop codons and unbiased weighting. Bottom right: The first two principal components for the scores computed under models using biased weighting and scale mean values assigned to stop codons. Blue: Non-thermophilic prokaryotes (N=418), Orange: Thermophilic prokaryotes (N=324).

representations but in this case for those based on sense codons (second principal components (p value): 2.587e-11 and 1.867e-07 for biased and unbiased weightings, respectively). The transversions was the substitution type for which the best separation was observed between thermophilic and non-thermophilic prokaryotes. This result was obtained for the mean change in the conformational entropy computed under the unbiased weighting (P=8.88E-12, Figure 5.9, tables V.15-V.18 in Appendix V).

The proteins of thermophiles are distinguished from those of non-thermophiles on the basis of some factors, such as, loop length, hydrophobic core compactness, aromatic side-chain stacking, salt bridges, hydrogen bonds, flexibility, conformational entropy, among others [51, 52 ,53]. In general, we observed that for the amino acid properties linked to three of these factors, namely, average long-range contacts (hydrophobic compactness), flexibility and conformational entropy, the genomic robustness values tend to be more relevant in thermophiles than in non-thermophiles (Appendix V). The thermophilic proteins are characterized by their rigidity and the compactness of their hydrophobic cores. The hydrophobic interactions in the protein interior are critical determinants of protein tertiary structure and stability [124]. Other studies also showed that the long-range contacts are important for stabilising the thermophilic proteins and the transition state structures of folded proteins [125]. Some evidences have pointed out that the conformational entropy is connected to an enhanced thermal stability. For example, the thermophilic proteins have an unfolded state with a reduced entropy and a

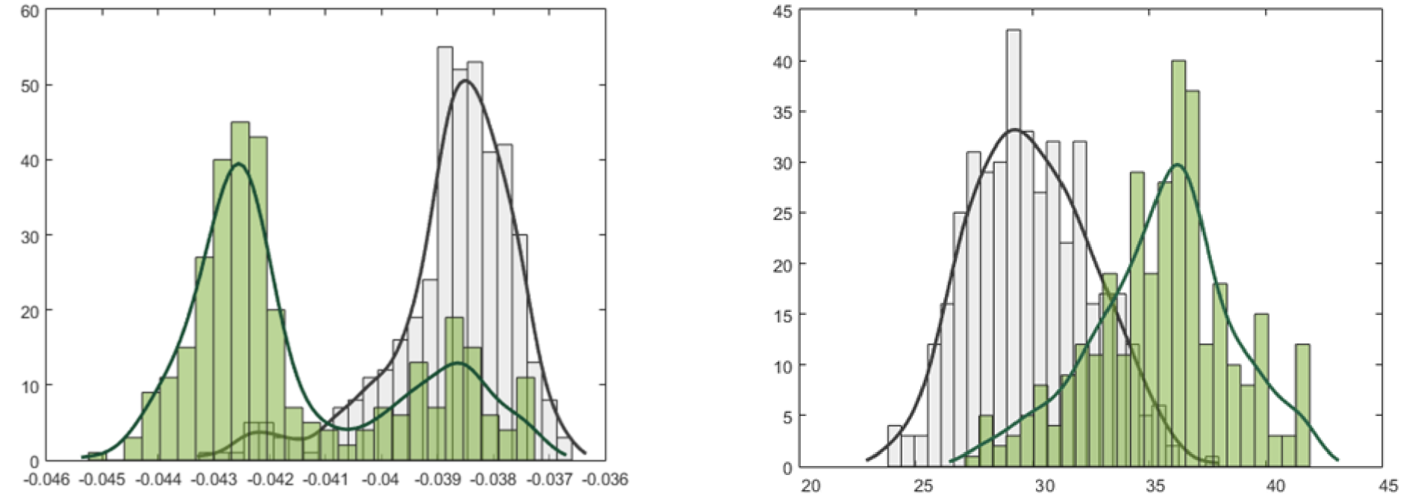


Figure 5.8. Histograms of the scores and Optimization percentages computed under different standard code models and amino acid properties. Green: Thermophilic prokaryotes (N=324), Grey: Non-thermophilic prokaryotes (N=418). B: The first codon position models based on the whole set of codons with unbiased weighting, Long-range contacts (p165) and scale means assigned to stop codons., B: The transversion models based on the whole set of codons with unbiased weighting, Conformational entropy (p102) and “Mean suppressor” values assigned to stop codons. (The reference of each amino acid index, in Table IX.1, Appendix IX).

residual structure more compact than mesophilic proteins [126, 127]. The conformational entropy is also closely related to flexibility. Besides being linked to thermal stability, the flexibility plays a crucial role in protein function [52]. Therefore, these amino acid properties are among the best-preserved ones by the genomic codon usage bias in thermophiles probably because they are strongly linked to the stability of thermophilic proteins.

The distribution of the codon usage bias measured by information entropy is essentially the same in both groups of genomes. In contrast, the differences between both growth temperature range groups are clearly visible in the histograms of scores and optimization percentage computed under several properties (histograms V.1-V.3 in Appendix V), These differences are because the more robust codons tend to be much more frequent, at the expense of the least robust codons, and this association between codon-choice patterns and robustness is stronger in thermophiles than in non-thermophiles. We could conclude that selective, or/and neutral, pressures influence the evolution of codon usage bias in such a way that the effect of the single-base changes, with respect to amino acid

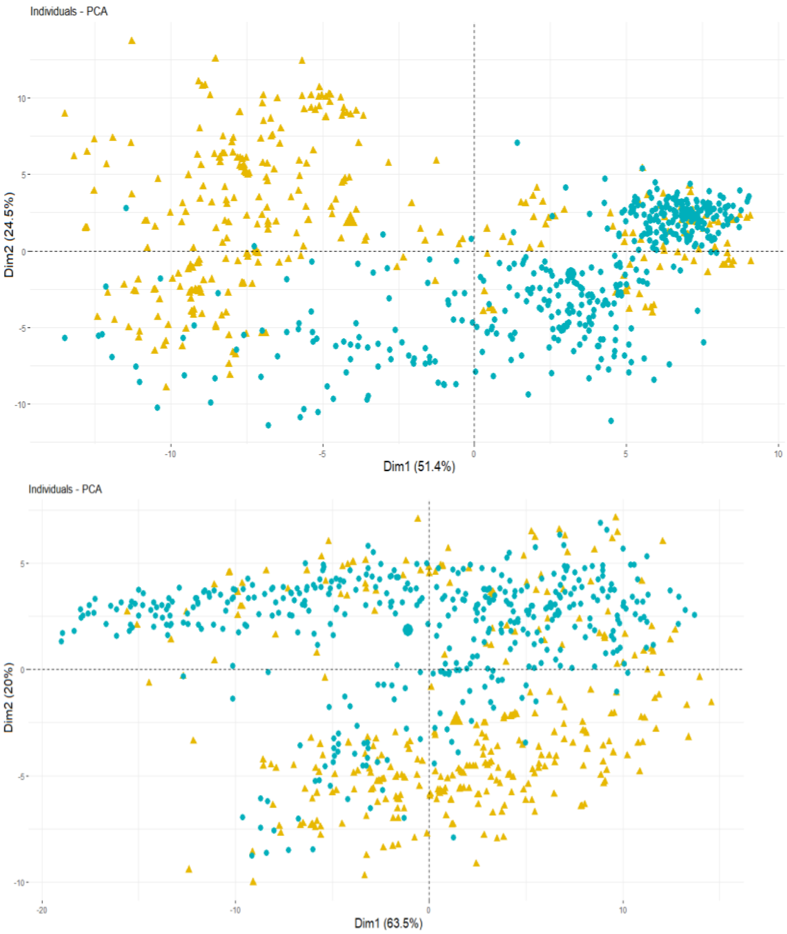


Figure 5.9 Principal component analysis of the scores and Optimization percentages for 84 amino acid properties and two partial genetic code models with unbiased weightings, Top: The first two principal components for scores computed under the transversion model based on sense codons, Bottom The first two principal components for the Optimization percentages computed under the transversion model based on the whole set of codons and scale means assigned to stop codons, Blue: Non-thermophilic prokaryotes (N=418), Orange: Thermophilic prokaryotes (N=324).

properties like the average long-range contacts and hydrophobicity, is more efficiently minimized in thermophilic prokaryotes than in non thermophilic prokaryotes.

Concerning the relationship between robustness and base composition for each codon position, we observed no separation between thermophiles and non-thermophiles at the third position (figure VI.2, Top left, in Appendix VI), corroborating previous evidences pointing to a weak selective pressure at this codon position[27]. As for the genomic robustness scores involving the first codon position, a separation was seen between the thermophiles and non-thermophiles, mainly, for the amino acid scales such as, average long-range contacts and hydrophobicity (figures VI.1 and VI.2, Appendix VI). The observed split-up between both groups of genomes at these codon positions but not at the third position, is consistent with the above results obtained by using the Three-level logistic mixed models.

5.3.3 Genomic robustness at the codon block level

The difference between the genomes of thermophiles and non-thermophiles with respect to the codon-block robustness could shed light on the relationship between the thermal adaptation and the ability to minimize the effect of errors at synonymous codon block level. For this analysis, the codon block robustness was computed as the mean phenotypic change in the average long-range contacts for the model based on sense codons using the biased and unbiased weightings (section 5.3.2). The above-mentioned property and genetic code representation were selected because of their strong discriminative power between thermophiles and non-thermophiles (table 5.10). Only the heterogeneous blocks were considered because only these blocks contribute to the difference between both groups with respect to genomic robustness (see section Methods). The more robust

[27] [27] The base composition at the third codon position is rather the result of the neutral mutational pressure because at this position occur most of the synonymous single-base changes.

the codon block, the stronger the association between the usage frequency and the robustness of its codons.

The codon blocks of the amino acids R, P, K, L and V showed to be significantly more robust in thermophiles than in non-thermophiles for average long-range contacts (tables 5.12 and 5.13, see tables VI.1 and VI.2 Appendix VI). The codon block of the amino acid Leucine (L) was observed to be significantly more robust in thermophiles than in non-thermophiles but only for the biased weightings. Since the thermophiles were found to be significantly more robust than non-thermophiles according to the unbiased and biased-weighted mean phenotypic changes for all codon blocks (table 5.14), we could conclude that the contributions to genomic robustness from the codon blocks encoding for the amino acids, R, K, P, V and L in thermophiles exceed those of the blocks more robust in non-thermophiles (Tables 5.12 and 5.13).

Table 5.12: Medians of the Unbiased-weighted mean change (UMC) in long-range contacts (p163) for each synonymous codon block (second and third columns). The first and the third quartiles are shown in parentheses. The first column: the amino acids encoded by codon blocks containing at least two codons with different robustness (hb). Model of the standard code based on the sense codons. The smaller UMC medians of both groups are shown in red letters. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). P values: Wilcoxon rank-sum test. False discovery rate:0.01, *: Significant.

hb	Thermophiles	Non-thermophiles	P values
A	0.1605(0.1600, 0.1609)	0.1606(0.1602, 0.1610)	0.0238
R*	0.3654(0.3565, 0.3816)	0.4417(0.4180, 0.4627)	2.4038e-90
G*	0.6728(0.6700, 0.6766)	0.6633(0.6601, 0.6694)	2.5050e-54
I*	0.9105(0.8721, 0.9250)	0.8467(0.8348, 0.8669)	1.3177e-47
L	0.4641(0.4581, 0.4691)	0.4646(0.4569, 0.4773)	0.0094
K*	0.4799(0.4136, 0.5312)	0.5256(0.3860, 0.5602)	0.0021
P*	0.0542(0.0512, 0.0558)	0.0552(0.0532. 0.0573)	1.1032e-09
S*	0.5525(0.5445, 0.5646)	0.5478(0.5399, 0.5551)	1.5879e-09
T*	0.1055(0.0984, 0.1132)	0.1015(0.0946, 0.1088)	6.9170e-09
V*	0.9503(0.9440, 0.9612)	0.9565(0.9466, 0.9675)	1.9157e-06

This finding implies that these codon blocks are the most efficient to reduce the effect of errors among all codon blocks in thermophilic genomes. The stronger ability of these codon blocks to minimize the effect of single-base changes is compatible with the important role played by their corresponding amino acids in thermophilic adaptation. Both factors, namely, the involvement in thermal adaptation and, to a lesser extent, the contribution to genomic robustness of these blocks could explain the high frequency of these amino acids observed in thermophilic genomes (table VI.1 in appendix VI). The higher frequency of the amino acids R, K, P, L and V has been considered as characteristic features of hyperthermophilic or thermophilic proteins[28]. [46, 126, 128, 129].

Table 5.13: Medians of the Biased-weighted mean change (BMC) in long-range contacts (p163) for each synonymous codon block (second and third columns). The first column: the amino acids encoded by codon blocks containing at least two codons with different robustness (hb). Model of the standard code based on the sense codons. The first and the third quartiles are shown in parentheses. The smaller BMC medians of both groups are shown in red letters. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). p values: Wilcoxon rank-sum test. False discovery rate:0.01,*: Significant.

hb	**Thermophiles**	**Non-thermophiles**	**p values**
A	0.0670(0.0670, 0.0671)	0.0671(0.0670, 0.0671)	0.0239
R*	0.1890(0.1801, 0.2235)	0.2887(0.2577, 0.3093)	4.3323e-76
G*	0.3399(0.3374, 0.3433)	0.3315(0.3286, 0.3369)	2.5050e-54
I*	0.3406(0.3368, 0.3421)	0.3342(0.3330, 0.3363)	1.3143e-47
L*	0.2214(0.2213, 0.2217)	0.2217(0.2215, 0.2223)	1.6615e-22
K*	0.1041(0.0975, 0.1092)	0.1087(0.0947, 0.1121)	0.0021
P*	0.0124(0.0121, 0.0125)	0.0125(0.0123, 0.0127)	1.1044e-09
S	0.2422(0.2408, 0.2435)	0.2423(0.2409, 0.2431)	0.3237
T*	0.0383(0.0352, 0.0415)	0.0378(0.0349, 0.0399)	0.0054
V*	0.3895(0.3837, 0.3997)	0.3950(0.3855, 0.4059)	4.9679e-06

It is known the high propensity of the charged amino acids, mainly the amino acid R, to participate in stabilizing salt-bridge interactions (ion pairs), specially, in the exposed part of thermophilic proteins. The amino acids, V and L, contribute to the hydrophobic interactions that play a crucial role in protein stability. The pyrrolidine ring of the amino acid, P, reduces the backbone conformational entropy of the unfolded state of the protein, which increases, in turn, the protein stability by decreasing the entropic difference between the folded and unfolded states [128].

Table 5.14: Genomic robustness (medians) of thermophiles and non-thermophiles defined as the Biased (BMC) and unbiased (UMC) weighted mean squared changes in long-range contacts (p163) under the model based on sense codons. The first and third quartiles are shown in parenthesis. p values: Wilcoxon rank-sum test.

	Thermophiles	**Non-thermophiles**	**P values**
BMC	0.1735(0.1720,0.1785)	0.1842(0.1806,0.1875)	6.6869e-70
UMC	0.4127(0.4110, 0.4153)	0.4197(0.4161,0.4239)	1.2336e-64

[28] The amino acid K are more frequent in hyperthermophilic proteins than in mesophiles but less frequent in thermophilic proteins than in mesophilic proteins. In the case of the amino acid, P, it is less common in hyperthermophilic proteins than in mesophilic proteins but more frequent in thermophilic proteins than in mesophilic proteins [126].

5.3.4 Codon robustness and synonymous codon usage

Significant differences have been observed between thermophiles and mesophiles with respect to synonymous codon frequency [48, 130]. It has been previously observed that the ability of codons to minimize the effect of errors explains, to some extent, these differences [50]. We explored the relationship between the codon robustness and these differences for the heterogeneous block codons. The codon robustness was computed by using the model with the best discriminative performance between both thermal categories, namely, the unbiased weighted mean change in average long-range contacts (p161) under the genetic code representation based on sense codons.

Significant differences were found between the thermophilic (thermophiles + hyperthermophiles) and non-thermophilic (mesophiles + psychrophiles) prokaryotes with respect to the usage frequency of some codons. The codons that have higher occurrence in thermophilic prokaryotes are, CUU (L), CUC (L), CUA(L), UCU(S), UCC(S), UCA(S), CCA(P), AGA(R), AGG(R), AUA (I), ACA (T), AAG(K), GUA(V), GUU (V), GCA(A), GCU (A), GGA(G), GGG(G). On the other hand, the more commonly used codons in non-thermophilic prokaryotes compared to the other group are, CUG (L), UCG(S), CCG (P), CGU(R), CGC(R), CGA(R), CGG(R), AUU(I), AUC(I), ACC(T), AAA(K), GUG(V), GUC(V), GCG(A), GCC(A), GGU(G), GGC(G) (table 5.15).

We found that 5 of the 6 codons for the arginine, were among the 10 top codons that showed the most significant differences in usage frequency between both groups of genomes (CGU, CGC, CGA, AGA and AGG). Two of these codons are among the 4 least robust codons, namely, CGU and CGC, and are much less frequent in thermophiles than in non-thermophiles, while the codons, AGG and AGA, which are more robust than the above two codons, are much more frequent in thermophiles compared to non-thermophiles (table 5.15). This explains why, when compared both groups of genomes with respect to the codon block for the arginine, the thermophiles showed to be much more robust than non-thermophiles. (tables 5.12 and 5.13).

Generally, a significant increase in the frequency of usage of at least one of the most robust codons at the expense of the least robust codons for a given block in thermophiles, will result in an increase in the robustness of this codon block in that group of genomes relative to the other. As described above, this is especially true for the set of synonymous codons specifying the arginine. This pattern was also observed, to a lesser extent, for the codon blocks corresponding to the amino acids, P, V and L. As for the amino acid, proline, it was seen that one of its two least robust codons, namely, CCG, showed to be remarkably less frequent in thermophiles than in non-thermophiles. This codon is less frequently used in thermophiles than the other more robust codons for this amino acid (CCU, CCC) compared to non-thermophiles. On the other hand, we observed that the codons, GUU and CUU, encoding for the amino acids, V and L, are more robust and more frequently used in thermophiles than their corresponding synonymous codons, GUG and CUG (table 5.15).

The robustness of synonymous codons is strongly associated with their usage frequencies at the genome level, mainly, in thermophilic prokaryotes. Our results indicate that the higher robustness of the thermophilic prokaryotes compared to that observed for non-thermophilic prokaryotes is mainly due to the heterogenous codon blocks corresponding to some of the most frequent amino acids in thermophilic or hyper-thermophilic proteins.

Table 5.15 The Unbiased-Weighted Mean change in long-range contacts (p163) for each codon (LR). LR was computed by using the standard code representation based on sense codons and a weighting with the synonymous codon frequency. The contiguous cells of the third column with the same color indicate codons forming homogeneous sub-blocks. the fourth and fifth columns show The Medians of the synonymous codon usage frequencies as well as the first and third quartiles in parentheses. The first column: The amino acids encoded by the heterogeneous blocks of the standard code. Non-Thermophiles: Non-Thermophilic prokaryotes (N=418), Thermophiles: Thermophilic prokaryotes (N=324). p values: Wilcoxon rank-sum test. *: Significant for a False discovery rate:0.01.

Amino Acids	Codons	LR	thermophiles	Non-thermophiles	p values
L	**UUA**	0.6606	0.1521(0.0427,0.2948)	0.1799(0.0111.0.3893)	0.6713
	UUG	0.3847	0.1395(0.1017,0.1963)	0.1439(0.0890,0.2042)	0.8968
	CUU*	0.6179	0.1908(0.1432,0.2384)	0.1209(0.0983,0.1600)	1.41E-31
	CUC*	0.6179	0.1488(0.0636,0.2515)	0.0936(0.0549,0.1815)	3.09E-05
	CUA*	0.8245	0.0683(0.0419,0.1030)	0.0620(0.0183,0.1080)	6.12E-04
	CUG*	0.6526	0.1570(0.0743,0.2679)	0.2314(0.0746,0.5172)	7.13E-06
S	**AGU**	1.6924	0.1479(0.0991.0.2023)	0.1705(0.0666,0.2348)	0.3163
	AGC	1.6924	0.2002(0.1409,0.2733)	0.2294(0.1445,0.3013)	0.0328
	UCU*	1.3973	0.1660(0.0892,0.2363)	0.1598(0.0503,0.2307)	0.0035
	UCC*	1.3973	0.1607(0.0895,0.2162)	0.1274(0.0643,0.2015)	7.41E-05
	UCA*	0.1517	0.1592(0.1021.0.2397)	0.1414(0.0570,0.2193)	3.76E-05
	UCG*	0.1816	0.1112(0.0593,0.1649)	0.1305(0.0829,0.2644)	1.95E-07
P	**CCU**	0.1301	0.2440(0.1477,0.3554)	0.2528(0.0926,0.3558)	0.0246
	CCC	0.1301	0.2308(0.1033,0.3148)	0.1741(0.0965,0.3127)	0.024
	CCA*	0.2551	0.2454(0.1402,0.3890)	0.2094(0.0774,0.3817)	0.0022
	CCG*	0.2551	0.2152(0.1039,0.3566)	0.2898(0.1418,0.5179)	1.14E-07
R	**CGU***	2.1368	0.0593(0.0382,0.1036)	0.2291(0.1282,0.3411)	6.29E-70
	CGC*	2.1368	0.0582(0.0279,0.2180)	0.3363(0.1692,0.5189)	1.08E-40
	CGA*	0.5428	0.0437(0.0230,0.0657)	0.0757(0.0448,0.1230)	2.75E-26
	CGG*	0.6231	0.0411(0.0183,0.1789)	0.1042(0.0414,0.1841)	7.36E-07
	AGA*	1.1619	0.3188(0.1121.0.5554)	0.0817(0.0182,0.1984)	1.46E-27
	AGG*	0.6578	0.2615(0.1616,0.3930)	0.0421(0.0266,0.0643)	3.00E-73
I	**AUU***	1.7514	0.3375(0.2075,0.4362)	0.4711(0.2048,0.5743)	6.07E-10
	AUC*	1.7514	0.2418(0.1322,0.4051)	0.3769(0.2187,0.7402)	1.16E-14
	AUA*	2.2895	0.4339(0.2334,0.5098)	0.1009(0.0387,0.2065)	1.32E-47
T	**ACU**	0.3315	0.2061(0.1198,0.2892)	0.1982(0.0672,0.2971)	0.091
	ACC*	0.3315	0.2711(0.1547,0.4051)	0.3532(0.2092,0.5191)	1.25E-07
	ACA*	0.5311	0.2827(0.1558,0.4103)	0.1879(0.0686,0.3196)	1.74E-10
	ACG	0.2829	0.1914(0.1001.0.3032)	0.2084(0.1454,0.2977)	0.0404
K	**AAA***	1.2751	0.5524(0.3655,0.6972)	0.6814(0.2874,0.7790)	0.0021
	AAG*	0.5663	0.4475(0.3027,0.6344)	0.3185(0.2209,0.7125)	0.0021
V	**GUU***	1.8816	0.3261(0.2244,0.4248)	0.2615(0.1053,0.3772)	1.37E-08
	GUC*	1.8816	0.1714(0.0880,0.2746)	0.2067(0.1117,0.3310)	7.25E-04
	GUA*	1.9094	0.2169(0.1102,0.3048)	0.1759(0.0554,0.3011)	5.40E-04
	GUG*	2.234	0.2426(0.1712,0.3659)	0.3077(0.1907,0.4414)	5.30E-06
A	**GCU***	0.4773	0.2514(0.1591.0.3363)	0.2366(0.0832,0.3205)	6.34E-05
	GCC	0.4773	0.2661(0.1390,0.3926)	0.2745(0.1650,0.4460)	0.1009
	GCA*	0.5068	0.2544(0.1574,0.3884)	0.2335(0.1108,0.3483)	2.24E-04
	GCG*	0.5068	0.1623(0.0943,0.2745)	0.2313(0.1439,0.3393)	5.26E-10
G	**GGU***	1.2619	0.2513(0.1750,0.3258)	0.2985(0.1642,0.3908)	6.31E-04
	GGC*	1.2619	0.2276(0.1275,0.3598)	0.3726(0.2134,0.5814)	7.37E-19
	GGA*	1.0564	0.3036(0.1968,0.4172)	0.1224(0.0785,0.2910)	3.56E-31
	GGG*	0.9547	0.1839(0.1292,0.2312)	0.1282(0.1011.0.1675)	2.25E-24

CHAPTER 6

DISCUSION

We observed that, among 235 amino acid indices, the 5 amino acid properties linked to the most relevant values of genetic code robustness, following the hydrophobicity/polarity, are the solvent accessible surface area, average long-range contacts, flexibility, Transmembrane helix and Small-linker propensities (see section 5.3.3). The amino acid property best preserved was the hydrophobicity/polarity. More specifically, the most relevant scales were, the Kyte's hydropathy index and Miyazawa's hydrophobicity, for the codon-based models and Polar requirement for the block-based models. It is interesting that the block-based model is precisely the representation used in previous works that apply, however, another estimation method based on randomly generated codes [25].

The standard genetic code was among the first three most robust natural codes with respect to the Cantelli's upper bounds computed by using the Miyazawa's hydrophobicity scale and different genetic code representations. This result is in agreement with previous studies by showing that the increase of robustness is not linked to the emergence of most alternative genetic codes [120,131]. We found that the Ciliate, Dasycladacean and Hexamita Nuclear Code was the most robust code or the third most robust code for two code representations, which is consistent with previous findings [131, 132]. We also corroborated that codon reassignments that create the Alternative yeast nuclear Code and Yeast mitochondrial code from the standard genetic code, are linked to the largest decrease in robustness [131]. The three codon positions have not the same contributions to genetic code robustness. Overall, the third codon position is the most robust codon position, followed by the 1st and 2nd positions in that order. These results agree with findings reported by other authors [79]. We also observed that the standard genetic code is more robust than most alternative genetic codes at the first codon position and with respect to transitions or transversions.

We also found that some nuclear codes are more robust than the standard genetic code at the first and second codon positions and most mitochondrial genetic codes are more robust than the standard code at the third codon position. The observation that some alternative genetic codes are more robust than the standard code at the first and third codon positions but not with respect to the whole code structure, suggests that the partial increment of robustness could have been an important factor in the fixation of codon reassignments giving rise to the emergence of new variant genetic codes. This is interesting because previous studies only considered the robustness computed from whole genetic code representations [120,131, 132].

In general, our results indicate that the natural genetic codes are more robust for amino acid properties strongly related to the protein stability and hint at the possibility that the load-minimization property might be an adaptation that takes place, mainly, at the level of one or two of the codon positions in alternative genetic codes. This adaptation involving the codon positions for which the protein translation errors are more frequent could be an important factor in the recent evolution of the standard genetic code and probably also played a crucial role at the transition from the RNA world to the modern DNA/RNA/protein world, when the fidelity of a primitive translation system was still very low (For more details see sections 2.1, 2.2 and 2.3.2) [39, 73].

We showed that there are two types of sets of synonymous codons in the standard genetic code, the homogeneous and heterogeneous codon blocks. To study the relationship between the robustness and the frequency of synonymous codons at the codon block level, only the heterogeneous codon blocks must be considered because only such blocks comprise at least two codons with different robustness values. We found that in the standard genetic code the set of heterogeneous codon blocks corresponds to the 10 amino acids, A, G, R, V, L, K, S, I, P and T for both weightings used and regardless of the amino acid property. Substituting two of these amino acids, K and R, for the acidic amino acids, D and E, leads to the known set of primitive amino acids [54]. This characteristic of the standard code linked to robustness might be a by-product or, on the contrary, another important factor in the evolution of the primordial genetic code to its present form [39]. The possibility of maximizing, at the level of the synonymous codon usage, the robustness to errors for the first amino acids incorporated into the genetic code could have been advantageous for primitive organisms with highly error-prone machineries for protein synthesis.

The synonymous codon usage is not random and, frequently, the alternative codons for the same amino acid occur with different frequencies. The synonymous codon usage pattern is unique to each species and is the result of a balance between neutral mutational processes and natural selection. Several factors have been shown to be related to the synonymous codon usage [3, 122,123]. The robustness to errors and growth temperature range group are two of these factors. Previous studies have shown that the synonymous codon usage and robustness to errors are correlated at the genome and gene level [40, 41, 43, 44, 50, 85, 86, 87]. Whereas, other studies have shown that in prokaryotic genomes there is no bias in the synonymous codon usage towards a higher frequency of the more robust codons [35, 42, 45]. However, it is not possible to know with certainty from these studies what is the general trend in prokaryotes, either because they used less than four prokaryotic genomes [35,43] or because they used a subset of genes representing complete genomes [42, 45] (For more details see section 2.4).

Regarding the thermophilic adaptation at the coding sequence level in prokaryotes, it is known that certain amino acids and codons tend to occur with high frequency in thermophilic prokaryotes [46,48, 49, 128, 129,130] and why some amino acids could confer selective advantages in high-temperature environments [128], but little is known about why certain synonymous codons tend to be used more frequently than others. To date, only one study has been conducted exploring the relationship between these synonymous codon usage preferences and the robustness to errors [50]. The authors concluded that the differences in the usage of synonymous codons between the mesophilic and thermophilic prokaryotes could be explained in terms of the synonymous codon usage robustness to errors and mRNA secondary-structure stability. However, in this study a small sample was used, the translational errors were not considered and in addition, the Mclachlan matrix, which is an amino acid substitution matrix, was applied, so no information was provided on the properties of the amino acids involved [50].

To assess the robustness of the synonymous codon usage, computed from complete genome sequences, a sample of thermophilic (324) and non-thermophilic prokaryotes (418) was chosen. The mean phenotypic change weighted with the relative frequency of synonymous codons was applied as measure of the ability to mitigate the impact of errors at the level of proteins. The optimization percentage and scores were computed for 84 amino acid indices, two weightings and genetic code representations as well as under three different methods to process the stop codons.

We have found that the synonymous codon usage of prokaryotic genomes is highly optimized to buffer the impact of errors. The highest degrees of error mitigation were observed for hydrophobicity and the other amino acid properties linked to protein folding and stability, indicating that the higher the robustness of synonymous codons in terms of these properties, the larger their frequencies at the genome level. The fact that several hydrophobicity indices and other properties related to protein stability are linked to the highest robustness values could be interpreted as an evidence of the action of selective pressures shaping the synonymous codon usage to minimize the effect of errors in proteomes. The robustness values for the weighting based on mistranslations were larger compared to those observed for the unbiased weighting. It is well known the high cellular cost of protein misfolding and aggregation caused by mistranslations and environmental factors like temperature. Consequently, increasing the frequency of the most robust codons at the expense of the least robust codons could be an evolutionary response to the observed high mistranslation rates [3, 133, 134, 135, 136]. Mitigating the effect of mistranslations would reduce the amount of severely misfolded proteins, thereby, facilitating the activity of the quality-control system based on chaperones and proteases constrained by metabolic costs [3, 134, 137].

We observed that synonymous codon usages of thermophilic genomes tend to be more robust to errors than those of non-thermophiles [50]. The thermostability and mutational robustness are known features of thermophilic proteins [90, 91]. Our results suggest that the high degree of optimization of synonymous codon usages in thermophilic genomes could be another mechanism to withstand high temperatures. Thus, the increased use of the most robust codons, especially at critical sites for protein stability could represent a general protection mechanism against the consequences of high rates of translation errors [91, 138], as was observed in conserved ligand-binding sites [85]. This observation explains why the codon block most optimized in thermophilic genomes to mitigate the effect of errors corresponds to arginine, one of the most important amino acids for thermostability (see section 5.3.3). It could also explain why we detected significantly higher robustness in thermophiles relative to non-thermophiles for the blocks corresponding to other amino acids, K, P, V and L, known to frequently occur in thermophilic and hyperthermophilic proteins [128, 129] (see section 5.3.3).

The high frequency of use of codons AGR (AGG, AGA) and the low frequency of codons CGY (CGU, CGC), are typical features of thermophilic genomes [46,48,49,130]. Our results indicate that the codons AGR are more frequent in thermophiles because they are more robust than the less frequent codons CGY. We also observed significant association between codon robustness and frequencies, for the codons, CCA (P), GUU (V) and CUU (L), which turned out to be more robust and more frequent than their respective synonymous codons, CCG, GUG and CUG. We also found that the association between the synonymous codon usage robustness and temperature range groups is stronger for the average long-range contacts, mainly, at the first codon position. The most significant differences in base composition between thermophiles and mesophiles among the three codon positions have been observed at this codon position (for the bases, A and C) [46].

The information flow from DNA to proteins has multiple error-prone steps. For example, considering only the protein synthesis which has one of the highest error rates, it was estimated that the 15% of average-length proteins would contain at least one amino acid substitution. There two main mechanisms or forces in the evolution of coding sequences, one relies on the error prevention and removal and the other, on error mitigation. The first mechanism is seen, for example, in substrate selection and proofreading mechanisms of DNA and RNA polymerases and its effect is reducing the error rates. Whereas, the mechanism for error-mitigation consists of reducing the effect of errors, an example of this is the complex network of chaperones and proteases that target misfolded proteins for chaperone-assisted refolding or degradation, although there are other examples like the intrinsic robustness of proteins and duplicate genes, among others [3, 7, 8, 9, 10, 12, 133, 134, 135, 139]. Both mechanisms have also been observed at the level of synonymous codon usage, the mechanism for error prevention, in this case, entails increasing the frequency of the translationally optimal codons that minimize the rate of amino acid misincorporations and the mechanism for error mitigation is based on increasing the use of the codons that reduce the effect of single-base changes in proteins. We have demonstrated that the synonymous codon usage robustness, mainly, to translational errors is a general trend in prokaryotic genomes and that this trend is stronger in the codon blocks corresponding to some of the most frequent amino acids in thermophilic and hyperthermophilic proteins. These results can be considered as evidences of selection on synonymous codon usage for maximizing the robustness in prokaryotes. However, the possibility of being a by-product of other mutational or selective pressures could not be ruled out.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

We have introduced some modifications to the statistical and optimization-based approaches to assess the genetic code and genome robustness in order to improve their efficiency and accuracy. The two main improvements are: 1) A method based on the first two moments of the unknown distribution of the weighted mean phenotypic change values for all possible amino acid-to-codon assignments. The Cantelli's upper bound and scores, defined from both moments, were used as measures of relevance of the genetic code and genome robustness values. This is a distribution-free method because it does not rely on any distribution assumption. 2) Thanks to the identification of a polynomially solvable instance of the Quadratic Assignment Problem, an exact algorithm to find the minimum genome robustness value was applied to compute the optimization percentage. In addition, we performed other minor improvements that decrease the number of operations required to compute the weighted mean phenotypic change and its distribution parameters, such as, mean and variance. More precisely, these improvements are based on, 1) equations for efficiently computing the variance and mean, 2) partitioning the graph representation of the genetic code into two components in the following four ways: 2.1) The vertices of one component represent the codons that belong to heterogeneous codon blocks and those of the other component represent the codons belonging to homogeneous codon blocks. 2.2) The vertices of one component represent the codons specifying different amino acids and those of the other component represent the codons that correspond to same amino acid. This partitioning is derived from the pairwise comparison between the amino acid-to-codon assignments corresponding to two genetic codes. 2.3) The vertices of one component represent sense codons and those of the other component represent stop codons. 2.4) One component contains the edges representing missense single-base changes and the other, the edges representing synonymous single base changes. All these modifications make the statistical and optimization-based approaches suitable for large-scale data analysis. For instance, the influence of the load minimization property of the genetic code on amino acid and synonymous codon usages can be efficiently tested by using these methods on large samples of biological sequences and amino acid properties. In addition, we have showed that the correspondence between our method and that based on the empirical sampling distribution of the weighted mean phenotypic change is high, mainly for amino acid properties related to the most relevant robustness values.

We applied these methods to answer two main questions: 1) *Is the increase in robustness to single-base changes important for the evolution of the alternative genetic codes?* 2) *Is the robustness to single-base changes important for the evolution of synonymous codon usage in prokaryotes?*

1) In general, our results indicate that several alternative genetic codes arise from the standard code not through codon reassignments that increase the robustness with respect to the entire code but by means of codon reassignments that increase the robustness with respect to the third and first codon positions, for the mitochondrial and nuclear genetic codes, respectively. Therefore, we can conclude that robustness with respect to substitution position could be important for the evolution of several alternative genetic codes. We have also found that robustness of the 23 natural genetic codes are highly optimized not only for hydrophobicity/polarity but also for other properties also linked to protein folding and stability, such as, solvent accessible surface area, average long-range contacts and flexibility.

2) We consider that robustness is important for the evolution of synonymous codon usage in prokaryotes based on three observations: The robustness is strongly associated with the frequency of

synonymous codon usage in prokaryotic genomes. The highest robustness values were observed for the weighting based on protein translation errors, one of the most frequent sources of errors in the information flow from DNA to proteins. The thermophilic prokaryotes are significantly more robust to errors than non-thermophilic prokaryotes, mainly at the level of those heterogeneous codon blocks that correspond to some of the amino acids that tend to be more frequent in thermophilic or hyper-thermophilic proteins. We have also shown that the high robustness values observed for these heterogeneous codon blocks in thermophilic genomes are due to the fact that the most robust codons, such as, AGG(R), AGA(R), CCA (P), GUU (V) and CUU (L), tend to be significantly more frequent than the least robust codons CGU(R), CGC(R), CCG(P), GUG(V) and CUG(L).

7.2 Future work

Several factors have been shown to be associated with the frequency of synonymous codon usage, such as, the base composition, the translation efficiency and accuracy, the RNA stability and the optimal growth temperature, among others. It is known, for example, that the highly expressed genes tend to prefer codons corresponding to abundant tRNAs. This correspondence between codon usage bias and tRNA content increases the protein translation efficiency and accuracy. We could explore the relationship between codon robustness and other factors like codon usage bias, mRNA secondary structures and tRNA gene copy number, in the highly expressed genes of thermophiles and non-thermophiles. This work might be useful, for example, to develop new methodologies for the optimization of heterologous gene expression.

We think that these methods to assess the genetic code and genome robustness together with other methods for calculating certain physical and chemical properties of new chemical compounds could be very useful for the new technologies to incorporate non-canonical amino acids into the genetic code.

REFERENCES

[1] Zaher HS, Green R. Fidelity at the molecular level: lessons from protein synthesis. C*ell,*136(4):746-762, 2009.

[2] Sniegowski P., Yevgeniy R. Mutation Rates: How Low Can You Go? *Current Biology*, 23(4): R147-R149, 2013.

[3] Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 10(10):715-724, 2009.

[4] Natali F., Rancati G.The Mutator Phenotype: Adapting Microbial Evolution to Cancer Biology. *Frontiers in Genetics.* 10. 713. 2019.

[5] Denamur E. and Matic I. Evolution of mutation rates in bacteria. *Mol. Microbiol*. 60: 820–827, 2006.

[6] Freeland S.J. The Darwinian Genetic Code: An Adaptation for Adapting? *Genetic Programming and Evolvable Machines*. 3(2):113-127, 2002.

[7] Guo HH, Choe J., Loeb LA. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci.USA*, 101 (25): 9205-9210, 2004.

[8] Navlakha S., He X., Faloutsos C. and Bar-Joseph Z. Topological properties of robust biological and computational networks*. J. R. Soc. Interface.* 11 :20140283, 2014.

[9] van Dijk AD, van Mourik S, van Ham RC. Mutational robustness of gene regulatory networks. *PLoS One.* 7(1): e30591, 2012.

[10] Masel J., Siegal M L. Robustness: mechanisms and consequences. *Trends in Genetics*. 25(9): 395–403, 2009.

[11] Castro-Chavez F. The rules of variation: Amino acid exchange according to the rotating circular genetic code, *J Theor Biol,*264:711-721, 2010.

[12] Fares MA. The origins of mutational robustness. *Trends Genet*.31(7):373-81, 2015.

[13] Agris PF, Eruysal ER, Narendran A, Väre VYP, Vangaveti S and Ranganathan SV. Celebrating wobble decoding: Half a century and still much is new. *RNA Biol*.15(4-5):537-553, 2018.

[14] Freeland S. J. and Hurst L. D. Load minimization of the genetic code: history does not explain the pattern. *Proc. R. Soc. Lond. B.* 265:2111-2119, 1998.

[15] Freeland SJ, Knight RD, Landweber LF, Hurst LD. Early fixation of an optimal genetic code. *Mol. Biol. Evol.* 17(4):511-8, 2000.

[16] Guilloux A, Jestin JL,The genetic code and its optimization for kinetic energy conservation in polypeptide chains, *Biosystems,*109:141-144,2012.

[17] Massey SE. Genetic Code Evolution Reveals the Neutral Emergence of Mutational Robustness, and Information as an Evolutionary Constraint. *Life (Basel).* 5(2): 1301–1332, 2015.

[18] Jestin, J.L., Kempf, A. Optimization Models and the Structure of the Genetic Code. *J Mol Evol*. 69, 452, 2009.

[19] Wong JT, Ng SK, Mat WK, Hu T, Xue H. Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life. *Life (Basel).*6(1):12, 2016.

[20] Santos J. and Monteagudo A. Simulated evolution applied to study the genetic code optimality using a model of codon reassignments. *BMC Bioinformatics.* 12: 56, 2011.

[21] Wiltschi, B., Budisa, N. Natural history and experimental evolution of the genetic code. *Appl Biotechnol*. 74, 739–753, 2007.

[22] Freeland SJ, Knight RD, Landweber LF. Measuring adaptation within the genetic code. *Trends Biochem Sci.* 25(2):44-5, 2000.

[23] Goodarzi H., Najafabadi HS, Torabi N On the coevolution of genes and genetic code. *Gene* 362:133–140, 2005.

[24] Koonin, E.V. Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code. *Life*, 7: 22-35, 2017.

[25] Haig D, Hurst LD. A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution*. 33:412–417, 1991.

[26] Freeland SJ, Hurst LD. The genetic code is one in a million. *Journal of Molecular Evolution*. 47(3):238– 248, 1998.

[27] Buhrman H., van der Gulik PTS, Kelk SM, Koolen, WM., and Stougie L. Some Mathematical Refinements Concerning Error Minimization in the Genetic Code*. IEEE/ACM Transactions on Computational Biology and Bioinformatics* Volume 8: 1358-1372, 2011.

[28] Di Giulio M, Capobianco MR and Medugno M. On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. *J of Theo Biol.*168:43–51,1994.

[29] Judson OP, Haydon D. The genetic code: what is it good for? An analysis of the effects of selection pressures on genetic codes. *J Mol Evol*. 49(5):539-50, 1999.

[30] Santos J. and Monteagudo A. Inclusion of the fitness sharing technique in an evolutionary algorithm to analyze the fitness landscape of the genetic code adaptability. *BMC Bioinformatics*. 18(1):195, 2017.

[31] BłaŻej P, Wnetrzak M, Mackiewicz D and Mackiewicz P. The influence of different types of translational inaccuracies on the genetic code structure. *BMC Bioinformatics.* 20(1):114, 2019.

[32] Wnętrzak M, Błażej P, Mackiewicz D and Mackiewicz P. The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm. *BMC Evol Biol*. 18(1):192, 2018.

[33] de Oliveira LL, de Oliveira PS and Tinós R. A multiobjective approach to the genetic code adaptability problem. *BMC Bioinformatics.* 16:52, 2015.

[34] Alff-Steinberger C. The genetic code and error transmission. *PNAS USA*, 64(2):584-91, 1969.

[35] Zhu CT, Zeng XB and Huang WD. Codon usage decreases the error minimization within the genetic code. *J Mol Evol.* 57(5):533-7, 2003.

[36] Gilis D, Massar S, Cerf NJ and Rooman M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biology.*2(11): research0049.1–0049.12, 2001.

[37] Koonin EV and Novozhilov AS. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct*. 2:24, 2007.

[38] Zhu W, Freeland S. The standard genetic code enhances adaptive evolution of proteins. *J. Theor Biol*.239(1):63-70, 2006.

[39] Koonin EV and Novozhilov AS. Origin and Evolution of the Universal Genetic Code. *Annu. Rev. Genet.* 51:45-62, 2017.

[40] Torabi N, Goodarzi H, Najafabadi HS, The case for an error minimizing set of coding amino

acids, *J Theor Biol*, 244:737-744,2007.

[41] Archetti, M. Selection on codon usage for error minimization at the protein level. *J. Mol. Evol.* 59, 400– 415, 2004.

[42] Archetti, M. Genetic robustness and selection at the protein level for synonymous *J. Evol. Biol.* 19 (2), 353, 2006.

[43] Najafabadi HS, Goodarzi H and Torabi N. Optimality of codon usage in *Escherichia coli* due to load minimization. *J Theor Biol*. 237(2):203-9, 2005.

[44] Najafabadi HS, Lehmann J and Omidi, M. Error minimization explains the codon usage of highly expressed genes in *Escherichia coli*. *Gene* 387:150–155, 2007.

[45] Marquez R, Smit S and Knight R. Do universal codon-usage patterns minimize the effects of mutation and translation error? *Genome Biol.* 6(11): R91, 2005.

[46] Singer GA and Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene.* 317(1-2):39-47, 2003.

[47] Ding, Y., Cai, Y., Han, Y. et al. Comparison of the structural basis for thermal stability between archaeal and bacterial proteins. *Extremophiles* 16, 67–78, 2012.

[48] Van der Linden, MG and Torres de Farias S. Correlation between codon usage and thermostability. *Extremophiles* 10:479–481, 2006.

[49] Lynn DJ, Singer GAC, and Hickey DA. Synonymous codon usage is subject to selection in thermophilic Bacteria. *Nucleic Acids Res.* 30(19): 4272–4277, 2002.

[50] Basak S., Roy S and Chandra Ghosh T. On the origin of synonymous codon usage divergence between thermophilic and mesophilic prokaryotes. *FEBS Lett* 581: 5825– 5830, 2007.

[51] Gromiha, M.M., Pathak, M.C., Saraboji, K., Ortlund, E.A. and Gaucher, E.A. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins*, 81: 715-721, 2013

[52] Karshikoff A., Nilsson L. and Ladenstein R. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS Journal.* 282:3899–3917, 2015.

[53] Hait, S, Mallik, S, Basu, S, Kundu, S. Finding the generalized molecular principles of protein thermal stability. *Proteins*; 1– 21, 2019.

[54] Bada JL. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chem. Soc. Rev*. 42, 2186-96, 2013.

[55] Freeland SJ, Wu T, Keulmann N. The case for an error minimizing standard genetic code. *Orig Life Evol Biosph.* 33:457–477, 2003.

[56] Yarus M. The Genetic Code and RNA-Amino Acid Affinities. *Life (Basel).* 7(2):13, 2012.

[57] Rodin AS, Szathmáry E and Rodin SN. On origin of genetic code and tRNA before translation. *Biol. Direct*. 6:14, 2011.

[58] Di Giulio, M. The Origin of the Genetic Code: Matter of Metabolism or Physicochemical determinism?. *J Mol Evol.* 77, 131–133, 2013.

[59] Di Giulio. An extension of the coevolution theory of the origin of the genetic code. *Biol. Direct.* 3:37, 2008.

[60] Di Giulio M. A discriminative test among the different theories proposed to explain the origin of the genetic code: The coevolution theory finds additional support. *Biosystems*. 169-170:1-4, 2018

[61] Takénaka A, Moras D, Correlation between equipartition of aminoacyl-tRNA synthetases and

amino-acid biosynthesis pathways, *Nucleic Acids Research*, 48: 3277–3285, 2020.

[62] Di Giulio, M. A Non-neutral Origin for Error Minimization in the Origin of the Genetic Code. *J Mol Evol* 86, 593–597, 2018.

[63] Massey SE. A sequential “2-1-3” model of genetic code evolution that explains codon constraints. *J. Mol. Evol.* 62:809–10, 2006.

[64] Massey SE. The neutral emergence of error minimized genetic codes superior to the Standard genetic code. *J. Theor. Biol.* 408:237–42, 2016.

[65] Higgs PG. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol. Direct.* 4:16, 2009.

[66] Fitch WM and Upper K. The phylogeny of tRNA sequences provides evidence for ambiguity Reduction in the origin of the genetic code. *Cold Spring Harb. Symp. Quant. Biol*. 52:759–67,1987

[67] Francis BR. Evolution of the genetic code by incorporation of amino acids that improved or changed protein function. *J. Mol. Evol.* 77:134–58, 2013.

[68] Ribas de Pouplana L, Turner RJ, Steer BA and Schimmel P. Genetic code origins: tRNAs older than their synthetases? *Proc. Natl. Acad. Sci. USA*, 95 (19): 11295-11300, 1998.

[69] Bezerra AR, Guimarães AR, Santos MA. Non-Standard Genetic Codes Define New Concepts for Protein Engineering. Life (Basel). 2015;5(4):1610-1628, 2015.

[70] Ambrogelly A, Palioura S and Soll D. Natural expansion of the genetic code. *Nat. Chem. Biol.* 3:29–35, 2007.

[71] Sengupta S and Higgs PG. 2005. A unified model of codon reassignment in alternative genetic codes. *Genetics* 170:831–40, 2005.

[72] Sengupta S, Higgs PG. Pathways of genetic code evolution in ancient and modern organisms. *J. Mol. Evol*. 80:229–43, 2015.

[73] Koonin EV and Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life.* 61(2):99-111, 2009.

[74] Goodarzi H, Najafabadi HS, Nejadb HA and Torabi N. The impact of including tRNA content on the optimality of the genetic code. *Bulletin of Mathematical Biology* 67: 1355–1368, 2005.

[75] Torabi N, Goodarzi H and Najafabadi HS. The case for an error minimizing set of coding amino acids. *J Theor Biol.* 244(4):737-44, 2007.

[76] Novozhilov AS and Koonin EV. Exceptional error minimization in putative primordial genetic codes. *Biol Direct.* 4:44, 2009.

[77] Buhrman H, van der Gulik PTS, Klau GW, Schaffner C, Speijer D and Stougie L. A Realistic Model Under Which the Genetic Code is Optimal. *J Mol Evol.* 77(4):170-84, 2013.

[78] Santos J, Monteagudo A. Study of the genetic code adaptability by means of a genetic algorithm *J Theor Biol.* 264(3):854-65, 2010.

[79] Błażej P, Wnętrzak M, Mackiewicz D and Mackiewicz P. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. *PLoS One.* 13(8): e0201715, 2018.

[80] Goldman N. Further Results on Error Minimization in the Genetic Code. *J Mol Evol* 37:662-664, 1993.

[81] Caporaso JG, Yarus M, Knight R. Error Minimization and Coding Triplet/Binding Site

Associations Are Independent Features of the Canonical Genetic Code. *J Mol Evol*, 61:597–607, 2005.

[82] QAPLIB [http://www.opt.math.tu-graz.ac.at/qaplib/]

[83] Attie O., Sulkow B., Di C and Qiu W. Genetic codes optimized as a traveling salesman problem. *PLoS One.* 14(10): e0224552, 2019.

[84] Błażej P, Kowalski DR, Mackiewicz D, Wnetrzak M, Aloqalaa DA and Mackiewicz P. The structure of the genetic code as an optimal graph clustering problem. *bioRxiv* May. 28, 2018.

[85] Bilgin, T., Kurnaz, I.A. & Wagner, A. Selection Shapes the Robustness of Ligand-Binding Amino Acids. *J Mol Evol*. 76, 343–349, 2013.

[86] Lauring, A., Acevedo, A., Cooper, S., and Andino, R. 2012. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an rna virus. *Cell Host & Microbe,* 12(5): 623-632, 2012.

[87] Firnberg E, Ostermeier M. The genetic code constrains yet facilitates Darwinian evolution. *Nucleic Acids Res*. 41(15):7420-7428, 2013.

[88] Schwersensky, M., Rooman M. and Pucci F. Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness. *bioRxiv* 2020.02.05.935809, 2020.

[89] Błażej P, Miasojedow B, Grabińska M, Mackiewicz P. Optimization of Mutation Pressure in Relation to Properties of Protein-Coding Sequences in Bacterial Genomes. *PLoS One*. 10(6): e0130411, 2015.

[90] Hormoz, S. Amino acid composition of proteins reduces deleterious impact of mutations. *Sci Rep.* 3, 2919, 2013.

[91] Finch AJ, Kim JR. Thermophilic Proteins as Versatile Scaffolds for Protein Engineering. *Microorganisms*. 6(4):97, 2018.

[92] Burkard RE, Cela ED, Pardalos PM and Pitsoulis LS. The quadratic assignment problem. *Handbook of Combinatorial Optimization*. 2:241-337, Kluwer Academic Publishers, 1998.

[93] Hubert, L. and Schultz, J. (1976), Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29: 190-241, 1976.

[94] Dutilleul P, Stockwell JD, Frigon D and Legendre P. The Mantel Test versus Pearson's Correlation Analysis: Assessment of the Differences for Biological and Environmental Studies. *Journal of Agricultural, Biological, and Environmental Statistics,* 5: 131-150, 2000.

[95] de Carvalho S.A., Rahmann S. Improving the Layout of Oligonucleotide Microarrays: Pivot Partitioning. In: Bücher P., Moret B.M.E. (eds) Algorithms in Bioinformatics. WABI 2006. *Lecture Notes in Computer Science*, vol 4175. Springer, Berlin, Heidelberg, 2006.

[96] Feizi S, Quon G., Mendoza M., Medard M., Kellis M. and Jadbabaie M. Spectral Alignment of Graphs. *IEEE Transactions on Network Science and Engineering*, 1-1. April 2019.

[97] El-Kebir M, Heringa J and Klau GW. Natalie 2.0: Sparse Global Network Alignment as a Special Case of Quadratic Assignment. *Algorithms* 8:1035-1051, 2015.

[98] Cela E. The quadratic assignment problem. Theory and algorithms. Kluwer Academic Pub., 1998.

[99] Sahni S. and Gonzalez T. P-Complete Approximation Problems. *J of the ACM*. 23:555-556, 1976.

[100] Nyberg A. The quadratic assignment problem. *Some Reformulations for the Quadratic*

Assignment Problem. PhD Thesis in Process Design and Systems Eng., Åbo, Finland, 2014.

[101] Burkard R., Dell'Amico M. and Martello S. Assignment problems. SIAM, Philadelphia, 2009.

[102] Abdel-Basset, Mohamed & Manogaran, Gunasekaran & Rashad, Heba & Zaied, Abdel Nasser. A comprehensive review of quadratic assignment problem: variants, hybrids and applications. *Journal of Ambient Intelligence and Humanized Computing.* 1-24, June, 2018.

[103] Tseng LY and Liang SC. A Hybrid Metaheuristic for the Quadratic Assignment Problem. *Computational Optimization and Applications*, 34:85–113, 2006.

[104] Li Y, Pardalos PM, Ramakrishnan KG and Resende MGC. Lower bounds for the Quadratic Assignment problem. *Annals of Operations Research*. 50:387–410, 1994.

[105] Christofides N and Gerrard M. A graph theoretic analysis of bounds for the quadratic Assignment problem. *Studies on graphs and discrete programming.* North-Holland Publishing Company, 61-68, 1981.

[106] Christofides N and Benavent E. An Exact Algorithm for the Quadratic Assignment Problem on a Tree. *Operations Research*, 37:760-768, 1989.

[107] Hoernes, T.P., Faserl, K., Juen, M.A. et al. Translation of non-standard codon nucleotides reveals minimal requirements for codon-anticodon interactions. *Nat. Commun.* 9, 4865, 2018.

[108] Mantel N. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research* 27:209-220,1967.

[109] Qiu P. A quarterly journal of methods applications and related topics *Journal of quality technology*, 50: 49-65, 2018.

[110] Pinelis I, de la Peña V., Ibragimov R, Osękowski A. and Shevtsova I. Inequalities and Extremal Problems in Probability and Statistics. Academic Press. 1st Ed., 2017.

[111] Zeeberg B. Shannon Information Theoretic Computation of Synonymous Codon Usage Biases in Coding Regions of Human and Mouse Genomes. *Genome Res.* 12: 944-955, 2002.

[112] Stroup WW. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC press, Taylor and Francis group, 2012.

[113] Jiming J. Linear and Generalized Linear Mixed Models and Their Applications. Springer, 2007.

[114] Bates D., Maechler M, Bolker B. and Walker S. Fitting Linear Mixed-Effects Models using lme4. J*ournal of Statistical Software*, 67(1), 1-48, 2015.

[115] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[116] Trinquier G. and Sanejouand YH. Whih effective property of amino acids is best preserved by the genetic code?. *Protein engineering*. 11:153-169, 1998.

[117] Gromiha, M.M., Pathak, M.C., Saraboji, K., Ortlund, E.A. and Gaucher, E.A. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins*, 81: 715-721,2013.

[118] Pace CN, Kailong F, Fryar KL, Landua J, Trevino SR, Shirley BA, Hendricks MM, Limura S, Gajiwala K, Scholtz M and Grimsley GR. Contribution of Hydrophobic Interactions to Protein Stability. *J Mol Biol.* 408(3): 514–528, 2011.

[119] Błażej P, Wnętrzak M, Mackiewicz P. The role of crossover operator in evolutionary-based approach to the problem of genetic code optimization. *BioSystems*. 150:61–72, 2016.

[120] Sammet SG, Bastolla U and Porto M. Comparison of translation loads for standard and

alternative genetic codes. *BMC Evolutionary Biology*. 10:178, 2010.

[121] Bada JL. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chem. Soc. Rev.* 42: 2186, 2013.

[122] Quax TE, Claassens NJ, Söll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell*. 59(2):149-161, 2015.

[123] Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol*;19(1):20-30, 2018.

[124] Gromiha MM, Pathak MC, Saraboji K, Ortlund EA, Gaucher EA. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins.* 81(4):715- 21, 2013.

[125] Sengupta D and Kundu S. Role of long- and short-range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. *BMC Bioinformatics,* 13: 142, 2012.

[126] Sawle L, Ghosh K. How do thermophilic proteins and proteomes withstand high temperature?. *Biophys J.* 101(1):217–227, 2011.

[127] Liu Z., Lemmonds S, Huang J., Tyagi M, Hong L and Jain N. Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. USA*, 115 (43) E10049-E10058, 2018.

[128] Zhou XX, Wang YB, Pan YJ and Li WF. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids.*34(1):25-33, 2008

[129] Taylor TJ and Vaisman II. Discrimination of thermophilic and mesophilic proteins. *BMC Struct Biol.* 10(Suppl 1): S5, 2010.

[130] Khan, M.F., Patra, S. Deciphering the rationale behind specific codon usage pattern in extremophiles. *Sci Rep.* 8, 15548, 2018.

[131] Kurnaz, M.L., Bilgin, T. & Kurnaz, I.A. Certain Non-Standard Coding Tables Appear to be More Robust to Error Than the Standard Genetic Code. *J Mol Evol* 70, 13–28, 2010.

[132] Morgens, D.W., Cavalcanti, A.R.O. An Alternative Look at Code Evolution: Using Non-canonical Codes to Evaluate Adaptive and Historic Models for the Origin of the Genetic Code.*J Mol Evol* 76, 71–80,2013.

[133] Warnecke T, Hurst LD. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat Rev Genet.* 2011;12(12):875-881, 2011.

[134] Kalapis D, Bezerra AR, Farkas Z, et al. Evolution of Robustness to Protein Mistranslation by Accelerated Protein Turnover. *PLoS Biol*.;13(11): e1002291, 2015.

[135] Mohler, K., Ibba, M. Translational fidelity and mistranslation in the cellular response to stress. *Nat Microbiol.* 2, 17117, 2017.

[136] Schramm FD, Schroeder K, Jonas K. Protein aggregation in bacteria. *FEMS Microbiology Reviews,* 44:54-72, 2020.

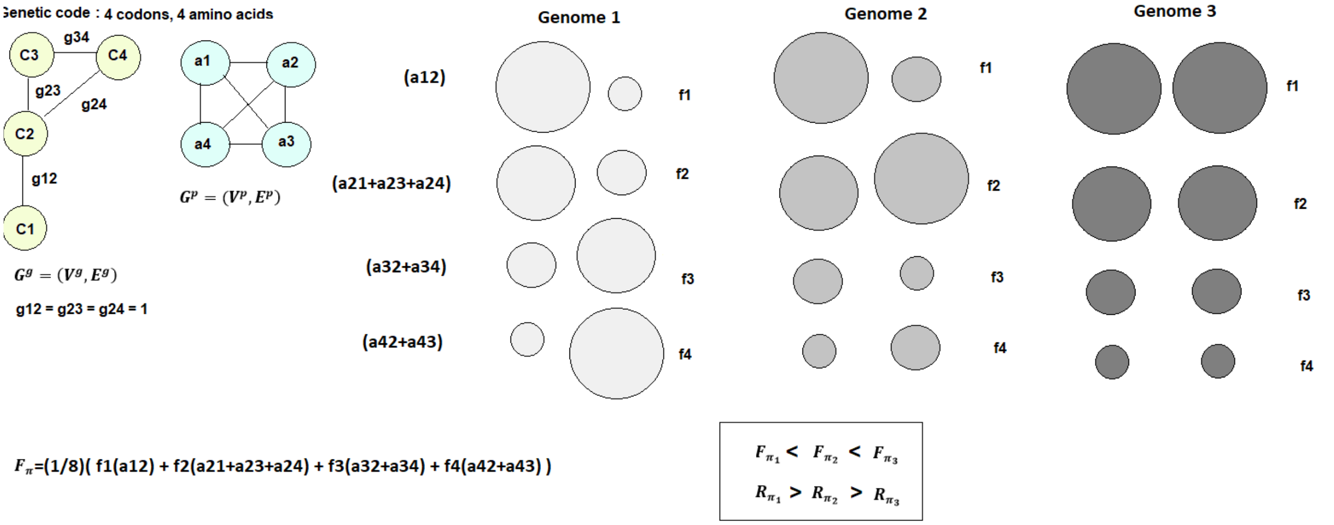
[137] Venev SV, Zeldovich KB. Thermophilic Adaptation in Prokaryotes Is Constrained by Metabolic Costs of Proteostasis. *Mol Biol Evol.* 2018;35(1):211-224, 2018.

[138] Magyar C, Gromiha MM, Sávoly Z, Simon I, The role of stabilization centers in protein thermal stability, *Biochemical and Biophysical Research Communications*. 471, 57-62, 2016

[139] Lauring, A., Frydman, J. & Andino, R. The role of mutational robustness in RNA virus evolution. *Nat Rev Microbiol.* 11, 327–336, 2013.

APPENDICES

Appendix I. Algorithms and methods



Genetic code : 4 codons, 4 amino acids
g34
C3
C4
g23
g24
C2
g12
C1
$G^g = (V^g, E^g)$
g12 = g23 = g24 = 1
a1
a2
a4
a3
$G^p = (V^p, E^p)$
Genome 1
Genome 2
Genome 3
(a12)
(a21+a23+a24)
(a32+a34)
(a42+a43)
f1
f2
f3
f4
F_π=(1/8)(f1(a12) + f2(a21+a23+a24) + f3(a32+a34) + f4(a42+a43))
$F_{\pi_1} < F_{\pi_2} < F_{\pi_3}$
$R_{\pi_1} > R_{\pi_2} > R_{\pi_3}$

by the proposition of Hardy, Littlewood, Polya, 1952

Figure I.1 Three hypothetical genomes that depend on a genetic code of four codons that code for different amino acids. Below each genome, on the left side, the mean phenotypic change in terms of amino acid distances, on the right, the codon frequency. The size of the circles is proportional to the values of the mean phenotypic change and codon usages. According to the proposition of Hardy, Littlewood, Polya (1952), if the more frequent codons are those with smaller mean phenotypic change (or larger robustness), then the value of the mean phenotypic change (bottom left) weighted with the genomic codon usage will be also smaller. This is the case for the genome 1, which is the most robust among the three genomes. Inside the box, genomes sorted in increasing order of the mean phenotypic change or robustness. Bottom left: Computing the weighted mean phenotypic change for this example.

Table I.1 Set of codons adjacent to each of the three stop codons UAA, UAG and UGA (standard genetic code). In red letters, the only position that differs between the stop codon and its neighbors.

UAA (Ochre)		**UAG (Amber)**		**UGA(Opal)**	
codons	**Amino acid**	**codons**	**Amino acid**	**codons**	**Amino acid**
AAA	Lys	AAG	Lys	AGA	Arg
CAA	Gln	CAG	Gln	CGA	Arg
GAA	Glu	GAG	Glu	GGA	Gly
UCA	Ser	UCG	Ser	UAA	None
UGA	None	UGG	Trp	UCA	Ser
UUA	Leu	UUG	Leu	UUA	Leu
UAC	Tyr	UAA	None	UGC	Cys
UAG	None	UAC	Tyr	UGG	Trp
UAU	Tyr	UAU	Tyr	UGU	Cys

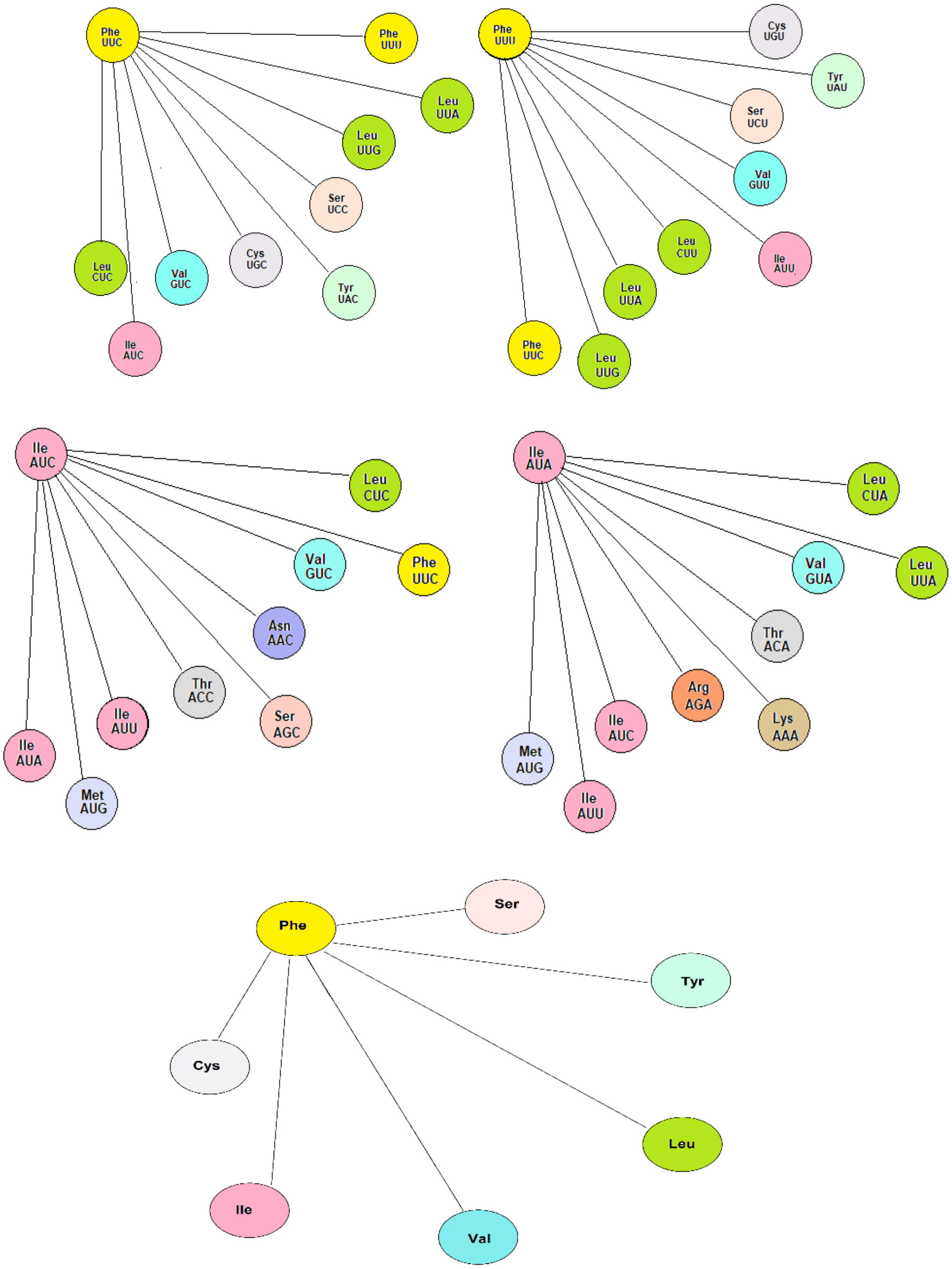
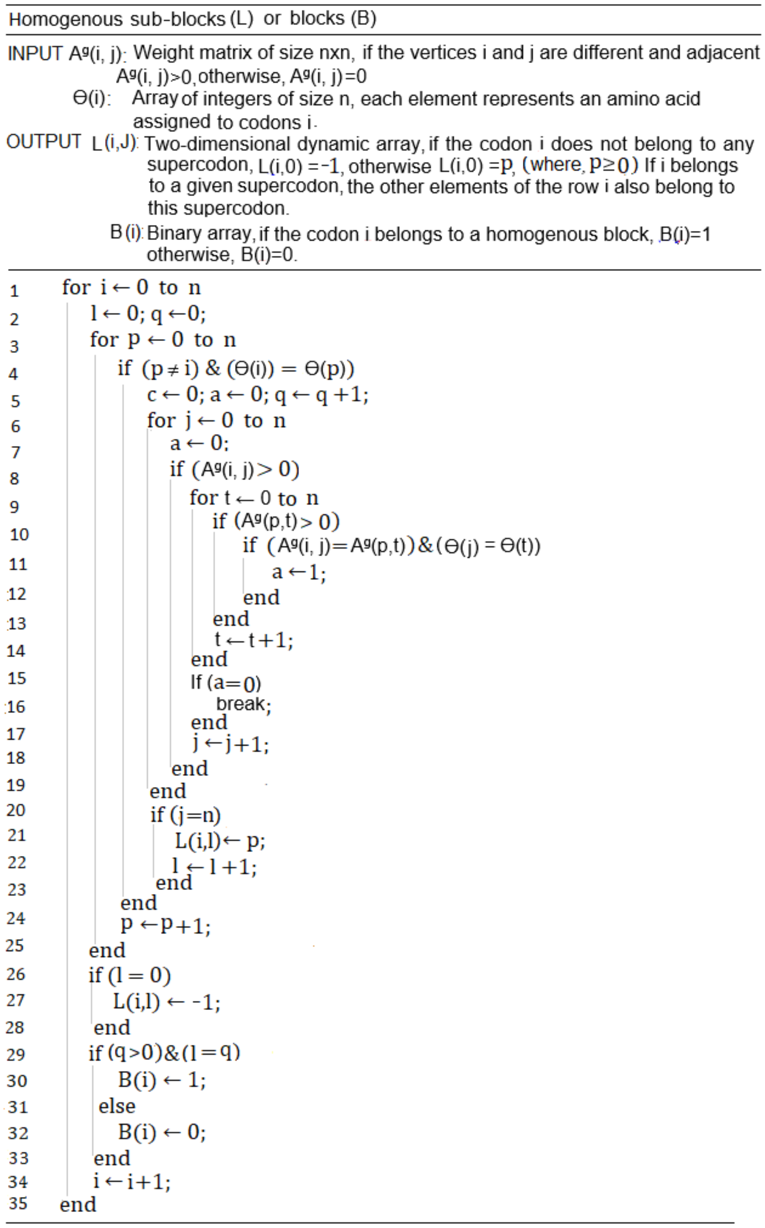


Figure I.2 Top: Two neighboring codons (UUU, UUC) that form a *homogeneous* block in the standard genetic code. They share the same amino acid neighborhood because their adjacent codons specify the same amino acids. Middle: two adjacent codons (AUC, AUA) that do not form a *homogeneous sub-block* in the standard genetic code. Bottom: The homogenous block that corresponds to the amino acid Phe and its neighborhood according to the block-based model.



Homogenous sub-blocks (L) or blocks (B)

INPUT $A^g(i, j)$: Weight matrix of size nxn, if the vertices i and j are different and adjacent $A^g(i, j)>0$, otherwise, $A^g(i, j)=0$

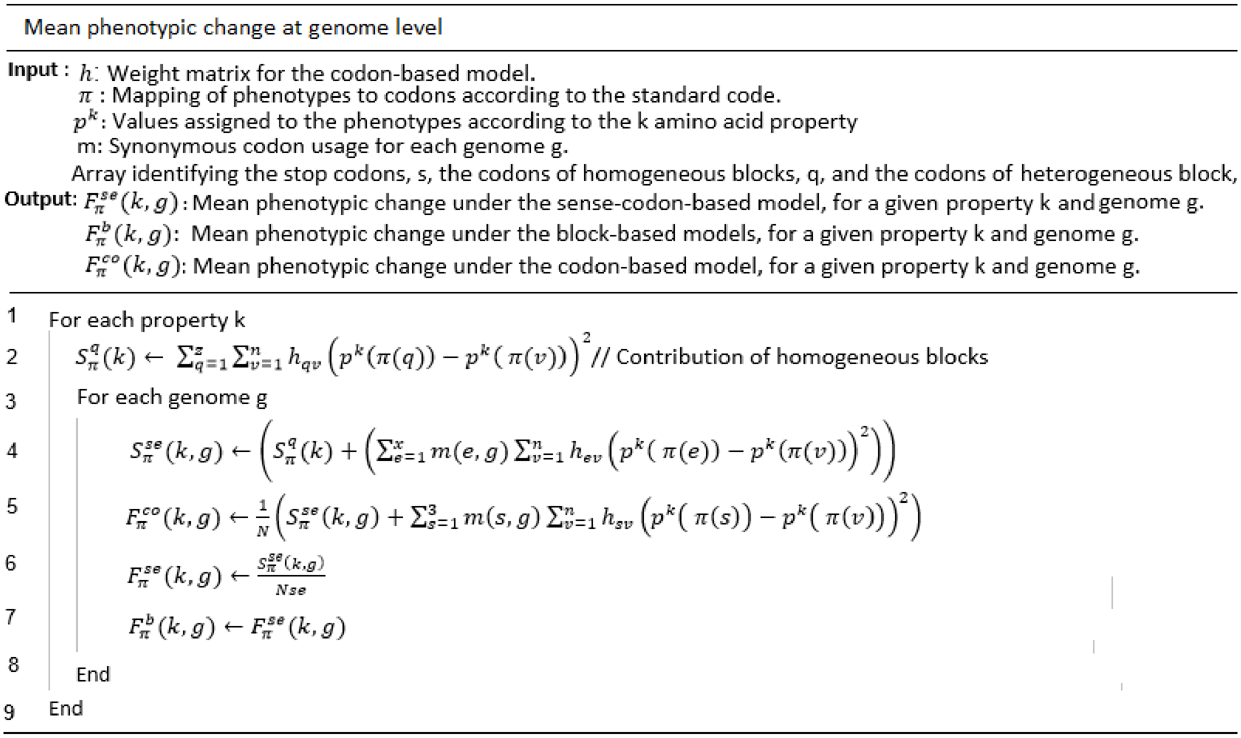
$\Theta(i)$: Array of integers of size n, each element represents an amino acid assigned to codons i.

OUTPUT L(i,J): Two-dimensional dynamic array, if the codon i does not belong to any supercodon, L(i,0) =-1, otherwise L(i,0) =p, (where, $p\geq0$) If i belongs to a given supercodon, the other elements of the row i also belong to this supercodon.

B(i): Binary array, if the codon i belongs to a homogenous block, B(i)=1 otherwise, B(i)=0.

```
for $i \leftarrow 0$ to n
    $l \leftarrow 0; q \leftarrow 0;$
    for $p \leftarrow 0$ to n
        if $(p \neq i)$ & $(\Theta(i)) = \Theta(p))$
            $c \leftarrow 0; a \leftarrow 0; q \leftarrow q+1;$
            for $j \leftarrow 0$ to n
                $a \leftarrow 0;$
                if $(A^g(i, j) > 0)$
                    for $t \leftarrow 0$ to n
                        if $(A^g(p,t) > 0)$
                            if $(A^g(i, j) = A^g(p,t))$ & $(\Theta(j) = \Theta(t))$
                                $a \leftarrow 1;$
                            end
                        end
                        $t \leftarrow t+1;$
                    end
                    If $(a=0)$
                        break;
                    end
                    $j \leftarrow j+1;$
                end
            end
            if $(j=n)$
                $L(i,l) \leftarrow p;$
                $l \leftarrow l+1;$
            end
        end
        $p \leftarrow p+1;$
    end
    if $(l = 0)$
        $L(i,l) \leftarrow -1;$
    end
    if $(q>0)$&$(l=q)$
        $B(i) \leftarrow 1;$
    else
        $B(i) \leftarrow 0;$
    end
    $i \leftarrow i+1;$
end
```

Pseudocode I.1 Algorithm to verify whether each codon represented in Ag fulfils the three requirements for being part of a homogeneous sub-block or block.



```
Mean phenotypic change at genome level

Input : $h$: Weight matrix for the codon-based model.
        $\pi$ : Mapping of phenotypes to codons according to the standard code.
        $p^k$: Values assigned to the phenotypes according to the k amino acid property
        m: Synonymous codon usage for each genome g.
        Array identifying the stop codons, s, the codons of homogeneous blocks, q, and the codons of heterogeneous block,
Output: $F_\pi^{se}(k,g)$: Mean phenotypic change under the sense-codon-based model, for a given property k and genome g.
        $F_\pi^{b}(k,g)$: Mean phenotypic change under the block-based models, for a given property k and genome g.
        $F_\pi^{co}(k,g)$: Mean phenotypic change under the codon-based model, for a given property k and genome g.

For each property k
    $S_\pi^q(k) \leftarrow \sum_{q=1}^{z}\sum_{v=1}^{n} h_{qv}\left(p^k(\pi(q)) - p^k(\pi(v))\right)^2$ // Contribution of homogeneous blocks
    For each genome g
        $S_\pi^{se}(k,g) \leftarrow \left(S_\pi^q(k) + \left(\sum_{e=1}^{x} m(e,g)\sum_{v=1}^{n} h_{ev}\left(p^k(\pi(e)) - p^k(\pi(v))\right)^2\right)\right)$
        $F_\pi^{co}(k,g) \leftarrow \frac{1}{N}\left(S_\pi^{se}(k,g) + \sum_{s=1}^{3} m(s,g)\sum_{v=1}^{n} h_{sv}\left(p^k(\pi(s)) - p^k(\pi(v))\right)^2\right)$
        $F_\pi^{se}(k,g) \leftarrow \frac{S_\pi^{se}(k,g)}{Nse}$
        $F_\pi^{b}(k,g) \leftarrow F_\pi^{se}(k,g)$
    End
End
```

Pseudocode I.2 Algorithm to compute the robustness or mean phenotypic change for a set of genomes according to a set of amino acid indices.

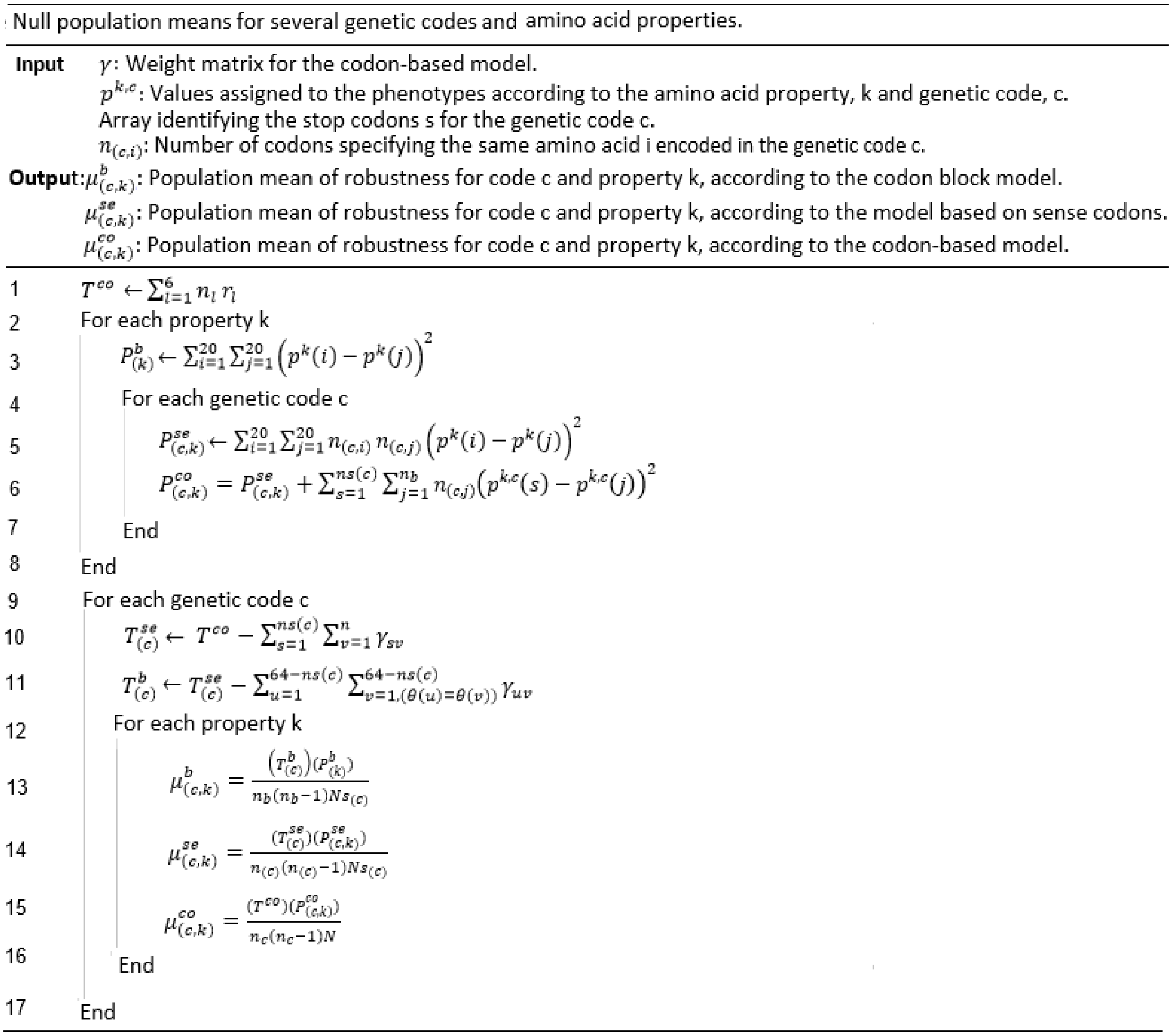
Definitions:$\pi(e), \pi(v), \pi(q), \pi(s)$: Mapping of phenotypes to codons v, q, e and s. $p^k(\pi(e)), p^k(\pi(q)), p^k(\pi(v))$: Values assigned to the amino acids $\pi(u), \pi(q), \pi(v)$ according to the k amino acid property. $m(e,g), m(s,g)$: Frequency of codons e and s for the genome g. h_{qv}, h_{ev}, h_{sv}: Weights on the edges between the codons q, e, s and their corresponding neighbors v. Z: Number of codons belonging to homogeneous blocks in the standard code. x: Number of codons belonging to heterogeneous blocks in the standard code. N: Total number of single base changes, Nse: Total number of single base changes involving only sense codons. n: Number of neighboring vertices.

```
Mean phenotypic change for several genetic codes and properties

Input: $\gamma$: Weight matrix for the codon-based model.
       $\pi$ : Mapping of phenotypes to codons according to the standard genetic code.
       $p^{k,c}$: Values assigned to the phenotypes according to the amino acid property, k, and genetic code, c.
       Array identifying the stop codons s of c.
       Array identifying the codons t to which the alternative genetic code $\beta$ and $\pi$ assign different phenotypes
Output: $F_{\pi}^{se}(k)$: Mean phenotypic change for $\pi$ according to the sense-codon-based model and property k.
        $F_{\pi}^{co}(k)$: Mean phenotypic change for $\pi$ according to the codon-based model and property k.
        $F_{\beta}^{co}(k,c)$: Mean phenotypic change for $\beta$ according to the codon-based model and property k.
        $F_{\beta}^{se}(k,c)$ : Mean phenotypic change for $\beta$ according to the sense-codon-based model and property k.

For each property k
    $S_{\pi}^{se}(k) \leftarrow \left(\sum_{u=1}^{n}\sum_{v=1}^{n}\gamma_{uv}\left(p^{k}(\pi(u)) - p^{k}(\pi(v))\right)^{2}\right)$
    $S_{\pi}^{co}(k) \leftarrow \left(S_{\pi}^{se}(k) + \sum_{s=1}^{ns}\sum_{v=1}^{n}\gamma_{sv}\left(p^{k}(\pi(s)) - p^{k}(\pi(v))\right)^{2}\right)$
    $F_{\pi}^{co}(k) \leftarrow \frac{S_{\pi}^{co}(k)}{N}$
    For each genetic code c
        $S_{\pi}^{t}(k,c) \leftarrow \left(\sum_{t=1}^{r}\sum_{v=1}^{n}\gamma_{tv}\left(p^{k,c}(\pi(t)) - p^{k,c}(\pi(v))\right)^{2}\right)$
        $S_{\beta}^{t}(k,c) \leftarrow \left(\sum_{t=1}^{r}\sum_{v=1}^{n}\gamma_{tv}\left(p^{k,c}(\beta(t)) - p^{k,c}(\beta(v))\right)^{2}\right)$
        $S_{\beta}^{co}(k,c) \leftarrow S_{\pi}^{co}(k) - S_{\pi}^{t}(k,c) + S_{\beta}^{t}(k,c)$
        $F_{\beta}^{se}(k,c) \leftarrow \frac{1}{Nse(c)}\left(S_{\beta}^{co}(k,c) - \left(\sum_{s=1}^{ns(c)}\sum_{v=1}^{n}\gamma_{sv}\left(p^{k,c}(\beta(s)) - p^{k,c}(\beta(v))\right)^{2}\right)\right)$
        $F_{\beta}^{co}(k,c) = \frac{S_{\beta}^{co}(k,c)}{N}$
    End
    $F_{\pi}^{se}(k) = \frac{S_{\pi}^{se}(k)}{Nss}$
End
```

Pseudocode I.3 Algorithm to compute the robustness or mean phenotypic change for a set of genetic codes with respect to several amino acid indices.

Definitions:

$\beta(t)$, $\pi(t)$: Codons t to which the alternative genetic codes β and standard genetic code π assign different phenotypes. $\beta(s)$ $\beta(v)$ $\pi(u)$ $\pi(s)$ $\pi(v)$: Phenotypes assigned to codons u, stop codons s and their neighboring codons v. $\gamma_{sv}, \gamma_{uv}, \gamma_{tv}$: Weights on edges between codons u, s, t and codons v. $F_{\pi}^{t}(k,c)$: Mean phenotypic change for the subset of codons assigned differently in the standard genetic code with respect to alternative codes. $F_{\beta}^{t}(k,c)$: Mean phenotypic change for the subset of codons assigned differently in the alternative with respect to the standard genetic code. ns number of stop codons. $F_{\beta}^{b}(k,c), F_{\pi}^{b}(k,c)$: The mean phenotypic change for π and β under the block model b. As explained in previous section, $F_{\beta}^{b}(k,c) = F_{\beta}^{se}(k,c)$ and $F_{\pi}^{b}(k,c) = F_{\pi}^{se}(k,c)$.
n: Number of neighboring vertices.

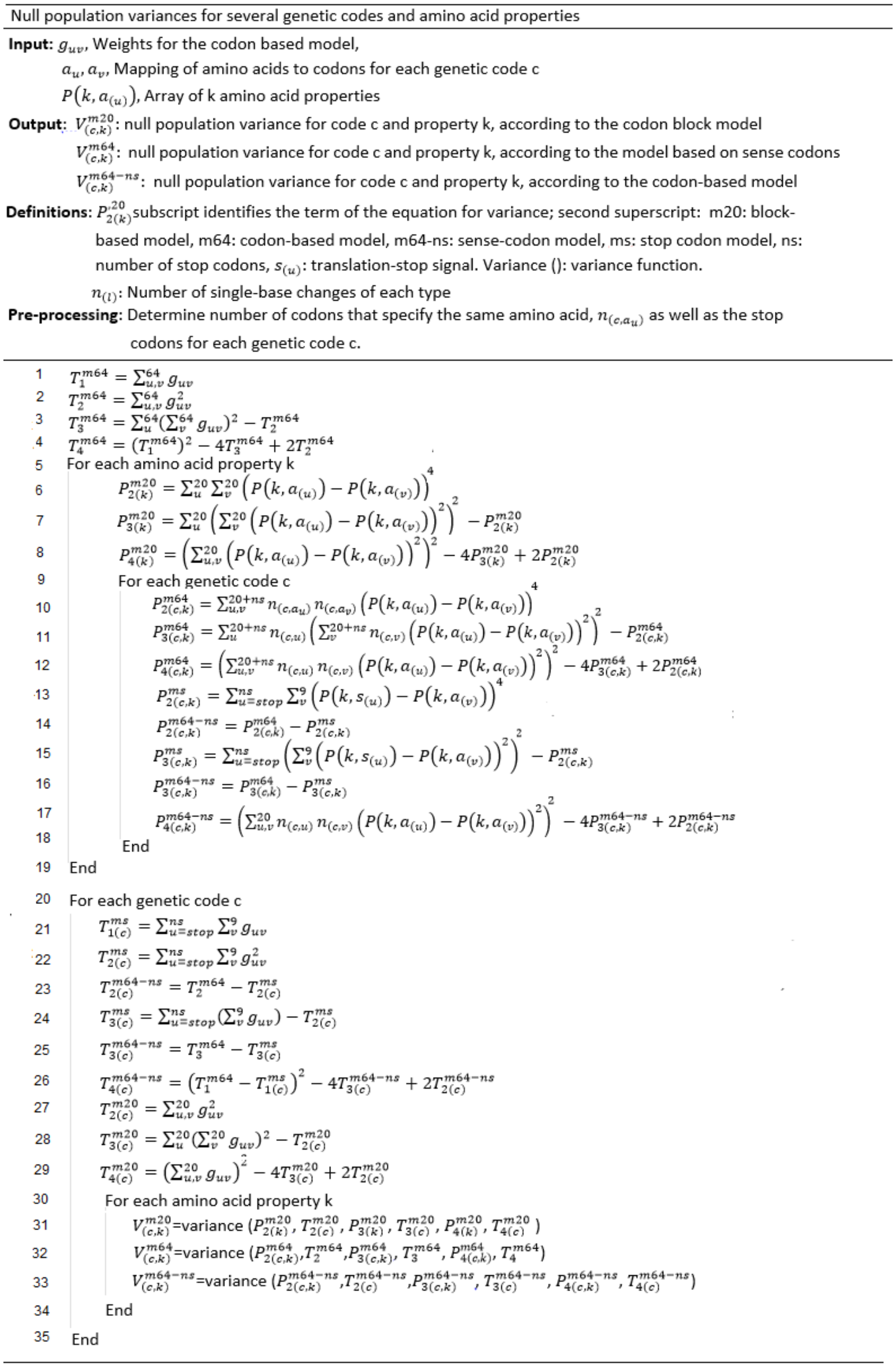


```
Null population means for several genetic codes and amino acid properties.
Input   $\gamma$: Weight matrix for the codon-based model.
        $p^{k,c}$: Values assigned to the phenotypes according to the amino acid property, k and genetic code, c.
        Array identifying the stop codons s for the genetic code c.
        $n_{(c,i)}$: Number of codons specifying the same amino acid i encoded in the genetic code c.
Output: $\mu^{b}_{(c,k)}$: Population mean of robustness for code c and property k, according to the codon block model.
        $\mu^{se}_{(c,k)}$: Population mean of robustness for code c and property k, according to the model based on sense codons.
        $\mu^{co}_{(c,k)}$: Population mean of robustness for code c and property k, according to the codon-based model.

$T^{co} \leftarrow \sum_{l=1}^{6} n_l\, r_l$
For each property k
    $P^{b}_{(k)} \leftarrow \sum_{i=1}^{20} \sum_{j=1}^{20} \left(p^k(i) - p^k(j)\right)^2$
    For each genetic code c
        $P^{se}_{(c,k)} \leftarrow \sum_{i=1}^{20} \sum_{j=1}^{20} n_{(c,i)}\, n_{(c,j)} \left(p^k(i) - p^k(j)\right)^2$
        $P^{co}_{(c,k)} = P^{se}_{(c,k)} + \sum_{s=1}^{ns(c)} \sum_{j=1}^{n_b} n_{(c,j)} \left(p^{k,c}(s) - p^{k,c}(j)\right)^2$
    End
End
For each genetic code c
    $T^{se}_{(c)} \leftarrow T^{co} - \sum_{s=1}^{ns(c)} \sum_{v=1}^{n} \gamma_{sv}$
    $T^{b}_{(c)} \leftarrow T^{se}_{(c)} - \sum_{u=1}^{64-ns(c)} \sum_{v=1,(\theta(u)=\theta(v))}^{64-ns(c)} \gamma_{uv}$
    For each property k
        $\mu^{b}_{(c,k)} = \frac{\left(T^{b}_{(c)}\right)(P^{b}_{(k)})}{n_b(n_b-1)Ns_{(c)}}$
        $\mu^{se}_{(c,k)} = \frac{(T^{se}_{(c)})(P^{se}_{(c,k)})}{n_{(c)}(n_{(c)}-1)Ns_{(c)}}$
        $\mu^{co}_{(c,k)} = \frac{(T^{co})(P^{co}_{(c,k)})}{n_c(n_c-1)N}$
    End
End
```

Pseudocode I.4 Algorithm to compute the null population means for a set of genetic codes with respect to a set of amino acid indices.

Definitions:

$P^{b}_{(k)}, T^{b}_{(c)}$: Terms for computing $\mu^{b}_{(c,k)}$ under the codon-block based model. $P^{co}_{(c,k)}, T^{co}$: Terms for computing $\mu^{co}_{(c,k)}$ under the codon-based model. $P^{se}_{(c,k)}, T^{se}_{(c)}$: Terms for computing $\mu^{se}_{(c,k)}$ under the model based on sense codons.n_l: Number of single-base changes of each type, l. r_l: Weights for single-base changes, s: Stop codons, ns(c): number of stop codons of the genetic code c. N: Number of single-base changes, Ns(c): Number of single-base changes involving sense codons in the genetic code c. nb: number of codon blocks, n(c): number of sense codons of the genetic code c.



Null population variances for several genetic codes and amino acid properties

Input: g_{uv}, Weights for the codon based model,
a_u, a_v, Mapping of amino acids to codons for each genetic code c
$P(k, a_{(u)})$, Array of k amino acid properties

Output: $V^{m20}_{(c,k)}$: null population variance for code c and property k, according to the codon block model
$V^{m64}_{(c,k)}$: null population variance for code c and property k, according to the model based on sense codons
$V^{m64-ns}_{(c,k)}$: null population variance for code c and property k, according to the codon-based model

Definitions: $P^{,20}_{2(k)}$subscript identifies the term of the equation for variance; second superscript: m20: block-based model, m64: codon-based model, m64-ns: sense-codon model, ms: stop codon model, ns: number of stop codons, $s_{(u)}$: translation-stop signal. Variance (): variance function.
$n_{(l)}$: Number of single-base changes of each type

Pre-processing: Determine number of codons that specify the same amino acid, $n_{(c,a_u)}$ as well as the stop codons for each genetic code c.

```
$T^{m64}_1 = \sum^{64}_{u,v} g_{uv}$
$T^{m64}_2 = \sum^{64}_{u,v} g^2_{uv}$
$T^{m64}_3 = \sum^{64}_{u}(\sum^{64}_{v} g_{uv})^2 - T^{m64}_2$
$T^{m64}_4 = (T^{m64}_1)^2 - 4T^{m64}_3 + 2T^{m64}_2$
For each amino acid property k
        $P^{m20}_{2(k)} = \sum^{20}_{u}\sum^{20}_{v}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^4$
        $P^{m20}_{3(k)} = \sum^{20}_{u}\left(\sum^{20}_{v}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^2\right)^2 - P^{m20}_{2(k)}$
        $P^{m20}_{4(k)} = \left(\sum^{20}_{u,v}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^2\right)^2 - 4P^{m20}_{3(k)} + 2P^{m20}_{2(k)}$
        For each genetic code c
            $P^{m64}_{2(c,k)} = \sum^{20+ns}_{u,v} n_{(c,a_u)}\, n_{(c,a_v)}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^4$
            $P^{m64}_{3(c,k)} = \sum^{20+ns}_{u} n_{(c,u)}\left(\sum^{20+ns}_{v} n_{(c,v)}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^2\right)^2 - P^{m64}_{2(c,k)}$
            $P^{m64}_{4(c,k)} = \left(\sum^{20+ns}_{u,v} n_{(c,u)}\, n_{(c,v)}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^2\right)^2 - 4P^{m64}_{3(c,k)} + 2P^{m64}_{2(c,k)}$
            $P^{ms}_{2(c,k)} = \sum^{ns}_{u=stop}\sum^{9}_{v}\left(P(k, s_{(u)}) - P(k, a_{(v)})\right)^4$
            $P^{m64-ns}_{2(c,k)} = P^{m64}_{2(c,k)} - P^{ms}_{2(c,k)}$
            $P^{ms}_{3(c,k)} = \sum^{ns}_{u=stop}\left(\sum^{9}_{v}\left(P(k, s_{(u)}) - P(k, a_{(v)})\right)^2\right)^2 - P^{ms}_{2(c,k)}$
            $P^{m64-ns}_{3(c,k)} = P^{m64}_{3(c,k)} - P^{ms}_{3(c,k)}$
            $P^{m64-ns}_{4(c,k)} = \left(\sum^{20}_{u,v} n_{(c,u)}\, n_{(c,v)}\left(P(k, a_{(u)}) - P(k, a_{(v)})\right)^2\right)^2 - 4P^{m64-ns}_{3(c,k)} + 2P^{m64-ns}_{2(c,k)}$
        End
End

For each genetic code c
    $T^{ms}_{1(c)} = \sum^{ns}_{u=stop}\sum^{9}_{v} g_{uv}$
    $T^{ms}_{2(c)} = \sum^{ns}_{u=stop}\sum^{9}_{v} g^2_{uv}$
    $T^{m64-ns}_{2(c)} = T^{m64}_2 - T^{ms}_{2(c)}$
    $T^{ms}_{3(c)} = \sum^{ns}_{u=stop}(\sum^{9}_{v} g_{uv}) - T^{ms}_{2(c)}$
    $T^{m64-ns}_{3(c)} = T^{m64}_3 - T^{ms}_{3(c)}$
    $T^{m64-ns}_{4(c)} = \left(T^{m64}_1 - T^{ms}_{1(c)}\right)^2 - 4T^{m64-ns}_{3(c)} + 2T^{m64-ns}_{2(c)}$
    $T^{m20}_{2(c)} = \sum^{20}_{u,v} g^2_{uv}$
    $T^{m20}_{3(c)} = \sum^{20}_{u}(\sum^{20}_{v} g_{uv})^2 - T^{m20}_{2(c)}$
    $T^{m20}_{4(c)} = \left(\sum^{20}_{u,v} g_{uv}\right)^2 - 4T^{m20}_{3(c)} + 2T^{m20}_{2(c)}$
     For each amino acid property k
         $V^{m20}_{(c,k)}$=variance ($P^{m20}_{2(k)}$, $T^{m20}_{2(c)}$, $P^{m20}_{3(k)}$, $T^{m20}_{3(c)}$, $P^{m20}_{4(k)}$, $T^{m20}_{4(c)}$ )
         $V^{m64}_{(c,k)}$=variance ($P^{m64}_{2(c,k)}$,$T^{m64}_2$,$P^{m64}_{3(c,k)}$, $T^{m64}_3$, $P^{m64}_{4(c,k)}$, $T^{m64}_4$)
         $V^{m64-ns}_{(c,k)}$=variance ($P^{m64-ns}_{2(c,k)}$,$T^{m64-ns}_{2(c)}$,$P^{m64-ns}_{3(c,k)}$, $T^{m64-ns}_{3(c)}$, $P^{m64-ns}_{4(c,k)}$, $T^{m64-ns}_{4(c)}$)
     End
End
```

Pseudocode I.5 Algorithm to compute the null population variances for a set of genetic codes with respect to several amino acid indices.

Definitions: variance: Equation for the population variance (eq. 14).$P^{b}_{2(k)}$, $P^{b}_{3(k)}$, $P^{b}_{4(k)}$: Right-handed terms in D_2, D_3, D_4 of equation 14, for the block model, b, and amino acid property, k.$T^{b}_{2(c)}$, $T^{b}_{3(c)}$, $T^{b}_{4(c)}$: Left-handed terms in D_2, D_3, D_4 of equation 14, for b and genetic code, c.$P^{se}_{2(c,k)}$, $P^{se}_{3(c,k)}$, $P^{se}_{4(c,k)}$: Right-handed terms in D_2, D_3, D_4 of equation 14, for the model based on sense codons, se, as well as, k and c.$T^{se}_{2(c)}$, $T^{se}_{3(c)}$, $T^{se}_{4(c)}$: Left-handed terms in D_2, D_3, D_4 of equation 14, for se and c.$P^{co}_{2(c,k)}$, $P^{co}_{3(c,k)}$, $P^{co}_{4(c,k)}$: Right-handed terms in D_2, D_3, D_4 of equation 14, for the codon-based model, co, as well as, c and k.T^{co}_{2}, T^{co}_{3}, T^{co}_{4}: Left-handed terms in D_2, D_3, D_4 of equation 14 for the codon-based model. T^{co}_{1},$P^{stp}_{2(c,k)}$, $P^{stp}_{3(c,k)}$, $T^{stp}_{2(c)}$, $T^{stp}_{3(c)}$, $T^{stp}_{1(c)}$: Auxiliary variables for computing the contribution of stop codons stp to right-handed and left-handed terms. ns(c): number of stop codons of c.

Appendix II. Robustness of the standard genetic code

Table II.1 The first 10 aa properties (from a Total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. It was used the **Unbiased-weighted mean phenotypic changes and codon-block representation**. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values Lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. Pr sim: Probability estimated by numerical simulation. Pr norm: Probability estimated by normal approximation. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, Table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**	**Pr.sim**	**Pr norm**
Polar requirement (p149)	0.8659	-2.2253	0.1680	0.4250	2.2537E-04	0.0130
Hydrophobicity (Wimley, p148)	0.7357	-2.0568	0.1912	0.4960	3.2338E-05	0.0199
Hydrophobicity (Meek, p130)	0.8781	-1.8719	0.2220	0.4710	5.2100E-04	0.0306
Medium thermodynamic stability (p188)	0.6057	-1.6512	0.2683	0.5589	1.6430E-04	0.0493
Protein-Protein interactions (p153)	0.9150	-1.6047	0.2797	0.5048	1.4190E-03	0.0543
Flexibility (2FN, MS, p209)	0.8546	-1.5268	0.3002	0.3847	2.3600E-05	0.0634
Flexibility (2FN, ML, p148)	0.9122	-1.4832	0.3125	0.4516	5.5400E-04	0.0690
Small linker propensity (p174)	0.8551	-1.4304	0.3283	0.4242	1.0600E-04	0.0763
Long-range contacts (p164)	0.9355	-1.3538	0.3530	0.4444	5.1880E-04	0.0879
Long-range contacts (p160)	1.0160	-1.3114	0.3677	0.4847	3.1284E-04	0.0949

2FN: Two flexible neighbors, MS: Mean scale parameter, ML: Mean location parameter

Table II.2 The first 10 aa properties (from a Total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. It was used the **Unbiased-weighted mean phenotypic changes and representation based on sense codons**. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values Lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, Table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**
Hydrophobicity (Miyazawa, p132)	0.9026	-8.8416	1.2631E-02	0.2202
Hydrophobicity (Cornette,p115)	1.0187	-8.5917	1.3366E-02	0.2089
Hydrophobicity (Kyte, p125)	1.1083	-8.5177	1.3596E-02	0.1758
Hydrophobicity (Wilson, p147)	1.0755	-8.5135	1.3609E-02	0.2605
Hydrophobicity (Parker, p135)	0.9681	-8.5024	1.3644E-02	0.2202
Transmembrane Alpha-Helix (p35)	1.0523	-8.3352	1.4189E-02	0.1960
Transmembrane Alpha-Helix (p28)	1.2221	-8.3344	1.4192E-02	0.2323
Hydrophobicity (Wilson, p147)	0.9657	-8.2757	1.4391E-02	0.2331
Long-range contacts (p164)	0.9355	-8.0984	1.5019E-02	0.1685
Solvent accesible Surface (p44)	0.9570	-8.0712	1.5118E-02	0.1903

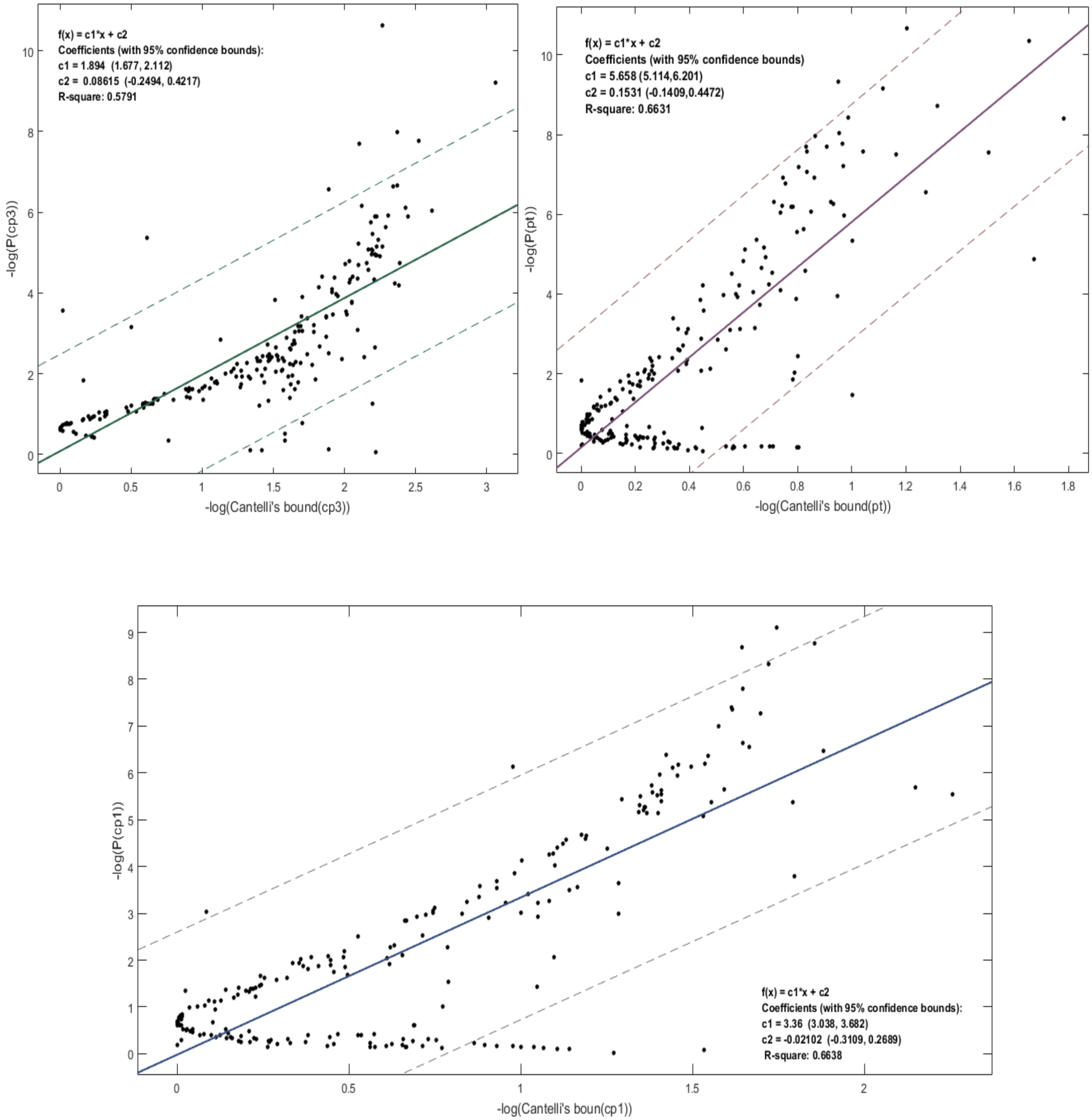


Figure II.1 Relationship between the negative logarithmic transformation of the Cantelli's bounds and p values computed by the permutation method by using samples of 10050000 codes for the standard genetic code and 235 amino acid properties. It was used the codon-block-based model with unbiased-substitution weighting. Top left: Third codon position, Top right: All codon positions, Bottom: First codon position.

Table II.3 Spearman rank correlation coefficients between Cantelli's upper bounds and empirical estimates of probability of obtaining codes more robust than the standard genetic code computed from a random sample of 10050000 codes (p values) for amino acid properties with p values smaller than the values shown in the first column. Cp1: Cantelli's upper bounds computed under the block-based representation of edges connecting first codon positions, Cp2: Cantelli's upper bounds computed under the block-based representation of edges connecting second codon positions, Cp3: Cantelli's upper bounds computed under the block-based representation of edges connecting third codon positions.Cpt : Cantelli's upper bounds computed under the whole block-based representation.

P values	**Cp1**	**Cp2**	**Cp3**	**Cpt**
$<1*10^6$	0.9645	0.9636	0.8691	0.9130
$<3*10^6$	0.9836	0.8254	0.8439	0.9436
$<5*10^6$	0.9881	0.9483	0.8715	0.9552
$<7*10^6$	0.9087	0.4429	0.8887	0.8758
$<9*10^6$	0.6607	0.0836	0.8608	0.6022

Table II.4 Biased-weighted mean phenotypic change (rob) under the codon-based models, with stop codons=scale mean. The first 10 aa properties (from a total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than that of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon blocks with more than 1 codon. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**
Hydrophobicity (Miyazawa, p132)	0.2942	-11.5415	7.4512E-03	0.1347
Hydrophobicity (Kyte, p125)	0.3509	-11.5015	7.5028E-03	0.1871
Transmembrane alpha-helix (p28)	0.3952	-11.3401	7.7162E-03	0.1927
hydrophobicity(Cowan,p117)	0.3397	-11.1858	7.9289E-03	0.1363
Transmembrane alpha-helix (p35)	0.3507	-11.1856	7.9291E-03	0.1444
Hydrophobicity(Parker, p135)	0.3392	-11.0920	8.0625E-03	0.1185
Long-range contacts (p164)	0.3170	-11.0837	8.0745E-03	0.1153
Solvent accesible surface (p44)	0.3286	-10.9339	8.2953E-03	0.1355
Polar requirement (p149)	0.2998	-10.9131	8.3267E-03	0.1427
Transmembrane helix turn (Wilson, p219)	0.3960	-10.9060	8.3374E-03	0.1387

Table II.5 Biased-weighted mean phenotypic change (rob) under the codon-based models, with stop codons=mean suppressor. The first 10 aa properties (from a total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than that of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**
Hydrophobicity (Miyazawa, p132)	0.2820	-11.7264	7.2198E-03	0.1387
Hydrophobicity (Kyte, p125)	0.3358	-11.7027	7.2488E-03	0.1815
Transmembrane alpha-helix (p28)	0.3811	-11.5252	7.4722E-03	0.2121
Long-range contacts (p164)	0.2981	-11.4091	7.6239E-03	0.1089
Transmembrane alpha-helix (p35)	0.3388	-11.3485	7.7048E-03	0.1508
hydrophobicity(Cowan,p117)	0.3291	-11.3367	7.7207E-03	0.1427
Solvent accesible surface (p44)	0.3089	-11.2582	7.8280E-03	0.1355
Hydrophobicity (Parker, p135)	0.3357	-11.1424	7.9903E-03	0.1371
Polar requirement (p149)	0.2880	-11.1216	8.0199E-03	0.1460
Transmembrane helix turn (Wilson, p219)	0.3868	-11.0195	8.1680E-03	0.1500

Table II.6 The first 10 aa properties (from a Total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. It was used the **Unbiased -weighted mean phenotypic changes**. Codon-based representation, codon stop=scale mean. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values Lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, Table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**
Hydrophobicity (Miyazawa, p132)	0.8919	-9.0042	1.2184E-02	0.1952
Hydrophobicity (Kyte, p125)	1.0809	-8.7702	1.2834E-02	0.1500
Hydrophobicity (Cornette,p115)	1.0094	-8.6967	1.3049E-02	0.1984
Transmembrane Alpha-Helix(p28)	1.1856	-8.6462	1.3200E-02	0.2016
Hydrophobicity (Parker, p135)	0.9668	-8.5717	1.3428E-02	0.1863
Hydrophobicity (Wilson, p147)	1.0773	-8.5286	1.3562E-02	0.1992
Transmembrane Alpha-Helix (p35)	1.0312	-8.5219	1.3583E-02	0.1653
Long-range contacts (p164)	0.9310	-8.2054	1.4635E-02	0.1468
Transmembrane Helix turn (Wilson, p219)	1.1428	-8.1344	1.4888E-02	0.1734
Solvent accesible Surface (p44)	0.9586	-8.0303	1.5270E-02	0.1653

Table II.7 The first 10 aa properties (from a Total of 235) in increasing order of Cantelli's bounds (CB) for the standard code. It was used **the Biased-weighted mean phenotypic change and codon-based representation,** stop codon=mean suppressor. Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values Lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, Table IX.1).

Amino acid properties	**rob**	**score**	**CB**	**Pr(AC)**
Hydrophobicity (Miyazawa, p132)	0.8658	-9.3064	1.1414E-02	0.2153
Hydrophobicity (Kyte, p125)	1.0544	-9.0531	1.2054E-02	0.1565
Hydrophobicity (Cornette,p115)	0.9815	-8.9945	1.2210E-02	0.2016
Transmembrane Alpha-Helix (p28)	1.1583	-8.9336	1.2375E-02	0.2145
Hydrophobicity (Parker, p135)	0.9473	-8.7788	1.2810E-02	0.2202
Transmembrane Alpha-Helix(p35)	1.0070	-8.7730	1.2826E-02	0.1806
Hydrophobicity (Wilson, p147)	1.0582	-8.7529	1.2884E-02	0.2444
Long-range contacts (p164)	0.9013	-8.6374	1.3227E-02	0.1548
Solvent accesible Surface (p44)	0.9258	-8.4711	1.3744E-02	0.1653
Small linker propensity (p174)	0.8234	-8.4511	1.3808E-02	0.1750

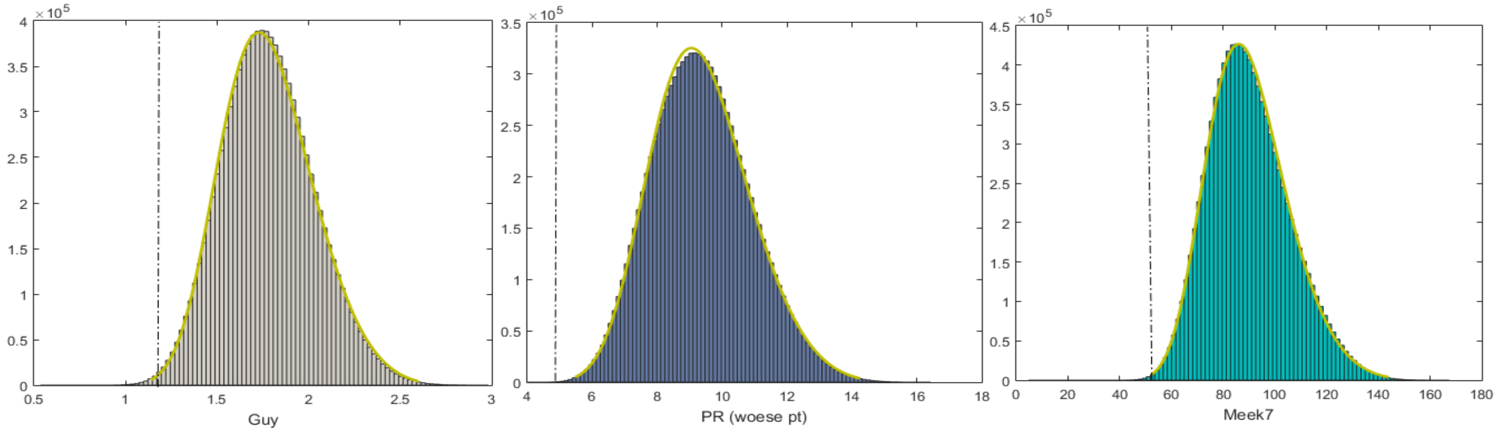


Figure II.2 Distributions of the unbiased-weighted mean change for the amino acid properties, Hydrophobicity/polarity scales: Left: Guy, middle: Polar Requirement, right: Meek PH 7.4, random sample of 107 codes. Dot-dashed line: Standard genetic code.

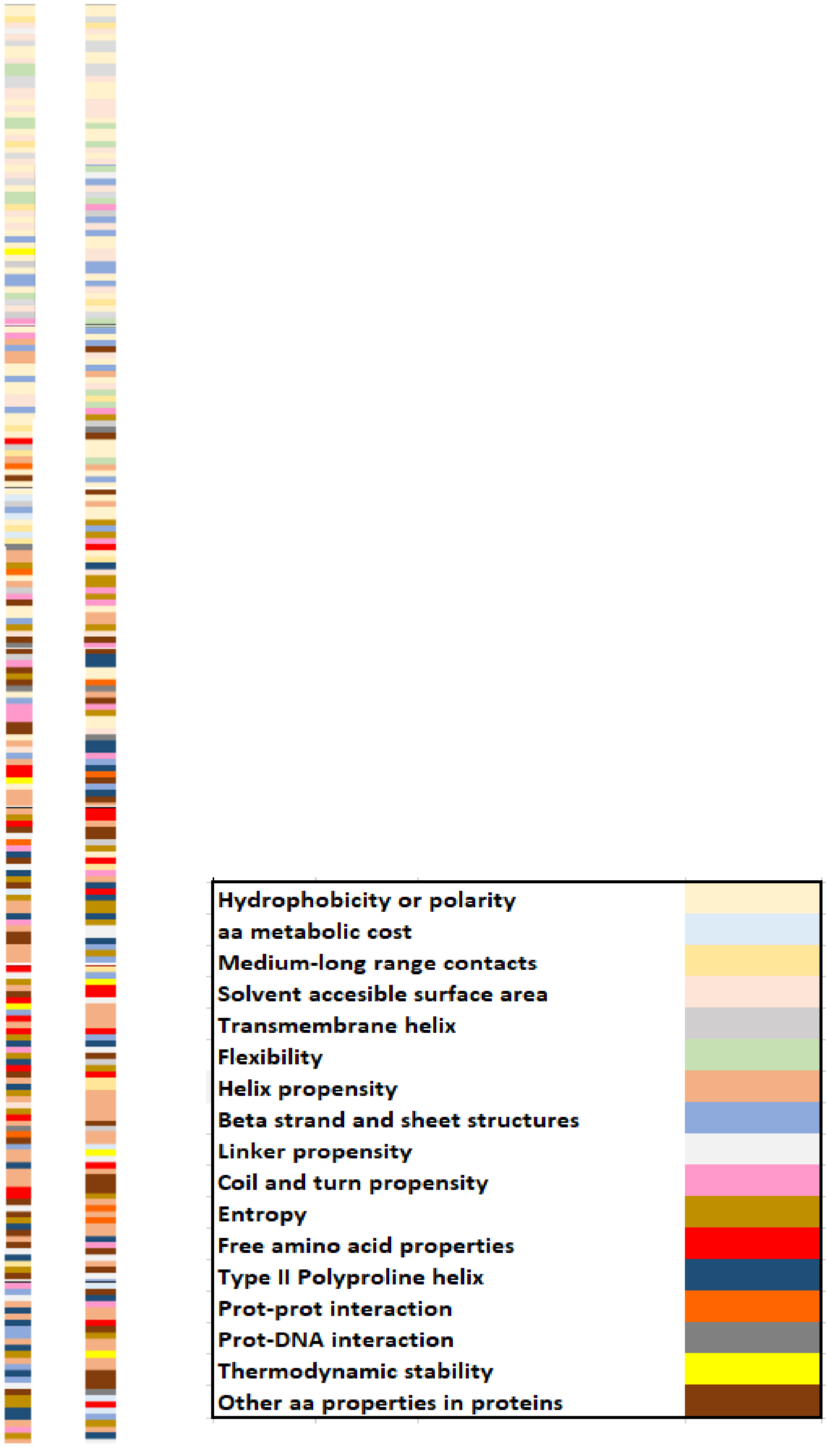


Figure II.3 Amino acid property sorted in increasing order of average ranks. The average ranks were calculated from lists of amino acid properties sorted in increasing order of Cantelli's upper bounds for the 23 genetic codes. The block-based (left bar) and codon-based (right bar) models were used with biased-weighted mean change in each of the 235 amino acid properties.

Appendix III. Robustness of the Natural genetic codes

Table III.1 Biased-weighted mean phenotypic change (rob) under the codon-based model with codon stop=mean suppressor. Scores for 23 genetic codes sorted in increasing order of their Cantelli's bounds (CB). The phenotype is expressed in terms of hydrophobicity (Miyazawa's contact energies), CB: Cantelli's bound, Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values lower than those of the standard code, These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon, The numbers in parentheses (In the footnotes and in first column of the table) indicate the NCBI translation table.

Genetic codes	rob	score	CB	Pr(AC)
The standard genetic Code (1)*	0.28203	-11.72635	0.007220	0.13871
Traustochytrium mitochondrial Code (23)	0.28187	-11.69421	0.007259	0.13952
The Invertebrate Mitochondrial Code (5)	0.29371	-11.64130	0.007325	0.12344
The Mold, Protozoan, and Coelenterate Mitochondrial Code (4)**	0.29217	-11.64034	0.007326	0.12698
The ascidian Mitochondrial Code (14)	0.29214	-11.63001	0.007339	0.12969
Trematode Mitochondrial Code (21)	0.29052	-11.62715	0.007343	0.12937
The Echinoderm and Flatworm Mitochondrial Code (9)	0.28957	-11.62656	0.007343	0.13226
The Euplotid Nuclear Code (10)	0.29576	-11.60929	0.007365	0.12742
The Vertebrate Mitochondrial Code (2)	0.29146	-11.60551	0.007370	0.14375
The Ciliate, Dasycladacean and Hexamita Nuclear Code (6)	0.29714	-11.59581	0.007382	0.16475
Pterobranchia Mitochondrial Code (24)	0.30263	-11.58188	0.007400	0.14683
Mesodinium Nuclear Code (29)	0.29290	-11.56558	0.007420	0.15328
Peritrich Nuclear Code (30)	0.30127	-11.56368	0.007423	0.17049
The Alternative Flatworm Mitochondrial Code (14)	0.29409	-11.56191	0.007425	0.13607
Cephalodiscidae Mitochondrial UAA-Tyr Code (33)	0.30797	-11.50799	0.007494	0.15806
Karyorelict Nuclear Code(27)***	0.30937	-11.49534	0.007511	0.15397
Candidate Division SR1 and Gracilibacteria Code (25)	0.29979	-11.49077	0.007517	0.15565
Blastocrithidia Nuclear Code (31)	0.31385	-11.46045	0.007556	0.16032
Pachysolen tannophilus Nuclear Code (26)	0.30987	-11.17548	0.007943	0.12742
Chlorophycean Mitochondrial Code(16)	0.33478	-11.11967	0.008023	0.15806
Scenedesmus obliquus Mitochondrial Code (22)	0.33302	-11.11534	0.008029	0.16210
The alternative yeast nuclear Code (12)	0.34563	-10.68904	0.008676	0.13790
The Yeast Mitochondrial Code (3)	0.35169	-10.10900	0.009691	0.13594

* The Bacterial, archaeal and plant plastid Code (11) has the same parameter values as the standard code,

** Full name: The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (4)

***The Condylostoma nuclear Code (28) has the same parameter values as the Karyorelict nuclear code,

Table III.2 Unbiased-weighted mean change (rob) and Scores for 25 genetic codes sorted in increasing order of their Cantelli's bounds (CB). **The codon-based representations for the genetic codes, stop codon=mean suppressor**. The phenotype is expressed in terms of Hydrophobicity (Miyazawa's contact energies Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values Lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers in parentheses (In the footnotes and first column of the table) indicate the NCBI translation table.

Genetic codes	**rob**	**score**	**CB**	**Pr(AC)**
Traustochytrium mitochondrial Code (23)	0.85300	-9.36144	0.011282	0.21210
The Vertebrate Mitochondrial Code (2)	0.86813	-9.32004	0.011381	0.21250
The standard genetic Code (1)*	0.86580	-9.30636	0.011414	0.21532
The ascidian Mitochondrial Code (14)	0.88825	-9.18577	0.011713	0.21875
The Mold, Protozoan, and Coelenterate Mitochondrial Code (4)**	0.89093	-9.18378	0.011718	0.21429
The Invertebrate Mitochondrial Code (5)	0.89649	-9.17636	0.011736	0.21875
The Echinoderm and Flatworm Mitochondrial Code (9)	0.88100	-9.17451	0.011741	0.21210
Trematode Mitochondrial Code (21)	0.88417	-9.17219	0.011747	0.21190
The Euplotid Nuclear Code (10)	0.89760	-9.15891	0.011781	0.21855
Pterobranchia Mitochondrial Code (24)	0.91742	-9.10677	0.011914	0.22063
The Ciliate, Dasycladacean and Hexamita Nuclear Code (6)	0.90343	-9.09904	0.011934	0.23115
The Alternative Flatworm Mitochondrial Code (14)	0.89021	-9.07784	0.011989	0.22295
Candidate Division SR1 and Gracilibacteria Code (25)	0.89093	-9.07509	0.011997	0.23871
Peritrich Nuclear Code (30)	0.91135	-9.07474	0.011997	0.23934
Mesodinium Nuclear Code (29)	0.88797	-9.06236	0.012030	0.23689
Cephalodiscidae Mitochondrial UAA-Tyr Code (33)	0.92801	-8.99894	0.012198	0.22903
Pachysolen tannophilus Nuclear Code (26)	0.86726	-8.95448	0.012318	0.22984
Karyorelict Nuclear Code (27)***	0.93378	-8.95350	0.012321	0.23810
Blastocrithidia Nuclear Code (31)	0.94249	-8.92413	0.012401	0.24603
Scenedesmus obliquus Mitochondrial Code (22)	0.93503	-8.82857	0.012667	0.24194
Chlorophycean Mitochondrial Code (16)	0.94537	-8.78141	0.012802	0.24919
The alternative yeast nuclear Code (12)	0.90696	-8.60083	0.013338	0.23629
The Yeast Mitochondrial Code (3)	0.86712	-8.10033	0.015012	0.23594

* The Bacterial, archaeal and plant plastid Code (11) has the same parameter values as the standard code.

** Full name: The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (4)

***The Condylostoma nuclear Code (28) has the same parameter values as the Karyorelict nuclear code.

Table III.3 Biased-weighted mean phenotypic change (rob) under the partial codon-based models for the standard code. Stop codons: Mean suppressor. The 10 amino acid properties correspond to those of table. p1: first codon position. p2: second codon position. p3: third codon position. rob: standard code robustness. Cb: Cantelli's upper bound. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

	p1			**p2**			**p3**		
Amino acid properties	**rob**	**score**	**cb**	**rob**	**score**	**cb**	**rob**	**score**	**cb**
Hydrophobicity (Miyazawa, p132)	0.3108	-6.9674	0.0202	0.4843	1.4680	0.3170	0.0510	-9.8135	0.0103
Hydrophobicity (Kyte, p125)	0.4303	-6.5378	0.0229	0.4294	-0.8661	0.5714	0.1477	-9.3821	0.0112
Transmembrane alpha-helix (p28)	0.4486	-6.6576	0.0221	0.5399	0.2433	0.9441	0.1548	-9.4166	0.0112
Long-range contacts (p164)	0.3828	-6.3023	0.0246	0.3676	-0.5141	0.7910	0.1438	-9.2975	0.0114
Transmembrane alpha-helix (p35)	0.4816	-5.9250	0.0277	0.4124	-0.6192	0.7228	0.1222	-9.4593	0.0111
Hydrophobicity (Cowan,117)	0.5131	-5.5656	0.0313	0.4010	-0.5805	0.7479	0.0733	-9.7114	0.0105
Solvent accesible Surface (44)	0.4591	-5.6809	0.0301	0.3440	-1.0748	0.4640	0.1235	-9.3841	0.0112
Hydrophobicity (Parker, 135)	0.3679	-6.6128	0.0224	0.4766	1.0436	0.4787	0.1627	-9.2237	0.0116
Polar requirement (149)	0.5087	-4.7334	0.0427	0.2825	-1.5205	0.3020	0.0729	-9.7542	0.0104
Transmembrane helix turn (Wilson, 219)	0.6383	-5.0493	0.0377	0.4221	-0.9087	0.5477	0.1000	-9.5909	0.0108

Table III.4 Unbiased- weighted mean change (rob) and Scores for 25 genetic codes sorted in increasing order of their Cantelli's bounds (CB). **The codon-based representation of the genetic codes stop codon=scale mean**. The phenotype is expressed in terms of Hydrophobicity (Miyazawa's contact energies). Pr(AC): Proportion of artificial genetic codes with Cantelli's bound values Lower than those of the standard code. These codes were generated by all possible reassignments of one codon in the synonymous codon sets with more than 1 codon. The numbers in parentheses (In the footnotes and first column of the table) indicate the NCBI translation table.

Genetic codes	**rob**	**score**	**CB**	**Pr(AC)**
The Ciliate, Dasycladacean and Hexamita Nuclear Code (6)	0.9108595	-9.018596	0.01214549	0.192623
The standard genetic Code (1)*	0.8919356	-9.004190	0.01218391	0.195161
Peritrich Nuclear Code (30)	0.9188763	-8.994349	0.01221026	0.203279
The Echinoderm and Flatworm Mitochondrial Code (9)	0.8979796	-8.991426	0.01221811	0.179839
The ascidian Mitochondrial Code (14)	0.9064351	-8.989967	0.01222202	0.185938
Trematode Mitochondrial Code (21)	0.9011511	-8.989617	0.01222296	0.180952
The Mold, Protozoan, and Coelenterate Mitochondrial Code (4)**	0.9091272	-8.989099	0.01222436	0.187302
The Alternative Flatworm Mitochondrial Code (14)	0.8990188	-8.983252	0.01224008	0.182787
The Invertebrate Mitochondrial Code (5)	0.9146736	-8.982887	0.01224106	0.187500
Mesodinium Nuclear Code (29)	0.8950918	-8.977502	0.01225557	0.206557
The Euplotid Nuclear Code (10)	0.9158360	-8.965212	0.01228878	0.188710
Karyorelict Nuclear Code (27)***	0.9337789	-8.953503	0.01232055	0.203968
Blastocrithidia Nuclear Code (31)	0.9424901	-8.924134	0.01240077	0.213492
Pterobranchia Mitochondrial Code (24)	0.9355838	-8.917568	0.01241881	0.191270
Cephalodiscidae Mitochondrial UAA-Tyr Code (33)	0.9368719	-8.907624	0.01244621	0.193548
The Vertebrate Mitochondrial Code (2)	0.9055493	-8.897803	0.01247335	0.204688
Candidate Division SR1 and Gracilibacteria Code (25)	0.9090002	-8.868576	0.01255467	0.218548
Pachysolen tannophilus Nuclear Code (26)	0.8933405	-8.642784	0.01321044	0.210484
Chlorophycean Mitochondrial Code (16)	0.9611024	-8.609776	0.01331057	0.215323
Scenedesmus obliquus Mitochondrial Code (22)	0.9564830	-8.565025	0.01344815	0.215323
Traustochytrium mitochondrial Code (23)	0.9062763	-8.499936	0.01365208	0.208871
The alternative yeast nuclear Code (12)	0.9329488	-8.292858	0.01433250	0.214516
The Yeast Mitochondrial Code (3)	0.8850806	-7.894931	0.01579032	0.212500

* The Bacterial, archaeal and plant plastid Code (11) has the same parameter values as the standard code.

** Full name: The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (4)

***The Condylostoma nuclear Code (28) has the same parameter values as the Karyorelict nuclear code.

Table III.5 Biased-weighted mean phenotypic change (rob) under the partial codon-based models for the standard code. Stop codons: Scale Mean. The 10 amino acid properties correspond to those of table. p1: first codon position. p2: second codon position. p3: third codon position. rob: standard code robustness. cb: Cantelli's upper bound. The numbers after p in parentheses (first column) indicate the position in the list of amino acid properties (Appendix, table IX.1).

	p1			**p2**			**p3**		
Amino acid properties	**rob**	**score**	**cb**	**rob**	**score**	**cb**	**rob**	**score**	**cb**
Hydrophobicity (Miyazawa, p132)	0.3383	-6.7348	0.0216	0.4867	1.5386	0.2970	0.0575	-9.7743	0.0104
Hydrophobicity (Kyte, p125)	0.4694	-6.2536	0.0249	0.4312	-0.8033	0.6078	0.1522	-9.3571	0.0113
Transmembrane alpha-helix (p28)	0.4752	-6.4689	0.0233	0.5435	0.3485	0.8917	0.1669	-9.3565	0.0113
Hydrophobicity (Cowan,p117)	0.5492	-5.2796	0.0346	0.4019	-0.5464	0.7701	0.0680	-9.7404	0.0104
Transmembrane alpha-helix (p35)	0.5240	-5.6039	0.0309	0.4133	-0.5908	0.7413	0.1149	-9.4967	0.0110
Hydrophobicity (Parker, p135)	0.3828	-6.4931	0.0232	0.4801	1.1155	0.4456	0.1546	-9.2688	0.0115
Long-range contacts (p164)	0.4430	-5.7532	0.0293	0.3667	-0.4531	0.8297	0.1414	-9.3081	0.0114
Solvent accesible surface (p44)	0.5149	-5.1806	0.0359	0.3438	-1.0222	0.4890	0.1272	-9.3590	0.0113
Polar requirement (p149)	0.5325	-4.4931	0.0472	0.2848	-1.4540	0.3211	0.0821	-9.6903	0.0105
Transmembrane helix turn (Wilson, p219)	0.6634	-4.8727	0.0404	0.4249	-0.8594	0.5752	0.0996	-9.5927	0.0108

aa	codon	Fmcodon	Fmccs	aa Neighborhood(p1-p2-p3)
F	TTT	0.3899	0.3899	leu Ile val ser tyr cys Leu Leu
	TTC	0.3899		leu Ile val ser tyr cys Leu Leu
L*	TTA	0.3497	0.3606	ile val ser stop stop phe phe
	TTG	0.3093		met val ser stop trp phe phe
	CTT	0.3653		phe ile val pro his arg
	CTC	0.3653		phe ile val pro his arg
	CTA	0.3823		Ile val pro gln arg
	CTG	0.3836		met val pro gln arg
S*	AGT	0.2526	0.3240	cys arg gly ile thr asn arg
	AGC	0.2526		cys arg gly ile thr asn arg
	TCT	0.3809		pro thr ala phe tyr cys
	TCC	0.3809		pro thr ala phe tyr cys
	TCA	0.3448		pro thr ala stop stop leu
	TCG	0.3384		pro thr ala stop trp leu
Y	TAT	0.2970	0.2970	his asn asp phe ser cys stop stop
	TAC	0.2970		his asn asp phe ser cys stop stop
Stop*	TAA			gln lys glu Leu Ser Tyr
	TAG			gln lys glu leu ser trp tyr
C	TGT	0.8955	0.8955	arg ser gly phe ser tyr stop trp
	TGC	0.8955		arg ser gly phe ser tyr stop trp
Stop*	TGA			arg arg gly leu ser trp cys
W	TGG	0.8122	0.8122	arg arg gly leu ser stop cys stop
P*	CCT	0.3080	0.3061	ser thr ala leu hist arg
	CCC	0.3080		ser thr ala leu hist arg
	CCA	0.3041		ser thr ala leu gln arg
	CCG	0.3041		ser thr ala leu gln arg
H	CAT	0.1909	0.1909	tyr asn asp leu pro arg gln
	CAC	0.1909		tyr asn asp leu pro arg gln
Q	CAA	0.1621	0.1621	lys glu stop leu pro arg his
	CAG	0.1621		lys glu stop leu pro arg his
R*	CGT	0.4189	0.2699	cys ser gly Leu Pro his
	CGC	0.4189		cys ser gly Leu Pro his
	CGA	0.0559		stop glys Leu pro gln
	CGG	0.3593		trp gly leu pro gln
	AGA	0.0873		gly stop ile thr lys ser
	AGG	0.2353		trp gly met thr lys ser
I*	ATT	0.4046	0.4114	leu val phe thr asn ser met
	ATC	0.4046		leu val phe thr asn ser met
	ATA	0.4250		leu leu val thr asn ser met
M	ATG	0.4560	0.4560	Ile val leu leu thr lys arg il
T*	ACT	0.2677	0.2733	ala pro ser ile asn ser
	ACC	0.2677		ala pro ser ile asn ser
	ACA	0.2721		ala pro ser ile arg lys
	ACG	0.2856		ala pro ser met arg lys
N	AAT	0.1858	0.1858	asp his tyr ile thr ser Lys
	AAC	0.1858		asp his tyr ile thr ser Lys
K*	AAA	0.1875	0.1896	gln glu stop ile thr arg asn
	AAG	0.1916		gln glu stop met thr arg asn
V*	GTT	0.2056	0.1989	ile leu phe ala asp gly
	GTC	0.2056		ile leu phe ala asp gly
	GTA	0.1883		ile leu leu ala glu gly
	GTG	0.1961		met leu leu ala glu gly
A*	GCT	0.1422	0.1419	val asp gly thr pro ser
	GCC	0.1422		val asp gly thr pro ser
	GCA	0.1416		val glu gly thr pro ser
	GCG	0.1416		val glu gly thr pro ser
D	GAT	0.1569	0.1569	asn his tyr val ala gly glu
	GAC	0.1569		asn his tyr val ala gly glu
E	GAA	0.0786	0.1896	lys gln stop val ala gly asp
	GAG	0.0786		lys gln stop val ala gly asp
G*	GGT	0.1945	0.1536	ser arg cys val ala asp
	GGC	0.1945		ser arg cys val ala asp
	GGA	0.0444		arg arg stop val ala glu
	GGG	0.1687		arg arg trp val ala glu

Table III.6 Neighborhood structure of the standard code. The third column shows the values of the weighted mean phenotypic change associated to nucleotide substitutions for each codon and the Miyazawa's contact energies; the boxes with only one color correspond to Sets of synonymous codons with the same amino acid neighborhood. The amino acids with (*) in the first column are coded by subsets of codons with different amino acid neighborhoods. The fourth column shows the weighted mean phenotypic change for each set of synonymous codons. The amino acids shown in the fifth column are encoded by the codons that differ by one nucleotide substitution in a given codon position (p1, p2, p3) with respect to the corresponding codons shown in the second column. The weighting scheme is based on the type and position of the nucleotide substitution.

Appendix IV. Robustness of archaeal and bacterial genomes

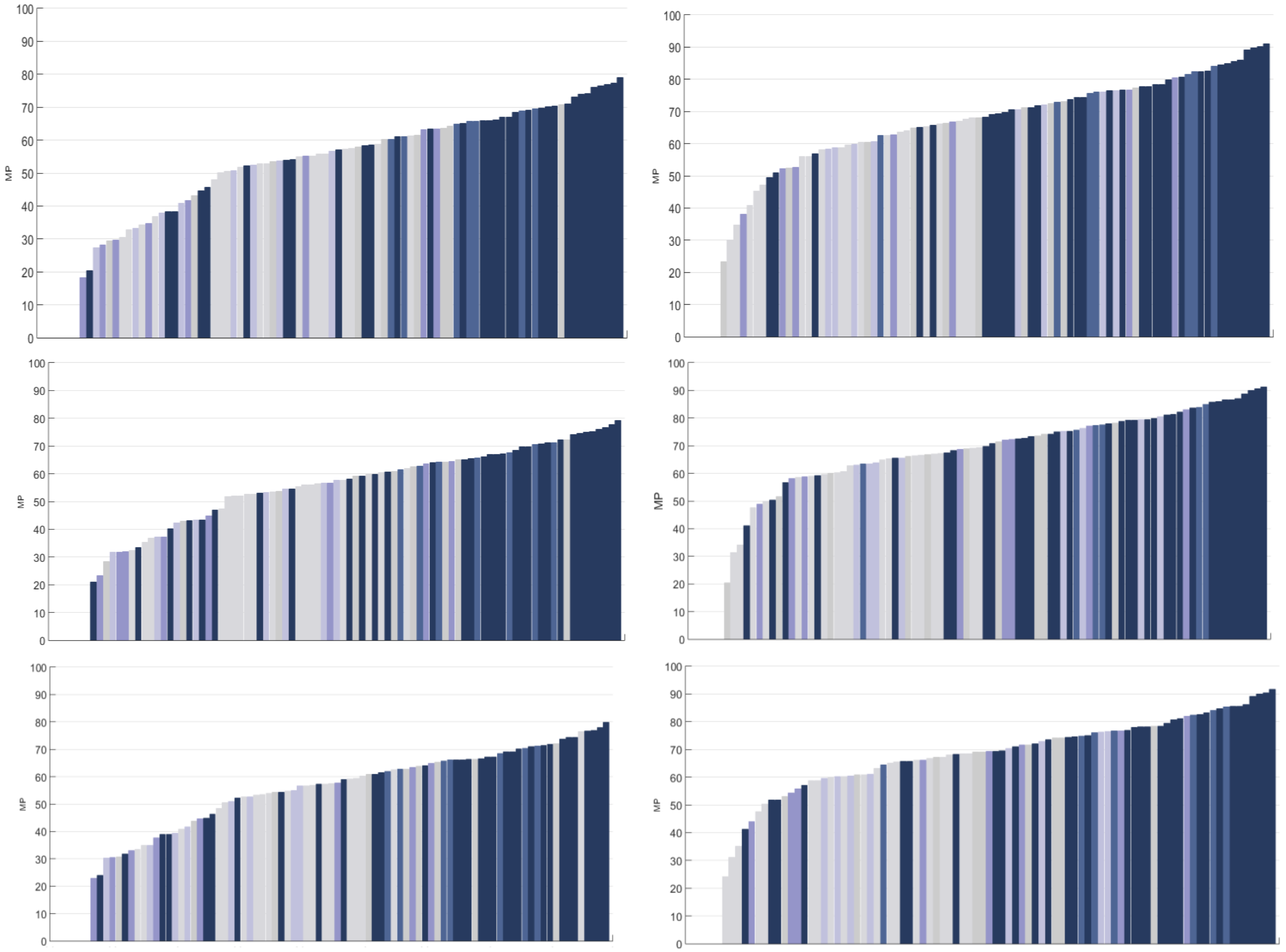


Figure IV.1 The 84 amino acid property scales sorted in order of increasing values of the median Optimization percentages for thermophilic (N=324) genomes. Each color corresponds to a given type of amino acid property. Each figure corresponds to the Minimization percentage medians computed under one of the standard code representations. Top left: The standard code models based on the whole set of codons with unbiased weighting and scale mean values assigned to stop codons. Top right: The standard code models based on the whole set of codons with biased weighting and scale mean values assigned to stop codons. Middle left: The standard code models based on sense codons and unbiased weighting, Middle right: The standard code models based on sense codons and biased weighting, Bottom left: The standard code models based on the whole set of codons with unbiased weighting and the values assigned to stop codons according to the "mean suppressor" method. Bottom right: The standard code models based on the whole set of codons with biased weighting and the values assigned to stop codons according to the "mean suppressor" method.

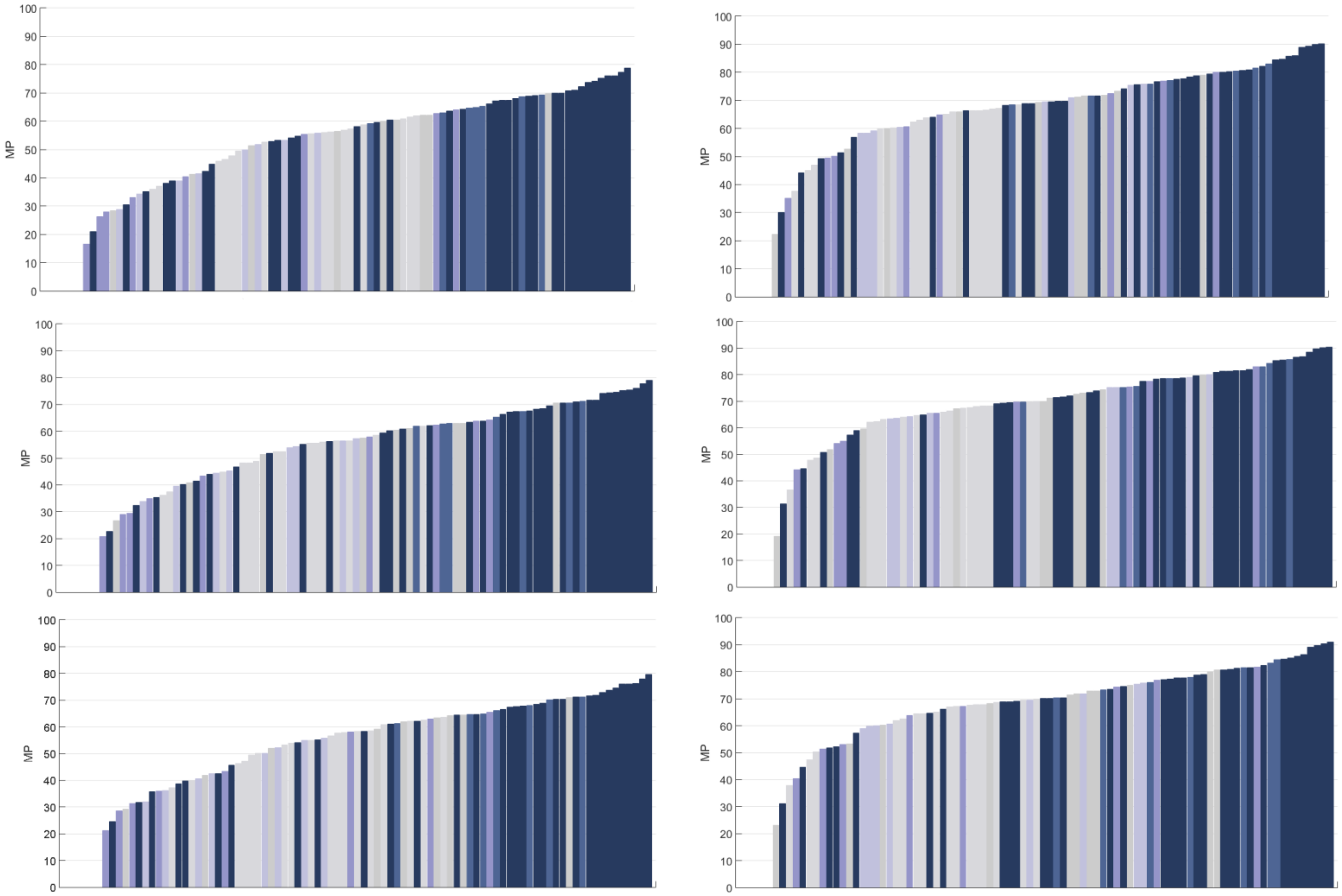


Figure IV.2 The 84 amino acid property scales sorted in order of increasing values of the median Optimization percentages for non-thermophilic (N=418) genomes. Each color corresponds to a given type of amino acid property. Each figure corresponds to the Minimization percentage medians computed under one of the standard code representations. Top left: The standard code models based on the whole set of codons with unbiased weighting and scale mean values assigned to stop codons. Top right: The standard code models based on the whole set of codons with biased weighting and scale mean values assigned to stop codons. Middle left: The standard code models based on sense codons and unbiased weighting, Middle right: The standard code models based on sense codons and biased weighting, Bottom left: The standard code models based on the whole set of codons with unbiased weighting and the values assigned to stop codons according to the "mean suppressor" method. Bottom right: The standard code models based on the whole set of codons with biased weighting and the values assigned to stop codons according to the "mean suppressor" method.

Table IV.1 The Spearman's rank correlation coefficients between median Optimization percentages and median scores computed under the whole (W) and partial (P) standard code models Thermophiles and Non-thermophiles. MS: Standard code representation based on sense codons and unbiased weighting, MSW: Standard code representation based on sense codons and biased weighting. M0W: The codon-based model of the standard code with biased weighting and scale mean values assigned to stop codons. M0: The codon-based model of the standard code with unbiased weighting and scale mean values assigned to stop codons. MMW: The codon-based model of the standard code with biased weighting and the values assigned to stop codons according to the "mean suppressor" method. MM: The codon-based model of the standard code with unbiased weighting and the values assigned to stop codons according to the "mean suppressor" method. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324).

Name	Thermic status	W Models	P Models				
			P1	P2	P3	Trans	Tranv
MS	Thermophiles	-0.9607	-0.9795	-0.8893	-0.8421	-0.9621	-0.7781
	Non-thermophiles	-0.9586	-0.9799	-0.8824	-0.8127	-0.9474	-0.7795
MSW	Thermophiles	-0.9472	-0.9099	-0.9795		-0.9575	-0.9360
	Non-thermophiles	-0.9442	-0.9298	-0.9829		-0.9197	-0.9410
M0	Thermophiles	-0.9405	-0.9706	-0.8528	-0.8496	-0.9776	-0.8966
	Non-thermophiles	-0.9273	-0.8994	-0.8672	-0.8187	-0.9592	-0.8943
M0W	Thermophiles	-0.9364	-0.9646	-0.9297		-0.9713	-0.9339
	Non-thermophiles	-0.9231	-0.9549	-0.9308		-0.9345	-0.9240
MM	Thermophiles	-0.9335	-0.9737	-0.8399	-0.8445	-0.9752	-0.8839
	Non-thermophiles	-0.9174	-0.9622	-0.8515	-0.8018	-0.9576	-0.8839
MMW	Thermophiles	-0.9247	-0.9561	-0.9788		-0.9640	-0.9268
	Non-thermophiles	-0.9299	-0.9444	-0.9788		-0.9322	-0.9276

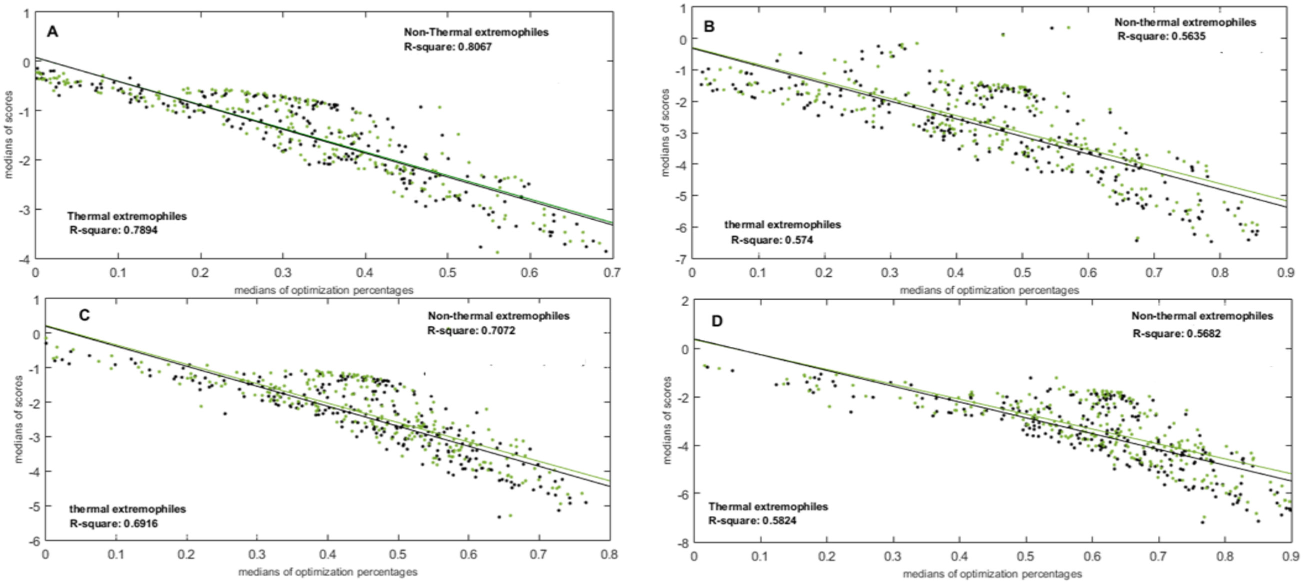
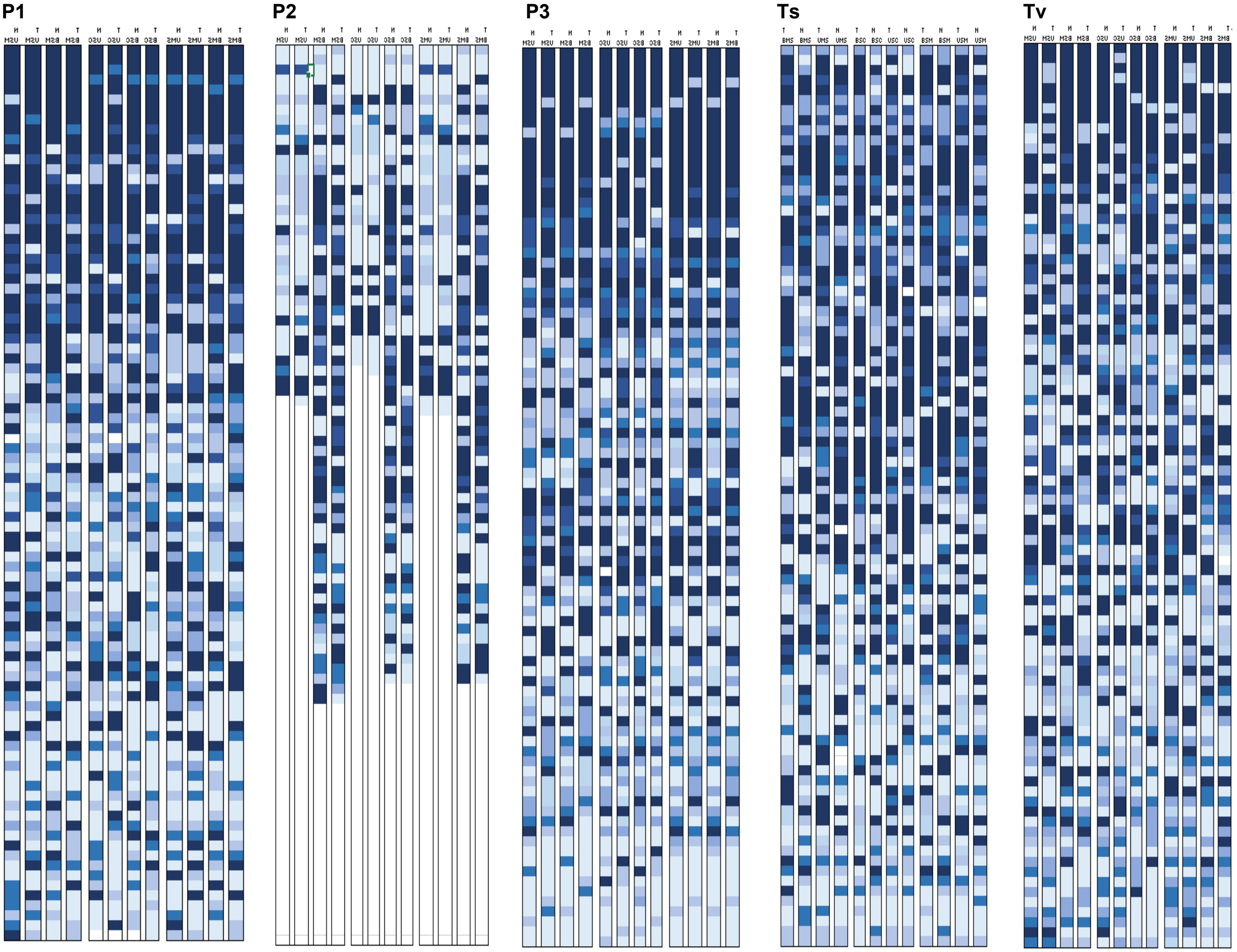


Figure IV.3 Medians of scores versus medians of optimization percentages for 235 aa properties and 742 genomes, Green: Non-thermophiles, Black: thermophiles. Weighting with synonymous codon usage. **A:** Unbiased-weighted mean phenotypic change, model based on sense codons. **B:** Biased-weighted mean phenotypic change, model based on sense codons, **C:** Unbiased-weighted Mean phenotypic change, codon-based model with codon Stop=Mean Suppressor, **D:** Biased-Weighted mean phenotypic change, codon-based model with Codon Stop= Mean suppressor.



Hydrophobicity/Polarity/Hydrophilicity	■
Accessible solvent area	■
Medium-Long range contacts	■
Flexibility	■
Entropy	■
Volume/Radius	■
Others	■
OP=0	□

Figure IV.4 The 84 amino acid property scales sorted in order of decreasing values (from top to bottom) of the median Optimization percentages (OP) for thermophilic (T) and non-thermophilic (N) genomes. Each color corresponds to a given type of amino acid property. The amino acid properties for which OP are equal to zero are represented by the color white (see the legend of the figure). Each of the 5 sets corresponds to the median Minimization percentages computed under one of the 5 types of partial standard code models: P1, first codon position; P2: second codon position; P3: third codon position, Ts: transitions and Tv, transversions. Each of the 12 columns for each set corresponds to one of the following models, from left to right: Column 1: The codon-based models of the standard code with biased weighting and the values assigned to stop codons according to the "mean suppressor" method (MMW) in thermophiles. Column 2: The MMW models in non-thermophiles. Column 3: The codon-based models of the standard code with unbiased weighting and the values assigned to stop codons according to the "mean suppressor" method (MM). Column 4: The MM models in non-thermophiles. Column 5: The standard code models based on sense codons and biased weighting (MSW) in thermophiles. Column 6: The MSW models in non-thermophiles. Column 7: The standard code models based on sense codons and unbiased weighting (MS) in thermophiles. Column 8: The MS models in non-thermophiles. Column 9: The codon-based models of the standard code with biased weighting and scale mean values assigned to stop codons (M0W) in thermophiles. Column 10: The M0W models in non-thermophiles. Column 11: The codon-based models of the standard code with unbiased weighting and scale mean values assigned to stop codons (M0) in thermophiles. Column 12: The M0 models in non-thermophiles. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324).

Appendix V. Comparison between thermophilic and non-thermophilic prokaryotes

Table V.1 The amino acid properties with significant coefficients (coeff) in order of increasing p values (4th and 5th columns) from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients (coeff) represent the fixed effects for the scores corresponding to the **unbiased-weighted mean phenotypic change**. It was used **the whole representation based on sense codons** of the standard code. Weighting with synonymous codon usage. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff**	**se**	**pvalue**	**AIC**
Long-range contacts (p163)	-0.2402(-0.2888 , -0.1950)	-0.1343(-0.1814 , -0.0959)	-255.32	10.78	2.20E-16	257.26
Long-range contacts (p161)	-0.3835(-0.4420 , -0.3277)	-0.2643(-0.3123 , -0.2344)	-186.40	40.81	2.20E-16	267.58
Long-range contacts (p160)	-0.5963(-0.6665 , -0.5414)	-0.4723(-0.5260 , -0.4326)	-84.26	20.64	4.74E-12	296.67
Long-range contacts (p164)	-1.1371(-1.2489 , -1.0055)	-1.0125(-1.1121 , -0.9047)	-33.11	11.53	5.51E-08	314.93
Thermodynamic stability(p177)	-0.2588(-0.2972 , -0.2135)	-0.2365(-0.2655 , -0.1849)	-137.66	40.76	1.99E-07	317.42
Conformational entropy (p101)	-0.3389(-0.3893 , -0.2980)	-0.2900(-0.3195 , -0.2510)	-62.83	23.14	2.45E-06	322.25
Solvent accesible surface (p48)	-2.3950(-2.6392 , -2.1655)	-2.2737(-2.4692 , -1.9859)	-8.32	2.95	9.20E-05	329.16
Hydrophobicity(Kyte, p125)	-3.7660(-4.0952 , -3.4296)	-3.6491(-3.9175 , -3.2072)	-5.39	1.90	2.24E-04	330.84
Polar requirement (p149)	-1.1778(-1.3139 , -1.0519)	-1.1224(-1.2279 , -0.9637)	-13.76	4.90	2.93E-04	331.34
Flexibility (2FN, MS, p209)	-1.9569(-2.1728 , -1.7647)	-1.8829(-2.0510 , -1.6615)	-9.36	3.46	3.17E-04	331.49

2FN: Two flexible neighbors

Table V.2 The amino acid properties with significant coefficients (coeff) in order of increasing p values (4th and 5th columns) from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients (coeff) represent the fixed effects for the scores corresponding to the **unbiased weighted mean phenotypic change**. The **whole codon-based representation of the standard code** with stop codon=scale mean. Weighting with synonymous codon usage. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff**	**se**	**pvalue**	**AIC**
Conformational entropy (p99)	0.0242(0.0022 , 0.0486)	0.0693(0.0484 , 0.0933)	-107.49	34.975	1.41E-06	321.19
Flexibility (2RN, MS, p211)	-2.5810(-2.8057 , -2.3109)	-2.4174(-2.6286 , -2.1643)	-11.76	4.57E+00	1.59E-06	321.42
Polarity (p151)	-3.2215(-3.4855 , -2.9387)	-2.8958(-3.1471 , -2.6356)	-8.88	3.34E+00	3.13E-06	322.72
flexibility (p186)	-1.3595(-1.4998 , -1.1909)	-1.3328(-1.4625 , -1.1838)	-17.76	6.53E+00	6.11E-06	324.00
Thermodynamic stability (p177)	-1.2845(-1.4398 , -1.1619)	-1.1851(-1.3153 , -1.0644)	-18.12	6.91E+00	6.94E-06	324.25
Flexibility (RFN, ML, p185)	-1.6215(-1.7875 , -1.4422)	-1.5765(-1.7300 , -1.4110)	-14.68	5.46E+00	1.04E-05	325.02
Hydrophobicity (Wilson, p147)	-0.6908(-0.7747 , -0.6096)	-0.6737(-0.7484 , -0.5927)	-31.72	1.21E+01	1.23E-05	325.35
Flexibility (2FN, ML, p184)	-1.0762(-1.1997 , -0.9446)	-1.0681(-1.1793 , -0.9571)	-20.32	7.24E+00	1.46E-05	325.67
Flexibility (2RN, ML, p183)	-1.6740(-1.8489 , -1.4867)	-1.6218(-1.7815 , -1.4336)	-13.26	5.00E+00	1.65E-05	325.90
Long-range contacts (p164)	-2.8056(-3.0894 , -2.5109)	-2.6905(-2.9449 , -2.3879)	-7.84	3.02E+00	2.99E-05	327.03

RFN : Rigid and Flexible neighbors, 2RN: Two rigid neighbors, 2FN: Two flexible neighbors, 2RN: Two rigid neighbors

Table V.3 The amino acid properties with significant coefficients (coeff) in order of increasing p values (4th and 5th columns) from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients (coeff) represent the fixed effects for the scores corresponding to the **biased-weighted mean phenotypic change**. The **whole representation based on sense codons of the standard code**. Double-weighting with synonymous codon usage and base change position/type. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418). Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coeff	se	pvalue	aic
Long-range contacts (p162)	-1.2300(-1.3808 , -1.0963)	-0.9895(-1.1125 , -0.8781)	-25.50	8.85	8.16E-08	315.7
Long-range contacts (p161)	-2.0717(-2.3010 , -1.8509)	-1.7714(-1.9566 , -1.5589)	-12.90	4.96	1.46E-06	321.3
Long-range contacts (p163)	-2.1552(-2.3889 , -1.9299)	-1.8456(-2.0289 , -1.6252)	-12.10	4.74	2.81E-06	322.5
Long-range contacts (p160)	-2.8946(-3.2306 , -2.6047)	-2.6063(-2.8400 , -2.2680)	-6.94	2.53	3.37E-05	327.3
Long-range contacts (p164)	-3.5130(-3.9060 , -3.1783)	-3.2257(-3.4944 , -2.8109)	-5.91	2.12	3.72E-05	327.5
Hydrophobicity(Ponnuswamy, p137)	-5.1128(-5.5916 , -4.6683)	-4.7962(-5.1369 , -4.2313)	-4.43	1.56	4.91E-05	328.0
Thermodynamic stability(p177)	-1.1080(-1.2111 , -0.9664)	-1.0733(-1.2072 , -0.9577)	-17.90	6.40	5.65E-05	328.2
Polarity (p151)	-4.1722(-4.5054 , -3.8238)	-3.9518(-4.2663 , -3.5489)	-5.72	2.00	6.20E-05	328.4
Solvent accesible surface (p48)	-4.4668(-4.9202 , -4.0949)	-4.1781(-4.4820 , -3.6718)	-4.63	1.64	8.46E-05	329.0
Hydrophobicity (Guy, p120)	-4.4076(-4.8609 , -4.0084)	-4.1434(-4.4671 , -3.6340)	-4.53	1.60	1.18E-04	329.6

Table V.4 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. **The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the unbiased-weighted mean phenotypic change computed using the model based on the sense codons of the standard code.** The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages. (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coeff	se	pvalue	aic
Long-range contacts (p163)	0,3166(0,3003, 0,3303)	0,2900(0,2786, 0,3021)	306,1	61,1	1,793E-11	299.27
Long-range contacts (p161)	0,3191(0,3048, 0,3337)	0,2948(0,2840, 0,3057)	362,27	58,38	3,071E-11	300.33
Long-range contacts (p160)	0,3731(0,3611, 0,3868)	0,3501(0,3358, 0,3631)	188,47	47,98	1,035E-06	320.59
Long-range contacts (p165)	0,4472(,43080, 46124))	0,4328(,41528, 44693)	157,72	51,89	4,938E-06	323.60
Conformational Entropy(p100)	0,2833(0,2707, 0,2982)	0,2660(0,2487, 0,2786)	124,91	46,91	5,108E-05	328.05
Thermodynamic stability (p177)	0,3690(0,3534, 0,38p164)	0,3610(0,3460, 0,3750)	110,07	42,2	0,0002282	330.87
Hydrophobicity(Ponnuswamy,p137)	0,6550(0,6398, 0,6701)	0,6378(0,6162, 0,6521)	97,46	30,89	0,0002592	331.11
Polarity(Zimmerman, p151)	0,5813(0,5598, 0,6042)	0,5512(0,5295, 0,5718)	76,79	28,04	0,000559	332.55
Conformational Entropy(p101)	0,4289(0,4094, 0,4474)	0,4090(0,3886, 0,4280)	73,63	25,55	0,0005609	332.56
Accessible surface area (p48)	0,6279(0,6137, 0,6403)	0,6185(0,5969, 0,6341)	82,39	30,24	0,002068	334.97

Table V.5 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the unbiased-weighted mean phenotypic change computed using the model based on the whole set of codons of the standard code and scale means assigned to stop codons. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages. (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coef**	**se**	**pvalue**	**aic**
Flexibility(S_RR, p211)	0,5072(0,4783, 0,5206)	0,4985(0,4821, 0,5125)	171,62	86,74	1,995E-07	317.42
Polarity(Zimmerman, p151)	0,5836(0,5573, 0,6107)	0,5468(0,5295, 0,5649)	115,96	40,26	1,541E-06	344.35
Thermodynamic stability (p177)	0,3687(0,3540, 0,3850)	0,3597(0,3443, 0,3727)	140,25	49,36	0,00001787	326.05
Polar requirement (Woese, p149)	0,4566(0,4439, 0,4690)	0,4489(0,4342, 0,4595)	131,2	41,36	0,000112	329.53
Flexibility(ML_Ft,p186)	0,3319(0,3063, 0,3466)	0,3433(0,3308, 0,3545)	127,83	41,72	0,0001226	329.70
Flexibility(ML_RF,p185)	0,3786(0,3554, 0,3945)	0,3892(0,3758, 0,4001)	130,79	41,85	0,0001242	329.73
Flexibility(ML_2R,p183)	0,4077(0,3823, 0,4233)	0,4150(0,3960, 0,4276)	111,3	36,13	0,0001526	333.22
Long-range contacts(p164)	0,5514(0,5274, 0,5663)	0,5541(0,5340, 0,5692)	98,68	33,97	0,0001707	333.22
Hydrophobicity(Cornette, p115)	0,7303(0,7132, 0,7472)	0,7231(0,6991, 0,7379)	91,81	32,56	0,0002831	331.28
Flexibility(S_RF, p211)	0,5664(0,5487, 0,5790)	0,5570(0,5388, 0,5695)	100,18	33,18	0,0005203	332.42

Table V.6 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the biased-weighted mean phenotypic change computed using the model based on the sense codons of the standard code. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages. (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff**	**se**	**pvalue**	**aic**
Long-range contacts (p160)	0,4886(0,4740, 0,4985)	0,4420(0,4259, 0,4665)	175,44	49,29	4,27E-07	318,89
Long-range contacts (p161)	0,5806(0,5690, 0,5911)	0,5411(0,5261, 0,5607)	177,28	52,45	1,61E-06	321,45
Long-range contacts (p163)	0,5875(0,5725, 0,6022)	0,5495(0,5306, 0,5686)	131,17	47,12	9,01E-06	324,75
Long-range contacts (p165)	0,7203(,70890, 73088))	0,6958(,67940, 70848))	145,65	51,57	7,05E-05	328,66
Long-range contacts (p160)	0,6866(0,6725, 0,6957)	0,6544(0,6349, 0,6715)	111,88	40,05	0,0002826	331,27
Hydrophobicity(Kyte, p125)	0,9050(0,8965, 0,9124)	0,9018(0,8906, 0,9093)	178,3	49,57	0,000622	332,75
Accessible surface area (p48)	0,7753(0,7651, 0,7831)	0,7557(0,7381, 0,7686)	123,46	40,45	0,001706	334,62
Hydrophobicity(Guy, p120)	0,7889(0,7784, 0,8016)	0,7743(0,7594, 0,7877)	102,68	38,76	0,002244	335,12
Accessible surface area (p44)	0,8501(0,8402, 0,8589)	0,8438(0,8316, 0,8525)	132,3	41,55	0,002432	335,27
Hydrophobicity(Cornette, p115)	0,8694(0,8590, 0,8783)	0,8681(0,8532, 0,8764)	130,4	48,83	0,002685	335,45

Table V.7 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the scores corresponding to the unbiased-weighted mean phenotypic change computed using the first codon position model based on the whole set of codons of the standard code and scale means assigned to stop codons. The second and third columns contain the median scores for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff**	**se**	**pvalue**	**AIC**
Flexibility (ML_FF,p184)	-0,4465(-0,5525, -0,3937)	-0,3664(-0,4567, -0,3029)	-158,29	52,89	2,20E-16	255,46
Flexibility (ML_RF,p185)	-0,9775(-1,0853, -0,9114)	-0,8856(-0,9486, -0,8173)	-81,3	20,53	2,20E-16	271,74
Long-range contacts (p165)	-1,5498(-1,7183, -1,4359)	-1,2478(-1,3460, -1,1388)	-41,33	13,6	2,80E-13	291,11
Long-range contacts (p162)	-0,2806(-0,3431, -0,1374)	0,04270(-0,0643, 0,15914)	-69,311	19,99	3,64E-13	291,63
Long-range contacts (p161)	-1,1648(-1,3046, -1,0222)	-0,7830(-0,9100, -0,6758)	-50,38	17,13	5,34E-13	292,38
Flexibility(ML_2R,p183)	-1,2890(-1,4092, -1,1683)	-1,1767(-1,2743, -1,0825)	-46,6	12,09	2,64E-12	295,51
Long-range contacts (p163)	-0,4747(-0,5299, -0,3119)	-0,0980(-0,2034, 0,00270)	-54	12,65	1,79E-11	299,26
Transmembrane Helix propensity (p139)	-1,1512(-1,2845, -1,0474)	-1,0455(-1,1321, -0,9555)	-48,12	13,69	1,02E-10	302,66
Hydrophobicity(Meek, p130)	-1,0963(-1,2223, -1,0014)	-1,0472(-1,1351, -0,9668)	-43,79	12,75	2,71E-09	309,07
Flexibility(S_RR,p211)	-0,4465(-0,5525, -0,3937)	-0,3664(-0,4567, -0,3029)	-24,32	7,724	2,80E-09	309,14

Table V.8 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the scores corresponding to the biased-weighted mean phenotypic change computed using the first codon position model based on the whole set of codons of the standard code and scale means assigned to stop codons. The second and third columns contain the median scores for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324) (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coef	se	p value	aic
Flexibility (ML_FF,p184)	-1,3503(-1,5021, -1,2630)	-1,2041(-1,3080, -1,1318)	-45,94	13,10973	3,53E-12	296,08
Long-range contacts (p162)	-0,3833(-0,4456, -0,1837)	0,05307(-0,0756, 0,18102)	-51,214	12,723041	4,65E-12	296,63
Long-range contacts (p165)	-1,7005(-1,8908, -1,5415)	-1,3170(-1,4423, -1,1877)	-38,63	11,65838	7,25E-12	297,5
Long-range contacts (p161)	-1,2554(-1,3967, -1,0848)	-0,7548(-0,9287, -0,6408)	-36,03	10,85555	1,20E-11	298,48
Long-range contacts (p163)	-0,6697(-0,7377, -0,4616)	-0,1710(-0,3143, -0,0604)	-46,56	9,984566	4,02E-10	305,35
Long-range contacts (p160)	-0,9010(-1,0044, -0,7235)	-0,4439(-0,6172, -0,3705)	-39,85	10,977796	6,64E-10	306,33
Flexibility(ML_RF,p185)	-1,6666(-1,8346, -1,5654)	-1,5191(-1,6359, -1,4120)	-31,35	10,27387	1,09E-09	307,29
Hydrophobicity(Meek, p130)	-0,9012(-1,0117, -0,8260)	-0,8560(-0,9218, -0,7879)	-50,43	15,0149	3,24E-09	309,42
Accessible surface area (p48)	-2,5879(-2,8071, -2,4024)	-2,1823(-2,3440, -2,0223)	-18,5	6,480696	5,73E-09	310,53
Flexibility(ML_2R,p183)	-1,8160(-2,0025, -1,6780)	-1,6940(-1,8251, -1,5531)	-23,31	8,291051	3,46E-08	314,03

Table V.9 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the unbiased-weighted mean phenotypic change computed using the first codon position model based on the whole set of codons of the standard code and scale means assigned to stop codons. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coef	se	pvalue	aic
Flexibility(ML_FF,p184)	0,1100(0,0897, 0,1968)	0,0863(0,0707, 0,1196)	476,03	54,42	2,20E-16	263,75
Flexibility(ML_Ft,p186)	0,2127(0,1889, 0,3581)	0,1911(0,1741, 0,2114)	361,55	56,16	2,20E-16	259,91
Flexibility(S_FF,p209)	0,3647(0,3477, 0,3805)	0,3390(0,3252, 0,3513)	300,1	61,72	2,20E-16	269,17
Flexibility(S_RR,p211)	0,4064(0,3776, 0,4271)	0,3519(0,3402, 0,3678)	346,1	48,03	2,20E-16	346,1
Flexibility(S_Ft,p227)	0,3452(0,1433, 0,3584)	0,1689(0,1401, 0,1965)	72,99	11,4	2,20E-16	245,43
Long-range contacts (p165)	0,3593(0,3377, 0,3716)	0,2973(0,2800, 0,3180)	466,8	67,38	2,48E-16	277,28
Long-range contacts (p164)	0,6898(0,6739, 0,7088)	0,6728(0,6600, 0,6850)	272,5	47,27	3,53E-13	291,57
Polar requirement (Woese, p149)	0,2165(0,2082, 0,2277)	0,2043(0,1980, 0,2090)	579	98,75	4,56E-13	292,07
Long-range contacts (p161)	0,3290(0,2926, 0,3445)	0,2278(0,2009, 0,2597)	262,62	64,48	3,19E-12	295,89
Accessible surface area (p48)	0,4541(0,4349, 0,4690)	0,4037(0,3881, 0,4245)	312,9	106	8,94E-12	297,91

Table V.10 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the biased-weighted mean phenotypic change computed using the first codon position model based on the whole set of codons of the standard code and scale means assigned to stop codons. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coef	se	pvalue	aic
Flexibility(ML_Ft, p186)	0,4075(0,3918, 0,4313)	0,3878(0,3740, 0,4015)	253,75	60,39	8,47E-13	293,29
Hydrophobicity(Cornette, p115)	0,7332(0,7229, 0,7461)	0,7188(0,7103, 0,7278)	369,3	54,7	5,99E-12	297,12
Flexibility(ML_RF,185)	0,3999(0,3854, 0,4224)	0,3789(0,3667, 0,3929)	245,19	58,16	1,49E-11	298,91
Long-range contacts(p165)	0,3986(0,3674, 0,4111)	0,3082(0,2901, 0,3368)	206,3	65,01	2,58E-11	299,99
Long-range contacts(p164)	0,6884(0,6727, 0,7063)	0,6704(0,6578, 0,6848)	228,7	52,25	9,17E-11	302,47
Long-range contacts(p161)	0,3608(0,3051, 0,3779)	0,2229(0,1908, 0,2664)	143,15	32,78	1,10E-09	307,32
Flexibility(ML_2R,p183)	0,4573(0,4370, 0,4802)	0,4430(0,4249, 0,4584)	182,89	41,16	1,18E-09	307,45
Flexibility(S_RR,p211)	0,5418(0,5262, 0,5596)	0,5288(0,5162, 0,5412)	216,4	50,66	6,19E-09	310,68
Accessible surface area (p48)	0,5024(0,4772, 0,5177)	0,4272(0,4108, 0,4580)	175,49	58,51	7,45E-09	311,04

Table V.11 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the scores corresponding to the unbiased-weighted mean phenotypic change computed using the second codon position model based on the sense codons of the standard code. The second and third columns contain the median scores for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418). Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff**	**se**	**pvalue**	**aic**
Conformational entropy (p98)	-0,8184(-0,9081, -0,6576)	-0,4696(-0,5783, -0,4060)	-48,89	10,38	3,65E-07	318,59
Conformational entropy (p92)	-0,8822(-0,9732, -0,7058)	-0,4875(-0,6091, -0,4138)	-42,58	8,40	8,80E-07	320,28
Conformational entropy (p97)	-1,5286(-1,6733, -1,3886)	-1,2860(-1,3888, -1,1925)	-21,46	8,74	2,12E-06	321,98
Conformational Entropy (p94)	-1,0841(-1,1873, -0,9501)	-0,8407(-0,9320, -0,7677)	-26,06	9,83	6,40E-06	324,09
Conformational Entropy (p90)	-0,9866(-1,0750, -0,8621)	-0,7690(-0,8507, -0,6906)	-26,03	9,65	1,04E-05	325,01
Thermodynamic stability (p104)	-0,5726(-0,6364, -0,5227)	-0,5364(-0,5812, -0,4843)	-42,88	17,62	1,67E-05	325,93
Hydrophobicity (Abraham, p108)	-0,0545(-0,1157, -0,0000)	0,02604(-0,0042, 0,0607)	-51,82	20,09	1,97E-04	330,59
Conformational Entropy (p91)	-0,8088(-0,8887, -0,7169)	-0,6957(-0,7690, -0,6078)	-18,91	6,75	4,10E-04	331,97
PK_aa-NH2 (p199)	-0,6594(-0,7389, -0,4570)	-0,3193(-0,4144, -0,2373)	-21,79	9,43	4,60E-04	332,19
Hydrophobicity (p147)	-0,8815(-0,9557, -0,8287)	-0,8582(-0,9144, -0,7752)	-24,57	9,60	1,23E-03	334,01

Table V.12 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the scores corresponding to the biased-weighted mean phenotypic change computed using the second codon position model based on the sense codons of the standard code. The second and third columns contain the median scores for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coeff**	**se**	**pvalue**	**aic**
Hydrophobicity (Cowan, p117)	-0,5951(-0,6564, -0,5328)	-0,5336(-0,5751, -0,4796)	-107,87	34,86	7,83E-13	293,13
Hydrophobicity (Abraham, p108)	-1,1849(-1,3057, -1,0822)	-1,0343(-1,1077, -0,9686)	-57,64	18,76	1,07E-12	293,75
Polarity(Zimmerman, p151)	-0,8278(-0,9737, -0,6931)	-0,7383(-0,8315, -0,5919)	-56,64	17,47	1,26E-10	303,08
Hydrophobicity (Karplus, p122)	-0,6839(-0,8034, -0,5774)	-0,5948(-0,6677, -0,5038)	-53,97	17,19	9,86E-10	307,1
Hydrophobicity(Meek, p129)	-0,4961(-0,5777, -0,4013)	-0,4561(-0,5212, -0,3570)	-83,97	23,91	6,85E-09	310,88
Hydrophilicity(Parker, p135)	-0,1333(-0,1990, -0,0738)	-0,1004(-0,1392, -0,0423)	-77,41	25,99	7,03E-09	310,93
Polar requirement (Woese, p149)	-0,5092(-0,5590, -0,4705)	-0,4831(-0,5230, -0,4398)	-73,44	29,23	1,67E-07	317,08
Hydrophobicity (Levitt, p128)	-1,1671(-1,2576, -1,0836)	-1,0078(-1,0672, -0,9329)	-36,52	13,63	1,96E-07	317,39
Conformational entropy (p98)	-1,1276(-1,2111, -1,0221)	-0,9844(-1,0375, -0,9279)	-43,03	16,23	2,60E-07	317,93
Hydrophobicity(Ooi, p133)	-0,5175(-0,5544, -0,4692)	-0,5033(-0,5449, -0,4607)	-60,62	20,66	5,36E-07	319,33

Table V.13 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the biased-weighted mean phenotypic change computed using the second codon position model based on the whole set of codons of the standard code and values assigned to stop codons according to the “Mean suppressor” method. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles		se	pvalue	aic
Hydrophobicity(Cowan, p117)	0,2480(0,2122, 0,2815)	0,2174(0,1949, 0,2390)	178,39	53,02	5,45E-12	296,94
Hydrophilicity(Roseman,p138)	0,2668(0,2400, 0,3007)	0,2001(0,1815, 0,2214)	263,92	91,52	5,67E-12	297,01
Hydrophobicity(Abraham, p108)	0,4114(0,3702, 0,4646)	0,3557(0,3320, 0,3825)	112,2	30,14	3,82E-10	305,26
Polarity(Zimmerman, p151)	0,2560(0,2103, 0,3099)	0,2269(0,1785, 0,2528)	132,44	47,06	2,61E-09	309
Hydrophobicity(Karplus, p122)	0,2437(0,2012, 0,2959)	0,2093(0,1723, 0,2370)	124,97	41,28	3,78E-09	309,72
Hydrophobicity(Meek, p129)	0,1385(0,1075, 0,1698)	0,1277(0,0967, 0,1514)	161,883	52,09	2,01E-08	312,98
Long-range contacts(p163)	0,0725(0,0465, 0,0893)	0,0640(0,0500, 0,0793)	230,69	71,36	4,16E-08	314,38
hydrophobicity(Sandberg, p233)	0,3417(0,2998, 0,3821)	0,3125(0,2799, 0,3436)	104,01	32,82	4,51E-08	314,54
Hydrophobicity(Fauchere, p119)	0,5333(0,4879, 0,5731)	0,4877(0,4522, 0,5212)	79,3	24,02	1,56E-07	316,94
Polar requirement (Woese, p149)	0,1963(0,1876, 0,2078)	0,1918(0,1804, 0,2000)	268,44	93,43	1,68E-07	317,09

Table V.14 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the biased-weighted mean phenotypic change computed using the second codon position model based on the sense codons of the standard code. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coef**	**se**	**pvalue**	**aic**
Hydrophilicity(Roseman,p138)	0,2690(0,2408, 0,3026)	0,2019(0,1846, 0,2229)	283,05	193,1	4,79E-12	296,7
Long-range contacts(p163)	0,0638(0,0354, 0,0841)	0,0536(0,0345, 0,0763)	208,71	49,17	1,45E-09	307,9
Polarity(Zimmerman, p151)	0,2790(0,2359, 0,3349)	0,2535(0,2098, 0,2759)	141,78	56,03	4,55E-09	310,1
Hydrophilicity(Parker, p135)	0,0589(0,0352, 0,0904)	0,0456(0,0196, 0,0620)	191,42	72,98	1,47E-08	312,4
Hydrophobicity(Karplus, p122)	0,2738(0,2309, 0,3262)	0,2407(0,2089, 0,2669)	123,76	40,16	2,18E-08	313,1
Hydrophobicity(Meek, p129)	0,1744(0,1425, 0,2029)	0,1639(0,1321, 0,1828)	202,13	69,9	2,44E-07	317,8
Hydrophobicity(Fauchere, p119)	0,5441(0,5043, 0,5786)	0,5059(0,4737, 0,5378)	81,73	27,64	2,01E-06	321,9
Conformational Entropy(p92)	0,3452(0,3163, 0,3709)	0,2890(0,2698, 0,3097)	113,65	41,43	4,12E-06	323,3
Hydrophobicity(Ooi, p133)	0,2267(0,2057, 0,2417)	0,2211(0,2034, 0,2393)	113,05	33,94	1,28E-05	325,4
Conformational Entropy(p98)	0,3334(0,3018, 0,3558)	0,2895(0,2734, 0,3081)	112,8	42,69	1,72E-05	326

Table V.15 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. **The coefficients represent the fixed effects for the scores corresponding to the unbiased-weighted mean phenotypic change computed using Transversion models based on the sense codons of the standard code**. The second and third columns contain the median scores for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coeff	se	pvalue	aic
Conformational entropy (p102)	-0,0396(-0,0408, -0,0377)	-0,0363(-0,0370, -0,0354)	-6354,20	1680,16	2,52E-12	295,43
Thermodynamic stability (p177)	-0,0323(-0,0343, -0,0305)	-0,0294(-0,0304, -0,0278)	-4483,10	822,56	1,66E-09	308,12
Long-range contacts (p162)	-0,0113(-0,0122, -0,0104)	-0,0107(-0,0122, -0,0100)	-4970,61	55,47	7,75E-09	311,12
Conformational Entropy(p100)	-0,0217(-0,0224, -0,0184)	-0,0171(-0,0177, -0,0166)	-5259,32	64,53	3,72E-08	314,16
Medium-range contacts (p182)	-0,0706(-0,0715, -0,0692)	-0,0680(-0,0687, -0,0670)	-5409,70	86,31	1,79E-07	317,21
Long-range contacts (p160)	-0,0152(-0,0162, -0,0142)	-0,0149(-0,0166, -0,0139)	-3302,86	64,53	4,07E-07	318,79
Long-range contacts (p163)	-0,0089(-0,0106, -0,0074)	-0,0088(-0,0115, -0,0076)	-1955,47	92,20	2,77E-06	322,49
Long-range contacts (p164)	-0,0325(-0,0341, -0,0312)	-0,0330(-0,0352, -0,0321)	-1988,44	76,74	9,40E-05	329,2
Hydrophobicity (Ponnuswamy, 137)	-0,0760(-0,0765, -0,0754)	-0,0753(-0,0760, -0,0747)	-3081,00	118,60	7,11E-05	328,68
Conformational entropy (p90)	-0,0606(-0,0621, -0,0565)	-0,0526(-0,0548, -0,0512)	-1505,23	125,1	3,16E-03	335,74

Table V.16 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the scores corresponding to the unbiased-weighted mean phenotypic change computed using Transversion models based on the whole set of codons of the standard code and values assigned to stop codons according to the "Mean suppressor" method. The second and third columns contain the median scores for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	Thermophiles	Non-thermophiles	coeff	se	pvalue	aic
Conformational entropy (p102)	-0,0423(-0,0428, -0,0398)	-0,0385(-0,0391, -0,0379)	-7595,00	69,66	8,88E-12	297,89
Conformational entropy (p101)	-0,0232(-0,0237, -0,0219)	-0,0213(-0,0218, -0,0206)	-6119,00	85,44	1,31E-08	312,13
Thermodynamic stability (p177)	-0,0329(-0,0344, -0,0311)	-0,0300(-0,0309, -0,0287)	-4715,20	73,95	1,34E-08	312,19
Medium-range contacts (p182)	-0,0681(-0,0688, -0,0666)	-0,0654(-0,0660, -0,0646)	-12318,90	38,86	1,90E-08	312,86
Flexibility(S_RR,p211)	-0,0361(-0,0382, -0,0353)	-0,0357(-0,0389, -0,0348)	-2663,20	102,29	5,74E-07	319,46
Hydrophobicity(Wilson,p147)	-0,0129(-0,0135, -0,0123)	-0,0131(-0,0143, -0,0126)	-2933,86	192,33	1,56E-03	334,45
Flexibility(ML_2R,p183)	-0,0201(-0,0214, -0,0190)	-0,0210(-0,0229, -0,0199)	-1709,58	105,55	1,59E-03	334,48
Flexibility(S_Ft, p212)	-0,0337(-0,0352, -0,0330)	-0,0342(-0,0362, -0,0335)	-1609,46	91,62	3,46E-03	335,91
Flexibility(S_RF,p210)	-0,0482(-0,0494, -0,0470)	-0,0479(-0,0494, -0,0474)	-1499,12	374,21	4,96E-03	336,56

Table V.17 TvMsanstopda The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the unbiased-weighted mean phenotypic change computed using Transversion models based on the sense codons of the standard code. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coef**	**se**	**pvalue**	**aic**
Conformational Entropy(p101)	0,2905(0,2720, 0,3136)	0,2653(0,2526, 0,2782)	170,64	52,33	2,063E-08	313,02
Long-range contacts(p162)	0,1608(0,1412, 0,1700)	0,1572(0,1446, 0,1673)	258,2	67,04	2,288E-08	313,22
Thermodynamic stability (p177)	0,3473(0,3165, 0,3727)	0,3059(0,2927, 0,3204)	150,21	52,14	7,051E-08	315,41
Long-range contacts(p161)	0,1975(0,1723, 0,2073)	0,1940(0,1796, 0,2036)	150,17	42,95	0,00000245	322,25
Conformational Entropy(p100)	0,2703(0,2464, 0,2938)	0,2207(0,2046, 0,2363)	115,6	37,79	0,000005631	323,85
Conformational Entropy(p102)	0,4550(0,4262, 0,4874)	0,4101(0,3900, 0,4343)	85,96	29,27	0,000007816	324,47
Long-range contacts(p165)	0,2864(0,2669, 0,3026)	0,2907(0,2771, 0,3028)	115,6	37,95	0,00007771	328,84
Hydrophobicity(Ponnuswamy,p137)	0,4422(0,4211, 0,4615)	0,4352(0,4046, 0,4545)	78,53	26,83	0,0005628	332,56
Flexibility(S_RR,p211)	0,3230(0,3079, 0,3417)	0,3281(0,3102, 0,3424)	86,23	35,1	0,004065	336,2

Table V.18 The amino acid properties corresponding to the coefficients with the smallest p values. These coefficients, (coeff) sorted in order of increasing p values, are from the Three-level logistic mixed models that best discriminate between thermophiles and non-thermophiles. The coefficients represent the fixed effects for the optimization ratios (OP=OR*100) corresponding to the unbiased-weighted mean phenotypic change computed using Transversion models based on the whole set of codons of the standard code and scale means assigned to stop codons. The second and third columns contain the median Minimization percentages for all properties in both groups. The first and third quartiles are shown in parentheses, se: standard error, AIC: Akaike Information criterion. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). OP: Optimization percentages (The numbers after p in parentheses (first column) indicate the position in the list of amino acid indices (Appendix, Table IX.1).

Amino acid properties	**Thermophiles**	**Non-thermophiles**	**coef**	**se**	**pvalue**	**aic**
Flexibility(S_RR,p211)	0,2572(0,2346, 0,2778)	0,2541(0,2375, 0,2689)	299,67	79,72	5,714E-11	301,54
Thermodynamic stability (p177)	0,3357(0,3083, 0,3627)	0,2965(0,2854, 0,3095)	274,07	105	3,559E-10	305,12
Flexibility(ML_2R,p183)	0,1670(0,1438, 0,1859)	0,1794(0,1588, 0,1934)	220,8	61,21	1,425E-07	316,77
Conformational Entropy(p101)	0,3420(0,3274, 0,3669)	0,3169(0,3039, 0,3308)	124,47	36,59	0,000007099	324,29
Polar requirement (Woese, p149)	0,2210(0,2106, 0,2346)	0,2285(0,2177, 0,2373)	201,69	59,87	0,00004167	327,66
Flexibility(ML_RF,p185)	0,0770(0,0559, 0,0989)	0,0982(0,0791, 0,1202)	186,1247	60,45	0,000044	327,77
Hydrophobicity(Cornette, p115)	0,6067(0,5858, 0,6237)	0,5927(0,5656, 0,6088)	95,59	31,67	0,00004971	328
Hydrophobicity(Meek, p130)	0,2894(0,2733, 0,3008)	0,2830(0,2722, 0,2946)	159,51	53,59	0,0001666	330,28
Hydrophobicity(Bull, p114)	0,6743(0,6531, 0,6900)	0,6618(0,6385, 0,6796)	78,59	29,67	0,0005524	332,53
Hydrophobicity(Lawson, p126)	0,2688(0,2532, 0,2858)	0,2381(0,2181, 0,2488)	101,39	35,36	0,0007292	333,04

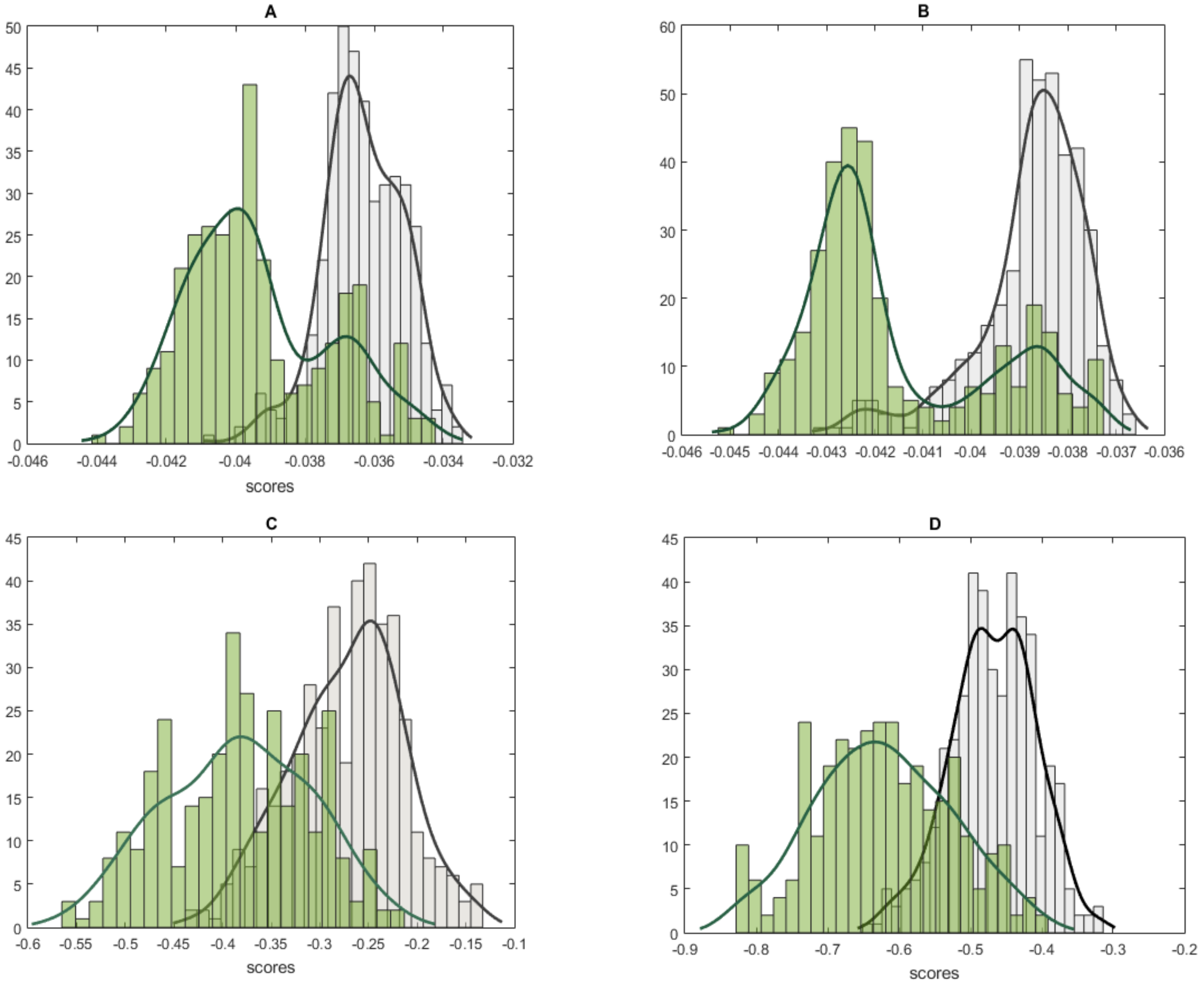


Figure V.1 Histograms of the scores computed under different standard code models and amino acid properties. Green: Thermophilic prokaryotes (N=324), Grey: Non-thermophilic prokaryotes (N=418). A: The transversion models based on sense codons, unbiased weighting and Conformational entropy (p102), B: The transversion models based on the whole set of codons with unbiased weighting, Conformational entropy (p102) and "Mean suppressor" values assigned to stop codons. C: The standard code models based on sense codons with unbiased weighting and Long-range contacts (p161). D: The second codon position models based on the whole set of codons, biased weighting, Hydrophilicity (Roseman, p138) and scale means assigned to stop codons.

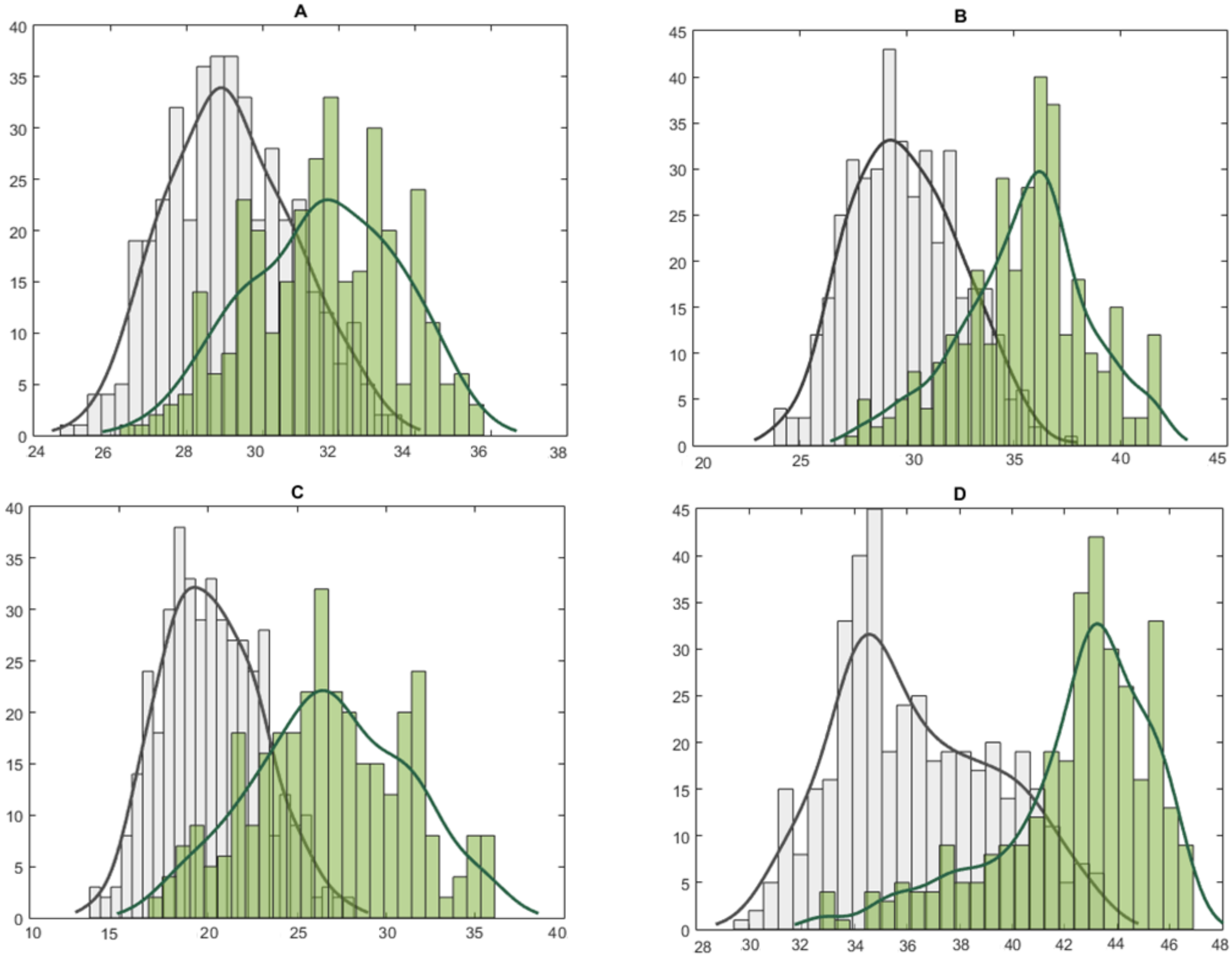


Figure V.2 Histograms of the Optimization percentages computed under different standard code models and amino acid properties. Green: Thermophilic prokaryotes (N=324), Grey: Non-thermophilic prokaryotes (N=418). A: The standard code models based on sense codons with unbiased weighting and long-range contacts (p163). B: The first codon position models based on the whole set of codons with unbiased weighting, Long-range contacts (p165) and scale means assigned to stop codons. C: The second codon position models based on the whole set of codons with biased weighting, Hydrophilicity (Roseman, p138) and values assigned to stop codons by using the "mean suppressor" method. D: The Transition models based on sense codons, unbiased weighting and Long-range contacts (p162).

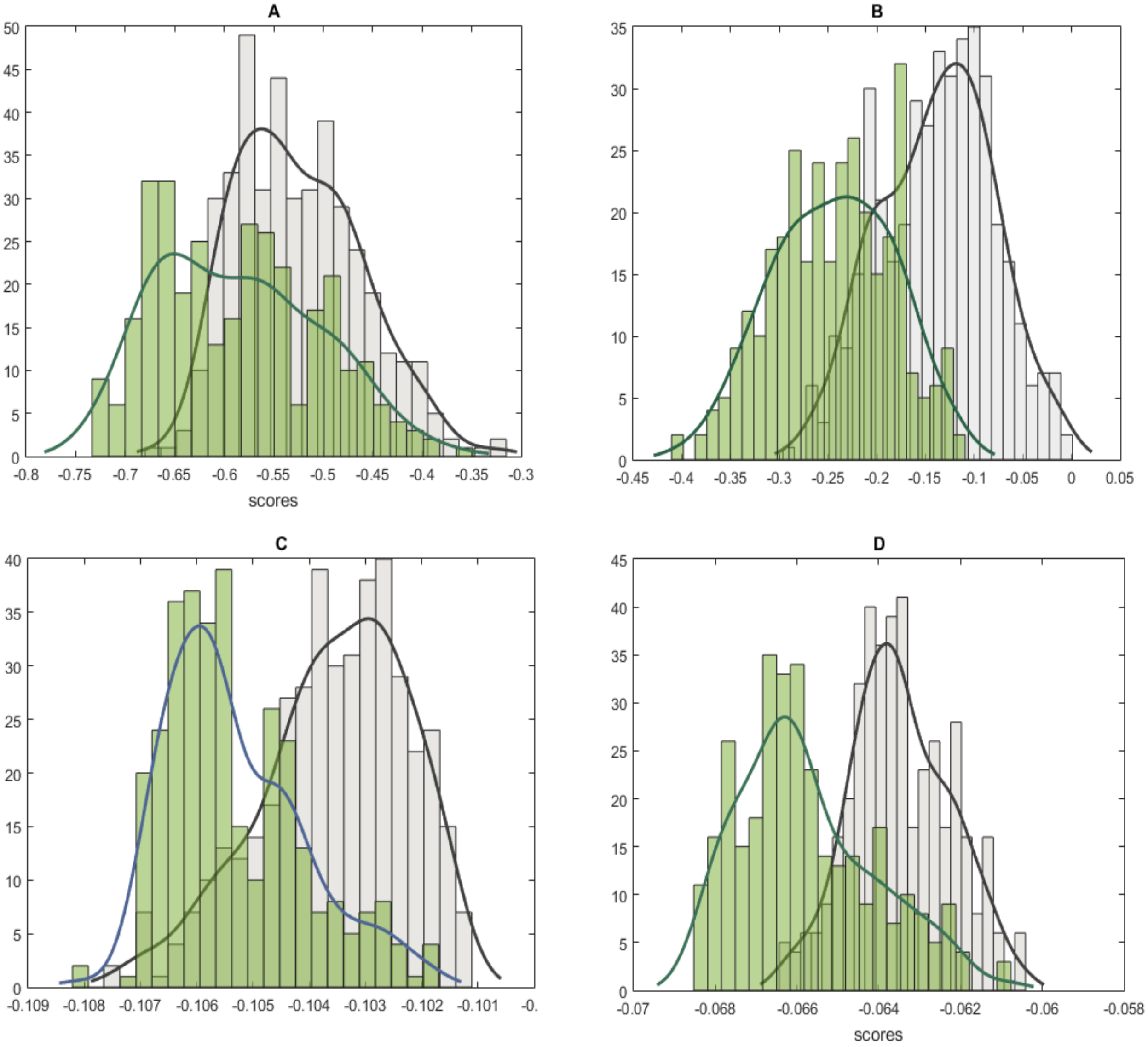


Figure V.3. Histograms of the scores computed under different standard code models and amino acid properties. Green: Thermophilic prokaryotes (N=324), Grey: Non-thermophilic prokaryotes (N=418). A: The second codon position models based on sense codons with biased weighting and Hydrophobicity (Cowan, p117), B: The standard code models based on sense codons with unbiased weighting and Long-range contacts (p163), C: The transversion models based on the whole set of codons with unbiased weighting, Hydrophobicity (Cornette, p115) and scale mean values assigned to stop codons, D: Transversion models based on the whole set of codons with unbiased weighting, Medium-range contacts (p182) and scale mean values assigned to stop codons.

Appendix VI. Robustness and base composition

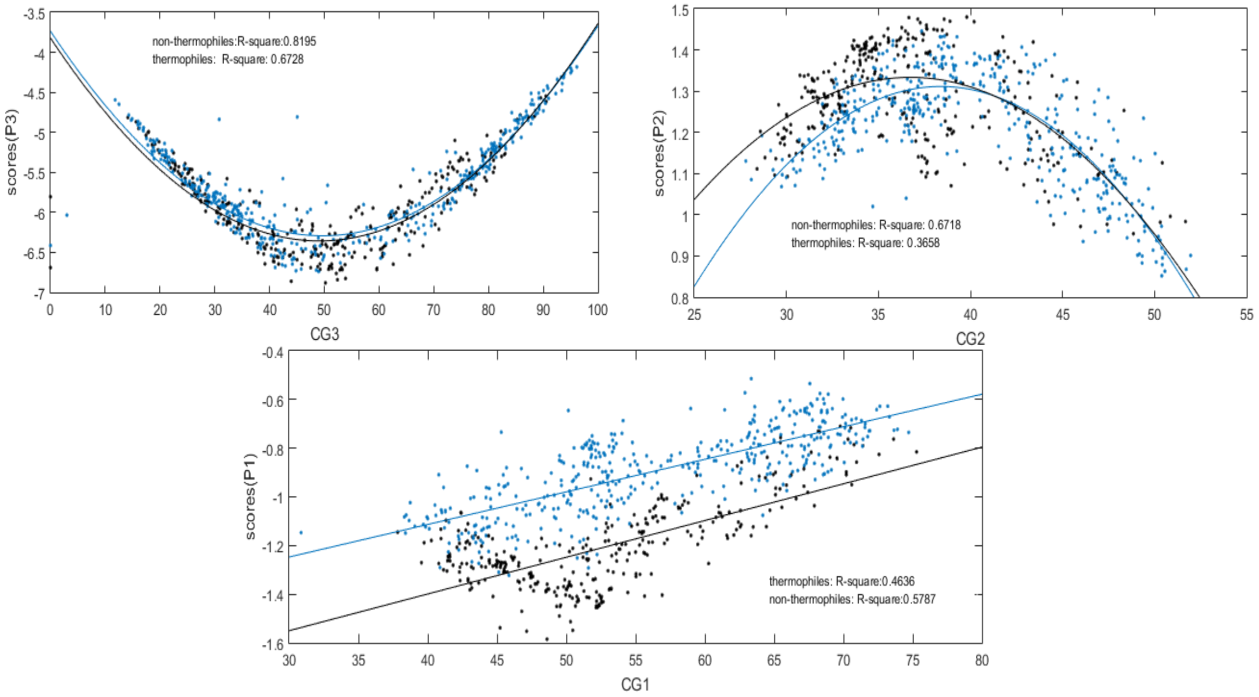


Figure VI.1 Top left: The scores corresponding to the unbiased-weighted mean change in hydrophobicity (Ponnuswamy, p137) for codon position 3 versus the GC content at this position. Top right: The scores corresponding to the unbiased-weighted mean change in hydrophobicity (Ponnuswamy, p137) for codon position 2 versus the GC content at this position. Bottom: The scores corresponding to the unbiased-weighted mean change in hydrophobicity (Ponnuswamy, p137) for codon position 1 versus GC content at this position. Representation based on sense codons of the standard code. Blue: genomes of Non-thermophiles, Black: genomes of thermophiles.

Table VI.1 Median amino acid composition for thermophiles and non-thermophiles. The prokaryotic genomes were classified into two categories: Non-thermophiles (mesophiles + psychrophiles, N=418) and thermophiles (Hyperthermophiles + thermophiles, N=324). The first and third quartiles shown in parentheses. AA: amino acid. *: significant according to Wilcoxon rank-sum test and False discovery rate: 0,01. #: p value between 0.05 and 0.07.

a	Thermophiles	Non-thermophiles	p values
L	0,1010(0,0961, 0,1078)	0,1026(0,0971, 0,1070)	0.6434
S*	0,0544(0,0495, 0,0594)	0,0604(0,0562, 0,0654)	3.0555e-27
P	0,0415(0,0369, 0,0455)	0,0405(0,0350, 0,0503)	0.9592
R	0,0541(0,0415, 0,0625)	0,0477(0,0396, 0,0659)	0.0686
I*	0,0747(0,0636, 0,0926)	0,0629(0,0490, 0,0747)	8.9875e-19
T*	0,0466(0,0445, 0,0493)	0,0541(0,0514, 0,0579)	9.7770e-69
K*	0,0704(0,0519, 0,0853)	0,0514(0,0329, 0,0675)	3.0097e-20
V*	0,0768(0,0707, 0,0843)	0,0690(0,0652, 0,0734)	1.7887e-25
A*	0,0703(0,0587, 0,0832)	0,0907(0,0737, 0,1197)	3.1018e-27
G	0,0715(0,0657, 0,0766)	0,0695(0,0642, 0,0823)	0.5715

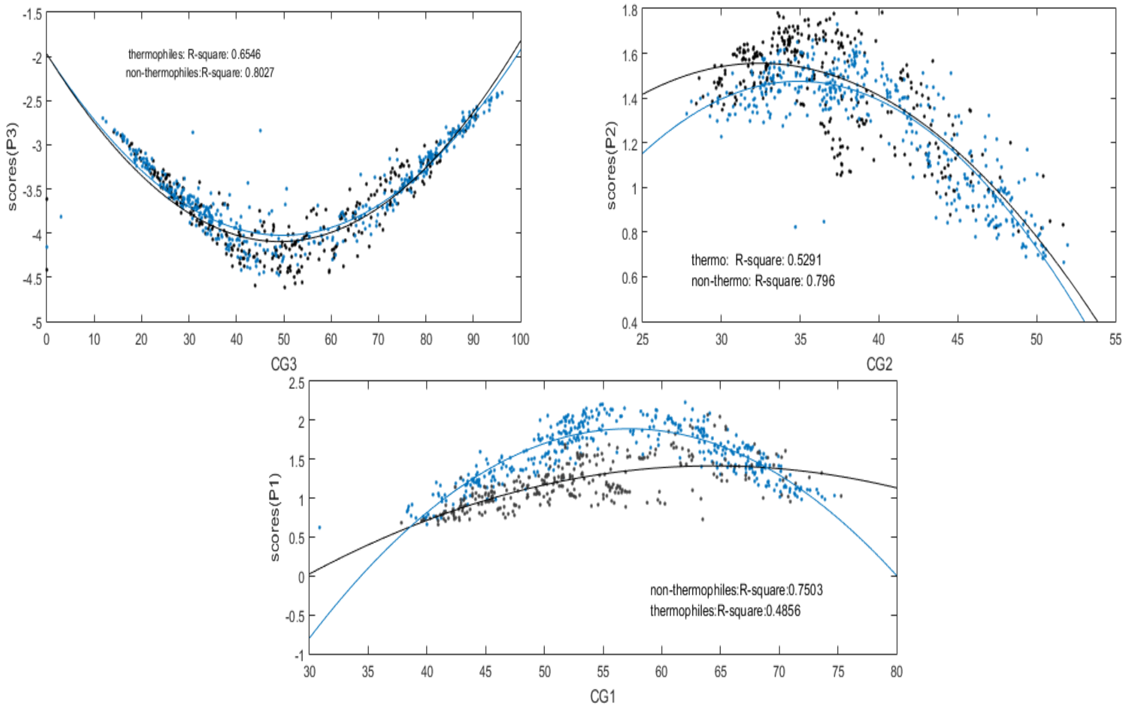


Figure VI.2 Top left: The scores corresponding to the unbiased -mean change in long-range contacts (161) for the codon position 3 versus the GC content at third codon position. Weighting with synonymous codon usage and base change position/type. Top right: The scores corresponding to the unbiased weighted mean change in long range contacts (161) for the codon position 2 versus the GC content at the second codon position, Bottom: The scores corresponding to the unbiased weighted mean change in long-range contacts (161) for the codon position 1 versus the GC content at first codon position. Representation based on sense codons of the standard code. Blue: genomes of Non-thermophiles, Black: genomes of thermophiles.

Table VI.2 Medians of the Unbiased-weighted mean change in long-range contacts (p161) for each synonymous codon block (third and fourth columns). The first and the third quartiles are shown in parentheses. The first column: the amino acids encoded by blocks with heterogeneous neighborhoods in the standard code (hb). Model of the standard code based on the sense codons. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). p values: Wilcoxon rank-sum test.

hb	**Thermophiles**	**Non-Thermophiles**	**p**
A	0.2421(0.2394,0.2439)	0.2425 (0.2406,0.2442)	0.0239
R	0.1796(0.1726,0.1989)	0.2582 (0.2343, 0.2781)*	3.2E-88
G	0.6352 (0.6326, 0.6388)	0.6265 (0.6235 ,0.6321)*	2.5E-54
I	0.8424 (0.8107,0.8541)	0.7908 (0.7811, 0.8072)*	1.3E-47
L	0.3683(0.3640.0.3707)	0.3693(0.3634,0.3751)	0.0024
K	0.4563(0.3970. 0.5014)	0.4964(0.3735, 0.5268)	0.0021
P	0.0702 (0.0688,0.0712)	0.0704 (0.0698, 0.0720)*	1.1E-09
S	0.6320(0.6239,	0.6267(0.6173,0.63p160)*	1.4E-10
T	0.1353(0.1259,0.1445)	0.1336(0.1253, 0.1399)	0.0048
V	0.7264(0.7182, 0.7349)	0.7294 (0.7229,0.7403)*	3.4E-06

Table VI.3 Medians of the biased-weighted mean change in long-range contacts (p161) for each synonymous codon block (third and fourth columns). The first and the third quartiles are shown in parentheses. The first column: the amino acids encoded by blocks with heterogeneous neighborhoods in the standard code (hb). Model of the standard code based on the sense codons. Non-thermophiles: Non-thermophilic prokaryotes (N=418), Thermophiles: thermophilic prokaryotes (N=324). p values: Wilcoxon rank-sum test.

hb	Thermophiles	Non-thermophiles	pvalue
A	0.1087 (0.1085,0.1089)	0.1088 (0.1086 0.1089)	0.0239
R	0.0912(0.0817,0.1282)	0.1935 (0.1623,0.2128)*	1.47E-75
G	0.3079(0.3064, 0.3100)	0.3028(0.3011,0.3061)*	2.51E-54
I	0.3056(0.3024,0.3067)	0.3004(0.2994,0.3020)*	1.32E-47
L	0.1735 (0.1730. 0.1739)	0.1738 (0.1732, 0.1742)*	1.66E-08
K	0.1282 (0.1223,0.1327)	0.1322 (0.1199,0.1352)	0.0021
P	0.0303 (0.0301, 0.0305)	0.0304 (0.0303, 0.0305)*	1.10E-09
S	0.2782 (0.2773 ,0.2795)	0.2776(0.2767,0.2785)*	2.72E-09
T	0.0750 (0.0701, 0.0792)	0.0743 (0.0703,0.0772)	0.0271
V	0.2657 (0.2617 0.2723)	0.2693 (0.2631 0.2764)*	3.16E-06

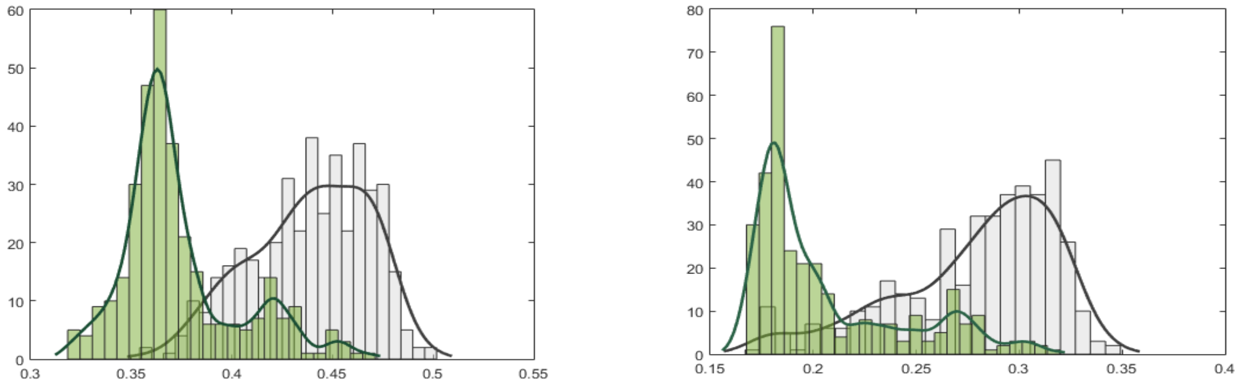


Figure VI.3 Histograms of the mean phenotypic change for the codon blocks corresponding to Arginine computed under models based on sense codons, Long-range contacts (p163) using either unbiased (left) or biased (right) weightings**.**

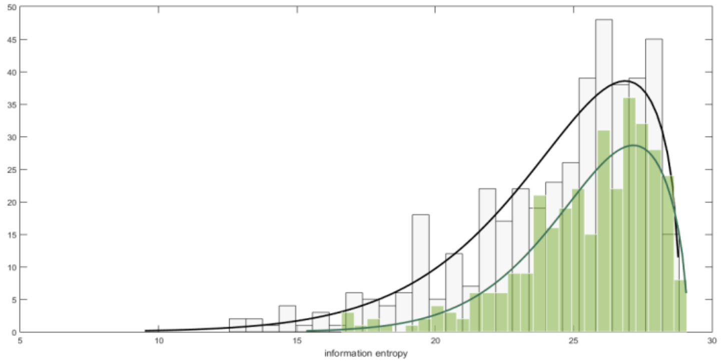


Figure VI.4 Histograms of information entropies computed for thermophilic and non-thermophilic genomes. Green: Thermophilic prokaryotes (N=324), Grey: Non-thermophilic prokaryotes (N=418).

Appendix VII. Developing the equation for the variance (terms $\boldsymbol{D_2}$ and $\boldsymbol{D_3}$)

Term $\boldsymbol{D_2}$

In the equation 14, we have actually the sum of two terms, D_2 and D_{2a}, corresponding to the two products of weights on the edges of E^g and E^p (eq 2 and 3), for example, for E^g these products are, $(y_{uv}\, y_{uv})$ and $(y_{uv}\, y_{vu})$. Since both graphs are undirected, D_2 is equal to D_{2a} (eq 1). As a result of this, this term is multiplied by two in (eq 14, see section Methods).

$$D_{2a} + D_2 = 2D_2 \tag{1}$$

$$D_2 = (T_2)(P_2)$$

$$T_2 = \left(\textstyle\sum_u^n \sum_v^n \gamma_{uv}^2\right) \tag{2}$$

$$P_2 = \left(\textstyle\sum_i^n \sum_j^n (p(i) - p(j))^4\right)$$

$$D_{2a} = (T_{2a})(P_{2a})$$

$$T_{2a} = \left(\textstyle\sum_u^n \sum_v^n \gamma_{uv}\, \gamma_{vu}\right) \tag{3}$$

$$P_{2a} = \left(\textstyle\sum_i^n \sum_j^n (p(i) - p(j))^2\, (p(j) - p(i))^2\right)$$

$$D_{2a} + D_2 = 2D_2$$

Since both weighted adjacency matrices are symmetrical, it follows that, $T_{2a} = T_2$ and $P_{2a} = P_2$. Hence, $D_2 = D_{2a}$ and thus, $D_2 + D_{2a} = 2D_2$.

Term $\boldsymbol{D_3}$

In both weight matrices, we have four products of the weights of two adjacent edges. For example, in the G^g weight matrix, we have the following products, $(\gamma_{uv}\gamma_{uh})$, $(\gamma_{vu}\gamma_{uh})$, $(\gamma_{uv}\gamma_{hu})$ and $(\gamma_{vu}\gamma_{hu})$. There are, hence, four terms, D_3, D_{3a}, D_{3b} and D_{3c}, each one corresponding to each weight product. Since both weight matrices are symmetrical, these four terms are equal (eq 16c, see Methods). As consequence, this term is multiplied by four in eq. 16c. (see Methods)

The term D_3 is defined as,

$$D_3 = (T_3)(P_3) \text{ , where} \tag{4}$$

$$T_3 = \left(\sum_u^n\left((\sum_v^n \gamma_{uv})^2 - \sum_v^{c_u} \gamma_{uv}^2\right)\right),$$

$$P_3 = \left(\sum_i^n \left(\left(\sum_j^n (p(i) - p(j))^2\right)^2 - \sum_j^n (p(i) - p(j))^4\right)\right)$$

If the first endpoints of both edges represent the same vertex u then,

$$D_3 = (T_3)(P_3), \tag{5}$$
$$T_3 = \sum_u^n \sum_{v,h,v\neq h}^n (\gamma_{uv}\gamma_{uh})$$
$$P_3 = \left(\sum_i^n \sum_{j,h,i\neq h}^n (p(i) - p(j)))^2 \, (p(i) - p(h))^2\right)$$

if the second endpoint of the first edge and the first endpoint of second edge represent the same vertex u then,

$$D_{3a} = (T_{3a})(P_{3a}), \tag{6}$$
$$T_{3a} = \sum_u^n \sum_{v,h,v\neq h}^n (\gamma_{vu}\gamma_{uh})$$
$$P_{3a} = \left(\sum_i^n \sum_{j,h,j\neq h}^n (p(j) - p(i))^2 \, (p(i) - p(h))^2\right)$$

if the second endpoint of the second edge and the first endpoint of first edge represent the same vertex u then,

$$D_{3b} = (T_{3b})(P_{3b}) \tag{7}$$
$$T_{3b} = \sum_u^n \sum_{v,h,v\neq h}^n (\gamma_{uv}\gamma_{hu})$$
$$P_{3b} = \left(\sum_i^n \sum_{j,h,j\neq h}^n (p(i) - p(j))^2 \, (p(h) - p(i))^2\right)$$

if the second endpoint of the second edge and the second endpoint of first edge represent the same vertex u then,

$$D_{3c} = (T_{3c})(P_{3c}) \tag{8}$$
$$T_{3c} = \sum_u^n \sum_{v,h,v\neq h}^n (\gamma_{vu}\gamma_{hu})$$
$$P_{3c} = \left(\sum_i^n \sum_{j,h,j\neq h}^n (p(j) - p(i))^2 \, (p(h) - p(i))^2\right)$$

Where, $v, j \neq h, \ \ \forall \ 1 \leq i, j, u, v, h \leq n \quad i, j, u, v, h, n \in N,$

Since both weight matrices are symmetrical, $T_3 = T_{3a} = T_{3b} = T_{3c}$ and $P_3 = P_{3a} = P_{3b} = P_{3c}$. Hence, $D_3 = D_{3a} = D_{3b} = D_{3c}$. That's why in our case, $D_3 + D_{3b} + D_{3c} + D_{3d} = 4D_3$. There are 4(n-3)! assignments for each three amino acids to each three vertices connected by two adjacent edges.

Appendix VIII Neighborhood structure of natural genetic codes

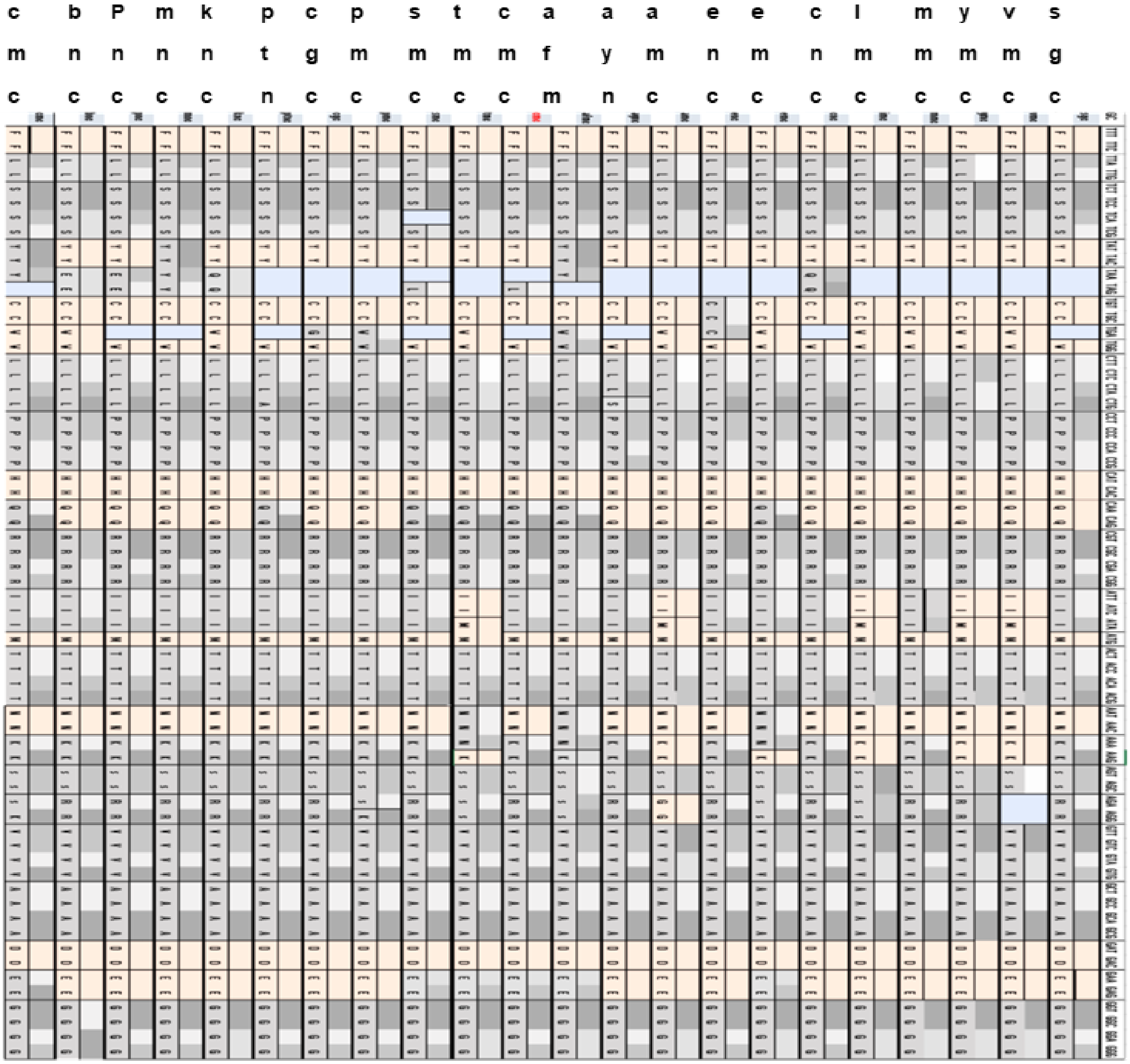

c b P m k p c p s t c a a a e e c l m y v s
m n n n n t g m m m m f y m n m n m m m m g
c c c c c n c c c c c m n c c c c c c c c c

Figure VIII.1 Neighborhood structures of 22 natural genetic codes. Each pair of Columns corresponds to a genetic code, each row corresponds to a codon. The left column of each pair shows the amino acid encoded by the codon of the corresponding row, each cell color (different shades of grey) of the right column indicates that the corresponding codon belongs to a homogeneous sub-block or is a singleton. Light orange: Indicates the presence of a homogeneous blocks, Light blue: Indicates the presence of stop codons. Mitochondria and plastid genetic codes, sgc: standard code, vmc: The Vertebrate Mitochondrial Code, ymc: The Yeast Mitochondrial Code, mmc: The Mold, Protozoan, and Coelenterate Mitochondrial Code, ivmc: The Invertebrate Mitochondrial Code, cnc: The Ciliate, Dasycladacean and Hexamita Nuclear Code, emc: The Echinoderm and Flatworm Mitochondrial Code, enc: The Euplotid Nuclear Code, amc: The Ascidian Mitochondrial Code, aync: The Alternative yeast nuclear code, afmc: The Alternative Flatworm Mitochondrial Code, cmc: Chlorophycean Mitochondrial Code, tmc: Trematode Mitochondrial Code, smc: Scenedesmus obliquus Mitochondrial Code, pmc: Pterobranchia Mitochondrial Code, cgc: Candidate Division SR1 and Gracilibacteria Code, ptnc: Pachysolen tannophilus Nuclear Code, knc: Karyorelict Nuclear Code, mnc: Mesodinium Nuclear Code, pnc: Peritrich Nuclear Code, bnc: Blastocrithidia Nuclear Code, cmtc: Cephalodiscidae Mitochondrial UAA-Tyr Code, tamc: Traustochytrium mitochondrial code.

Appendix IX Amino acid properties, genomes and empirical estimates of robustness

Table IX.1 List of amino acid properties scales. Pnum: Numerical identifiers of the amino acid property scale, Loc, L: Local amino acid property: Amino acid property scale defined from specific proteins, protein regions/domains and sites. Glo, G: Global amino acid property: Amino acid property scale defined for the amino acid regardless the biological context. aa: amino acid, B: beta, A: Alpha. Asa: accessible Surface area.

Pnum	Loc/Glo	Amino acid property scale names	References
1	L	**Scores for adenine-protein interaction**	Mandel-Gutfreund and Margalit, Nucleic Acids Research, 1998, Vol. 26, No. 10 2306–2312
2	L	**aa Helix Propensities in B/A proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
3	L	**aa Alpha Helical Propensities**	Muñoz V, Serrano L. J Mol Biol. 1995 Jan 20;245(3):275-96.
4	L	**aa Helix Propensities in alfa proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
5	L	**aa B-sheet Propensities in B+A proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
6	L	**aa B-sheet Propensities in B proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
7	L	**aa B-sheet Propensities in B/A proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
8	L	**aa Alpha Helical Propensities**	Blaber M1, Zhang XJ et al. J Mol Biol. 1994 Jan 14;235(2):600-24.
9	L	**Buried Alpha Helix solvent accessibilities**	Michael J. Thompson and Richard A. Goldstein PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
10	L	**aa Coil Propensities in B+A proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
11	L	**aa Coil Propensities in B/A proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
12	L	**Coil Propensities in B-sheet proteins jiang 1997**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
13	L	**Helix Propensities in pept. whithout Helix-stabilizing schain interactions.**	A. Chakrabartty, T. Kortemme, and R. L. Baldwin, Protein Science (1994), 3:843-852.
14	L	**Normalized frequency of Alpha-Helix**	Chou PY, Fasman GD (1974) Biochemistry. 13 (2): 222–245.
15	L	**Coil Accessible Surface area**	Fan Jiang Protein Engineering vol. 16 no. 9 pp. 651-657, 2003
16	L	**aa Alpha Helix Propensities**	Deléage G1, Roux B. Et al. Protein Eng. 1987 Aug-Sep;1(4):289-94.
17	L	**aa Rotational Potentials in Alpha-Helix**	Bahar,I, Kaplan Mand. Jernigan R.L PROTEINS: Structure, Function, and Genetics 29:292–308 (1997)
18	L	**Exposed Alpha Helix solvent accessibilities**	Michael J. Thompson' and Richard A. Goldstein PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
19	L	**Helix Propensities in B+A proteins**	Bo Jiang Tao Guo Lei-Wei Peng Zhi-Rong Sun, Peptide Science 45(1): 35-49, December 1997
20	L	**conformational preference parameter for membrane-Buried helices**	Rao M.J.K., Argos P. Biochim. Biophys. Acta 869:197-214(1986).
21	L	**Free energy for a-Helical conformation**	Victor Munoz and Luis Serrano, PROTEINS: Structure, Function, and Genetics 20:301-311 (1994)
22	L	**Thermodynamic scale for the aa Helix-forming tendencies**	O'Neil KT, DeGrado WF. Science. 1990 Nov 2;250(4981):646-51.
23	L	**aa Rotational Potentials in Alpha-Helix**	I. Bahar,M. Kaplan,and R.L. Jernigan PROTEINS: Structure, Function, and Genetics 29:292–308 (1997)
24	L	**aa Rotational Potentials in Alpha-Helix**	I. Bahar,M. Kaplan,and R.L. Jernigan PROTEINS: Structure, Function, and Genetics 29:292–308 (1997)
25	L	**aa Rotational Potentials in Alpha-Helix**	I. Bahar,M. Kaplan,and R.L. Jernigan PROTEINS: Structure, Function, and Genetics 29:292–308 (1997)
26	L	**Helicity in water, 0222nm Circular dicroism (CD) spectra is used as a measure of Helicity (model peptides).**	Liu LP, Deber CM. Biopolymers. 1998;47(1):41-62.
27	L	**Helicity in n-butanol, 0222nm CD spectra is used as a measure of Helicity (model peptides).**	Liu LP, Deber CM. Biopolymers. 1998;47(1):41-62.
28	L	**statistical transmembrane Alpha-Helix Propensities in single-spanning proteins**	Liu LP, Deber CM. Biopolymers. 1998;47(1):41-62.
29	L	**Free energy for a Helical region based on psi-phi matrices**	Victor Munoz and Luis Serrano PROTEINS: Structure, Function, and Genetics 20:301-311 (1994)
30	L	**Helix-coil stability constants**	Altmann KH1, Wójcik J, Vásquez M, Scheraga HA. Biopolymers. 1990;30(1-2):107-20.
31	L	**Helix-forming tendency in thermostable proteins**	Gregory L. Warren Gregory A. Petsko Protein Engineering, Design and Selection, Volume 8, Issue 9, September 1995, Pages 905–913
32	L	**Helix propensity scale**	Jianxin Yang, Erik J. Spek, Youxiang Gong, et al Protein Science (1997). 6:1264-1272.
33	L	**Helix propensity scale**	RichardsonJ.S. and Richardson,D.C. (1988) Science, 240, 1648-1652.

34	L	**Helix propagation Propensities**	Rohl CA, Chakrabartty A, Baldwin RL. Protein Sci. 1996 Dec;5(12):2623-37.
35	L	**transmembrane Alpha-Helix propensity**	Gromiha MM. Protein Eng. 1999 Jul;12(7):557-61
36	L	**Helix propensity scale**	Williams R.W., Chang A., Juretic D. and Loughran.S; Biochim Biophys Acta. 1987 Nov 26;916(2):200-4.
37	L	**Asa in coil structures**	Fan Jiang Protein Engineering vol. 16 no. 9 pp. 651-657, 2003
38	L	**Total Asa in folded beta s structures**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
39	L	**Hydrophilic Asa in folded beta s structures**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
40	L	**Hydrophobic Asa in folded beta s structures**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
41	L	**Total Asa in folded coil structures**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
42	L	**Hydrophilic Asa in folded coil structures**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
43	L	**Hydrophobic Asa in folded coil structures**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
44	G	**Hydrophilic Asa in folded proteins**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
45	G	**Hydrophobic Asa in folded proteins**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
46	G	**Total Asa in folded proteins**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
47	L	**Asa in Alpha Helix**	Fan Jiang Protein Engineering vol. 16 no. 9 pp. 651-657, 2003
48	G	**accessible Surface area**	Uttamkumar Samanta Ranjit P. Bahadur Pinak Chakrabarti Protein Engineering vol.15 no.8 pp.659–667, 2002
49	L	**accessible Surface area in Beta strands**	Fan Jiang Protein Engineering vol. 16 no. 9 pp. 651-657, 2003
50	G	**Total accessible Surface area**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
51	G	**Hydrophilic accessible Surface area**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
52	G	**Hydrophobic accessible Surface area**	Laurence Lins, Annick Thomas, AND RoberT Brasseur; Protein Sci. 2003 Jul; 12(7): 1406–1417.
53	L	**Optimized beta-structure-coil equilibrium constant**	Bull. Inst. Chem. Res., Kyoto Univ. 63, 82-94 (1985)
54	L	**Buried Beta sheet solvent accessibility**	Michael J. Thompson and Richard A. Goldstein PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
55	L	**B sheet propensity**	Deléage G, Roux B. Protein Eng. 1987 Aug-Sep;1(4):289-94.
56	L	**Propensity of Amino acid residues to occur in Isolated E-strand**	Narayanan Eswar, C.Ramakrishnan and N.Srinivasan Protein Engineering vol. 16 no. 5 pp. 331±339, 2003
57	L	**Propensity of Amino acid residues to occur in Edge β-strand**	Narayanan Eswar, C.Ramakrishnan and N.Srinivasan Protein Engineering vol. 16 no. 5 pp. 331±339, 2003
58	L	**Exposed Beta sheet solvent accessibility**	Michael J. Thompson' and Richard A. Goldstein PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
59	L	**Propensity of Amino acid residues to occur in Inner β-strand**	Narayanan Eswar, C.Ramakrishnan and N.Srinivasan Protein Engineering vol. 16 no. 5 pp. 331±339, 2003
60	L	**B sheet propensity**	Rune Linding*, Robert B. Russell Nucleic Acids Research, 2003, Vol. 31, No. 13 3701–3708
61	L	**B sheet propensity**	Minor DL Jr, Kim PS. Nature. 1994 Feb 17;367(6464):660-3.
62	L	**B sheet propensity**	Minor DL Jr, Kim PS. Nature. 1994 Feb 17;367(6464):660-3.
63	L	**Free energy for B-strand region**	Victor Munoz and Luis Serrano PROTEINS: Structure, Function, and Genetics 20:301-311 (1994)
64	L	**Free energy for B-strand region**	Victor Munoz and Luis Serrano PROTEINS: Structure, Function, and Genetics 20:301-311 (1994)
65	L	**Free energy for B-strand conformation**	Victor Munoz and Luis Serrano PROTEINS: Structure, Function, and Genetics 20:301-311 (1994)
66	L	**B sheet conformational parameters**	Chou PY, Fasman GD. Annu Rev Biochem. 1978;47:251-76.
67	L	**B strand preference inside local bending**	Carola Daffner, Gareth Chelvanayagam and Patrick Argos Protein Science (1994), 32376-882.
68	L	**B strand preference next to local bending**	Carola Daffner, Gareth Chelvanayagam and Patrick Argos Protein Science (1994), 32376-882.
69	L	**B strand preference**	Carola Daffner, Gareth Chelvanayagam and Patrick Argos Protein Science (1994), 32376-882.
70	L	**B sheet propensity**	Zimmerman JM, Eliezer N, Simha R. J Theor Biol. 1968 Nov;21(2):170-201.

71	L	**Cytosine_protein interaction**	Mandel-Gutfreund and Margalit*, Nucleic Acids Research, 1998, Vol. 26, No. 10 2306–2312
72	L	**Lipid accessibilities within the transmembrane Helix 1**	Larisa Adamian, Vikas Nanda, William F. DeGrado and Jie Liang; PROTEINS 59:496–509 (2005)
73	L	**Buried coil solvent accessibility**	Michael J. Thompson' and Richard A. Goldstein; PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
74	L	**Exposed coil solvent accessibility**	Michael J. Thompson' and Richard A. Goldstein; PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
75	L	**Random coil propensity**	Deléage G, Roux B. Protein Eng. 1987 Aug-Sep;1(4):289-94.
76	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
77	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
78	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
79	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
80	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
81	L	**Amino Acid Propensities in Polyproline II Helices L3**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
82	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
83	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
84	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
85	L	**Amino Acid Propensities in Polyproline II Helices L+3**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
86	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
87	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
88	L	**Position-Dependent Propensities for Polyproline II Helices**	Cubellis, M. V., Caillez, F. , Blundell, T. L. and Lovell, S. C. (2005); PROTEINS: Structure, Function, and Genetics 58: 880-892.
89	L	**Difference between Side-chain conformational entropies of Amino acids in the a-Helical and the coil states**	Avbelj F, Fele L. J Mol Biol. 1998 Jun 12;279(3):665-84.
90	G	**Conformational entropy differences between free and Buried states of aa Side-chains**	Abagyan R, Totrov M. J Mol Biol. 1994 235:235:983-1002.
91	G	**absolute entropy**	ANDREW J. DOIG AND MICHAEL J.E. STERNBERG; Protein Science (1995), 4:2247-2251.
92	L	**The Side-chain conformational entropies of Amino acids in the coil states**	Avbelj F, Fele L.; J Mol Biol. 1998 Jun 12;279(3):665-84.
93	L	**Backbone Entropy of All Residues in the Coil Library**	Jha AK1, Colubri A, Zaman MH, Koide S, Sosnick TR, Freed KF.; Biochemistry. 2005 Jul 19;44(28):9691-702.
94	G	**Mean changes in Side-chain conformational entropy**	Andrew J. Doig AND Michael J.E. Sternberg; Protein Science (1995), 4:2247-2251.
95	L	**The Side-chain conformational entropies of Amino acids in the a-Helical structures**	Avbelj F, Fele L.; J Mol Biol. 1998 Jun 12;279(3):665-84.
96	G	**Side-chain conformational entropy**	Koehl P, Delarue M.;J Mol Biol. 1994 Jun 3;239(2):249-75.
97	G	**Side-chain conformational entropy**	Kon Ho Lee, Dong Xie, Ernesto Freire, and L. Mario Amzel; PROTEINS: Structure, Function, and Genetics 20:68-84 (1994)
98	G	**Side-chain conformational entropy**	Pickett SD, Sternberg MJ.;J Mol Biol. 1993 Jun 5;231(3):825-39.
99	G	**Sequence-dependence of backbone entropy OPLS-AA-01**	Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR; J Mol Biol. 2003 Aug 15;331(3):693-711.
100	G	**Sequence-dependence of backbone entropy AMBER 94**	Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR; J Mol Biol. 2003 Aug 15;331(3):693-711.
101	G	**Sequence-dependence of backbone entropy G-S-94**	Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR; J Mol Biol. 2003 Aug 15;331(3):693-711.
102	G	**Sequence-dependence of backbone entropy OPLS-UA**	Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR; J Mol Biol. 2003 Aug 15;331(3):693-711.
103	L	**Scores for guanine-protein interaction**	Mandel-Gutfreund and Margalit, Nucleic Acids Research, 1998, Vol. 26, No. 10 2306–2312
104	G	**High thermodynamic stability**	James O. Wrabl, Scott A. Larson, and Vincent J. Hilser; Protein Sci. 2001 May; 10(5): 1032–1045.
105	L	**Alpha Helix 1 (p15) Propensities**	Shortle D. Protein Sci. 2002 Jan;11(1):18-26.

106	L	**Alpha Helix 2 (p15) Propensities**	Shortle D. Protein Sci. 2002 Jan;11(1):18-26.
107	L	**Alpha Helix 3 (p15) Propensities**	Shortle D. Protein Sci. 2002 Jan;11(1):18-26.
108	G	**Hydrophobicity**	Abraham, D. J., and Leo, A. J. (1987) Proteins: Struct., Funct., Genet. 2, 130-152.
109	G	**Lipid accessibilities within the transmembrane Helix 2**	Larisa Adamian,Vikas Nanda,William F. DeGrado, and Jie Liang;PROTEINS: Structure, Function, and Bioinformatics 59:496–509 (2005)
110	L	**The free energy difference due to the main-chain conformational entropy between the a-Helical and the coil states**	Avbelj F, Fele L; J Mol Biol. 1998 Jun 12;279(3):665-84.
111	G	**Surface propensity scale**	Thijs Beuming and Harel Weinstein; Bioinformatics Vol. 20 no. 12 2004, pages 1822–1835
112	G	**Metabolic costs of Amino acid biosynthesis, numbers of available hydrogen atoms in NADH, NADPH, and FADH2**	Hiroshi Akashi and Takashi Gojobori; Proc Natl Acad Sci U S A. 2002 Mar 19; 99(6): 3695–3700.
113	G	**scaled Side Chain Hydrophobicity**	Shaun D. Black and Diane R. Mouldt; ANALYTICAL BIOCHEMISTRY 193, 72-82 (1991)
114	G	**Hydrophobicity, free energies of transfer of aa from the solution to the Surface**	Bull HB, Breese K.; Arch Biochem Biophys. 1974 Apr 2;161(2):665-70.
115	G	**Hydrophobicity, for detecting amphipathic structures in proteins**	James L. Cornette, Kemp B. Cease, Hanah Margalitt; J. Mol. Biol. (1987) 195, 659-685
116	G	**Hydrophobicity, (retention times on HPLC, ph 3)**	Cowan R, Whittaker RG.;Pept Res. 1990 Mar-Apr;3(2):75-80.
117	G	**Hydrophobicity, (retention times on HPLC, ph 7.5)**	Cowan R, Whittaker RG.;Pept Res. 1990 Mar-Apr;3(2):75-80.
118	G	**Consensus normalized Hydrophobicity scale**	Eisenberg, D. et al.The Hydrophobic moment detects periodicity in protein Hydrophobicity Proc. Natl. Acad. Sci. USA 81, 140-144 (1984)
119	G	**Hydrophobic parameter pi**	Fauchere J.-L., Pliska V.E. Eur. J. Med. Chem. 18:369-375(1983).
120	G	**Hydrophobicity,Partition energy**	H R Guy; Biophys J. 1985 Jan; 47(1): 61–70.
121	G	**Hydrophilicity**	THOMAS P. HOPP AND KENNETH R. WOODS;
122	G	**pure Hydrophobicity scale**	P. Asndrew Karplus; Protein Science(1997). 61302-1307.
123	G	**Hydrophobicity**	W.R. Krigbum and Akira Komoriya; Biochimica et Biophysica Acta, 576 (1979) 204--228
124	G	**Hydrophobicity**	G. Rose, A. Geselowitz, G. Lesser et al. Hydrophobicity of Amino Acid Residues in Globular Proteins, Science 229(1985)834-838.
125	G	**Hydropathy index**	Jack Kyte and Russell F. Doolitle; J. Mol. Biol. (1982) 157, 105-132
126	G	**Hydrophobicity, free energy of transfer of aa from cyclohexylpyrrolidone to water**	Erlinda Q. Lawson, Albert J. Sadler et al. JBiolChem Vol. 259, No. 5, Issue of March 10- p p . 2910-2912, 1984
127	G	**Hydrophobicity, free energy of transfer of aa from etanol to water**	Erlinda Q. Lawson, Albert J. Sadler et al. JBiolChem Vol. 259, No. 5, March 10. 2910-2912, 1984
128	G	**Hydrophobic parameter**	Michael Levitt; J. Mol. Biol. (1976) 104, 59-107
129	G	**Hydrophobicity, ph 2.1 (retention times on HPLC)**	James L. Meek ; Proc. Natl. Acad. Sci. USA Vol. 77, No. 3, pp. 1632-1636, March 1980
130	G	**Hydrophobicity, ph 7.4 (retention times on HPLC)**	James L. Meek; Proc. Natl. Acad. Sci. USA Vol. 77, No. 3, pp. 1632-1636, March 1980
131	G	**Hydrophobicity, partition coefficients (interior and Surface aa)**	Susan Millerl, Joel Janin’, Arthur M. Lesk, and Cyrus Chothia; J. Mol. Biol. (1987) l%, 641-656
132	G	**Hydrophobicity, contact energies derived from protein structure**	Sanzo Miyazawa and Robert L. Jernigan; Macromolecules 1985, 18, 534-552
133	G	**Hydrophobicity, Total free energy of hydration**	Tatsuo Ooi, Motohisa Oobatake et al.; Proc. Natl. Acad. Sci. USA Vol. 84, pp. 3086-3090, May 1987
134	L	**Helix propensity**	C.Nick PaceJ.Martin Scholtz; Biophysical Journal Volume 75, Issue 1, July 1998, Pages 422-427
135	G	**Hydrophilicity, HPLC**	Parker JM, Guo D, Hodges RS.; Biochemistry. 1986 Sep 23;25(19):5425-32.
136	G	**Average ratios between residue occurrences at the intra- and extracellular Sides of the membrane**	Persson B1, Argos P.; Protein Sci. 1996 Feb;5(2):363-71.
137	G	**Hydrophobicity,Accessibility reduction ratio**	P.K.Ponnuswamy, M. Prabhakaran and P. Manavalan; Biochimica et Biophysica Acta, 623 (1980) 301--316
138	G	**Hydrophilicity of polar Amino acid Side-chains is markedly reduced by flanking peptide bonds**	Roseman, M.A. J. Mol. Biol. 200, 513-522 (1988)
139	L	**distribution of Amino acid residues in transmembrane a-Helix bundles**	Fadel A. Samatey, Chuanbo Xut, AND Jean-Luc Potot; Proc. Natl. Acad. Sci. USA Vol. 92, pp. 4577-4581, May 1995
140	G	**Side-chain contribution to protein stability, based on the stability change of mutant proteins**	Takano, K., Yutani, K. Protein Eng. 14, 525-528 (2001)

141	L	**Transmembrane central aa frequence**	Yitzhak Pilpel, Nir Ben-Tal and Doron Lancet; J. Mol. Biol. (1999) 294, 921-935
142	L	**Transmembrane extracelular aa frequence**	Yitzhak Pilpel, Nir Ben-Tal and Doron Lancet; J. Mol. Biol. (1999) 294, 921-935
143	L	**Transmembrane intracelular aa frequence**	Yitzhak Pilpel, Nir Ben-Tal and Doron Lancet; J. Mol. Biol. (1999) 294, 921-935
144	L	**Transmembrane both termini aa frequence**	Yitzhak Pilpel, Nir Ben-Tal and Doron Lancet; J. Mol. Biol. (1999) 294, 921-935
145	L	**Transmembrane Total aa frequence**	Yitzhak Pilpel, Nir Ben-Tal and Doron Lancet; J. Mol. Biol. (1999) 294, 921-935
146	G	**Transfer free energy to lipophilic phase**	von Heijne, G. (1981) Eur. J. Biochem. 116,419-422.
147	G	**Hidrophobicity (HPLC)**	K J Wilson, A Honegger, R P Stötzel, and G J Hughes ; Biochem J. 1981 Oct 1; 199(1): 31–41.
148	G	**Hydrophobicity, Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water**	Wimley WC, White SH.; Nat Struct Biol. 1996 Oct;3(10):842-8.
149	G	**Polar requirement**	Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC (1966). Cold Spring Harbour Symp Quant Biol 31:723–736
150	G	**Hydrophobicity, Hydration potential**	Wolfenden R, Andersson L, Cullis PM, Southgate CC.;Biochemistry. 1981 Feb 17;20(4):849-55.
151	G	**Polarity**	Zimmerman, J.M., Eliezer, N. and Simha, R.;J. Theor. Biol. 21, 170-201 (1968)
152	L	**Propensities of Amino acids for all protein-protein interfaces**	Ozlem Keskin, Chung-jung TsaI, Haim Wolfson AND Ruth ; Protein Sci. 2004 Apr;13(4):1043-55.
153	L	**Propensities of Amino acids for type I protein-protein interfaces**	Ozlem Keskin, Chung-jung TsaI, Haim Wolfson AND Ruth ; Protein Sci. 2004 Apr;13(4):1043-55.
154	L	**Propensities of Amino acids for type II protein-protein interfaces**	Ozlem Keskin, Chung-jung TsaI, Haim Wolfson AND Ruth ; Protein Sci. 2004 Apr;13(4):1043-55.
155	L	**Propensities of Amino acids for type III protein-protein interfaces**	Ozlem Keskin, Chung-jung TsaI, Haim Wolfson AND Ruth ; Protein Sci. 2004 Apr;13(4):1043-55.
156	L	**aa isoelectric point**	Zimmerman, J.M., Eliezer, N. and Simha, R.; J. Theor. Biol. 21, 170-201 (1968)
157	L	**L1 (p15) Propensities**	Shortle D.; Protein Sci. 2002 Jan;11(1):18-26.
158	L	**L2 (p15) Propensities**	Shortle D.; Protein Sci. 2002 Jan;11(1):18-26.
159	L	**L3 (p15) Propensities**	Shortle D.; Protein Sci. 2002 Jan;11(1):18-26.
160	G	**Average 11-20 Long-range contacts per residue**	M. Michael Gromihaa,U, S. Selvaraj; Biophysical Chemistry 77, 1999. 49-68
161	G	**Average 21-30 Long-range contacts per residue**	M. Michael Gromihaa,U, S. Selvaraj; Biophysical Chemistry 77, 1999. 49-68
162	G	**Average 31-40 Long-range contacts per residue**	M. Michael Gromihaa,U, S. Selvaraj; Biophysical Chemistry 77, 1999. 49-68
163	G	**Average 41-50 Long-range contacts per residue**	M. Michael Gromihaa,U, S. Selvaraj; Biophysical Chemistry 77, 1999. 49-68
164	G	**Average p50 Long-range contacts per residue**	M. Michael Gromihaa,U, S. Selvaraj; Biophysical Chemistry 77, 1999. 49-68
165	G	**Average long-range contacts per residue**	M. Michael Gromihaa,U, S. Selvaraj; Biophysical Chemistry 77, 1999. 49-68
166	L	**LH (p15) Propensities**	Shortle D.; Protein Sci. 2002 Jan;11(1):18-26.
167	L	**Medium Length linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
168	L	**Length 1 linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
169	L	**Length 2 linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
170	L	**Length 3 linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
171	L	**Total linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
172	L	**Helical linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
173	L	**non-Helical linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
174	L	**small Length linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003
175	L	**Long Length linker propensity**	Richard A.George and Jaap Heringa; Protein Engineering vol.15 no.11 pp.871–879, 2003

176	L	**Propensity of Amino acid residues to occur loops**	Narayanan Eswar, C.Ramakrishnan and N.Srinivasan; Protein Engineering vol. 16 no. 5 pp. 331-339, 2003
177	L	**Low thermodynamic stability**	James O. Wrabl, Scott A. Larson,and Vincent J. Hilser; Protein Sci. 2001 May; 10(5): 1032–1045.
178	L	**m1 (p15) Propensities**	Shortle D. Protein Sci. 2002 Jan;11(1):18-26.
179	L	**m2 (p15) Propensities**	Shortle D. Protein Sci. 2002 Jan;11(1):18-26.
180	L	**m3(p15) Propensities**	Shortle D. Protein Sci. 2002 Jan;11(1):18-26.
181	G	**The free energy difference due to the main-chain conformational entropy between the beta strand and the coil states**	Avbelj F, Fele L.; J Mol Biol. 1998 Jun 12;279(3):665-84.
182	G	**Average Medium-range contacts in globular proteins**	Gromiha MM1, Selvaraj S. Prog Biophys Mol Biol. 2004 Oct;86(2):235-77.
183	G	**Two rigid neighbors, Mean Location parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G; Protein Sci. 2003 May;12(5):1060-72.
184	G	**Two flexible neighbors, Mean Location parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G; Protein Sci. 2003 May;12(5):1060-72.
185	G	**One rigid and one flexible neighbor, Mean Location parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G; Protein Sci. 2003 May;12(5):1060-72.
186	G	**Mean Location parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G; Protein Sci. 2003 May;12(5):1060-72.
187	G	**Amino acid melting point**	Fasman, G.D., ed.; "Handbook of Biochemistry and Molecular Biology", 3rd ed., Volume 1, CRC Press, Cleveland (1976)
188	G	**Medium thermodynamic stability**	James O. Wrabl, Scott A. Larson and Vincent J. Hilser; Protein Sci. 2001 May; 10(5): 1032–1045.
189	G	**Amino acid molecular weight**	Fasman, G.D., ed.; "Handbook of Biochemistry and Molecular Biology", 3rd ed., Volume 1, CRC Press, Cleveland (1976)
190	L	**Free energy of residues at the N terminal position 1 in α-helices**	Claire L Wilson, Simon J. Hubbard and Andrew J. Doig ; Protein engineering vol 7 no 7 pp545-554, 2002
191	L	**Free energy of residues at the N terminal position 2 in α-helices**	Claire L Wilson, Simon J. Hubbard and Andrew J. Doig ; Protein engineering vol 7 no 7 pp545-554, 2002
192	L	**Free energy of residues at the N-cap position in α-helices**	Claire L Wilson, Simon J. Hubbard and Andrew J. Doig ; Protein engineering vol 7 no 7 pp545-554, 2002
193	L	**Average partner number**	Uttamkumar Samanta, Ranjit P.Bahadur and Pinak Chakrabarti;Protein Engineering vol.15 no.8 pp.659–667, 2002
194	L	**Amino acid optical rotation**	Fasman, G.D., ed.; "Handbook of Biochemistry and Molecular Biology", 3rd ed., Volume 1, CRC Press, Cleveland (1976)
195	L	**other (p15) Propensities**	Shortle D1.; Protein Sci. 2002 Jan;11(1):18-26.
196	L	**Metabolic costs of Amino acid biosynthesis,numbers of High-energy phosphate bonds in ATP and GTP molecules**	Hiroshi Akashi and Takashi Gojobori; Proc Natl Acad Sci U S A. 2002 Mar 19; 99(6): 3695–3700.
197	L	**phi (p15) Propensities**	Shortle D1.; Protein Sci. 2002 Jan;11(1):18-26.
198	G	**Amino acid PK-C**	Fasman, G.D., ed.; "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
199	G	**Amino acid PK-N**	Fasman, G.D., ed.; "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
200	L	**Propensity of Amino acid residues to occur polyproline type Helix**	Narayanan Eswar, C.Ramakrishnan and N.Srinivasan; Protein Engineering vol. 16 no. 5 pp. 331-339, 2003
201	L	**Amino acid Propensities in polyproline II helices**	Stapley BJ, Creamer TP.; Protein Sci. 1999 Mar;8(3):587-95.
202	L	**r1 (p15) Propensities**	Shortle D1.; Protein Sci. 2002 Jan;11(1):18-26.
203	L	**r2 (p15) Propensities**	Shortle D1.; Protein Sci. 2002 Jan;11(1):18-26.
204	L	**r3 (p15) Propensities**	Shortle D1.; Protein Sci. 2002 Jan;11(1):18-26.
205	G	**Sidechain Radii**	Adrian P. Cootes,Paul M.G. Curmi,2 Ross Cunningham; PROTEINS: Structure, Function, and Genetics 32:175–189 (1998)
206	L	**cis-trans prolyl isomerisation Rate constants**	Reimer U, Scherer G, Drewello M, Kruber S, Schutkowski M, Fischer G.; J Mol Biol. 1998 Jun 5;279(2):449-60.
207	L	**trans-cis prolyl isomerisation Rate constants**	Reimer U, Scherer G, Drewello M, Kruber S, Schutkowski M, Fischer G.; J Mol Biol. 1998 Jun 5;279(2):449-60.
208	G	**Refractivity**	McMeekin, T.L., Groves, M.L. and Hipp, N.J.; "Amino Acids and Serum Proteins" , American Chemical Society, Washington, p. 54 (1964)
209	G	**Two flexible neighbors, Mean Scale parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G.;Protein Sci. 2003 May;12(5):1060-72.
210	G	**One rigid and one flexible neighbor, Mean Scale parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G.;Protein Sci. 2003 May;12(5):1060-72.

211	G	**Two rigid neighbors, Mean scale parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G.;Protein Sci. 2003 May;12(5):1060-72.
212	G	**Total Mean scale parameters (fit of the B-factors to a Gumbel distribution)**	Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G.;Protein Sci. 2003 May;12(5):1060-72.
213	G	**Attenuation of the non local electrostatic energies**	Avbelj F, Fele L.; J Mol Biol. 1998 Jun 12;279(3):665-84.
214	L	**Scores for thymine-protein interaction**	Mandel-Gutfreund and Margalit, Nucleic Acids Research, 1998, Vol. 26, No. 10 2306–2312
215	G	**Total metabolic costs of Amino acid biosynthesis**	Hiroshi Akashi and Takashi Gojobori; Proc Natl Acad Sci U S A. 2002 Mar 19; 99(6): 3695–3700.
216	L	**Buried Turn propensity**	Michael J. Thompson' and Richard A. Goldstein; PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
217	L	**Exposed Turn propensity**	Michael J. Thompson' and Richard A. Goldstein; PROTEINS: Structure, Function, and Genetics 25:38-47 (1996)
218	L	**Averaged Turn Propensities in transmembrane Helices**	Monné M, Nilsson I, Elofsson A, von Heijne G.; J Mol Biol. 1999 Nov 5;293(4):807-14.
219	L	**Normalized Turn potential in transmembrane Helices**	Monné M, Nilsson I, Elofsson A, von Heijne G.; J Mol Biol. 1999 Nov 5;293(4):807-14.
220	L	**Turn propensity**	Deléage G, Roux B.; Protein Eng. 1987 Aug-Sep;1(4):289-94.
221	L	**Turn propensity**	CAROLA DAFFNER, GARETH CHELVANAYAGAM,' AND PATRICK ARGOS; Protein Science (1994), 32376-882.
222	L	**Potentials for position i of the type VIII Turn**	Harri Santa, Markku Ylisirnio, Tommi Hassinen; Protein Engineering vol.15 no.8 pp.651–657, 2002
223	L	**Potentials for position i+1 of the type VIII Turn**	Harri Santa, Markku Ylisirnio, Tommi Hassinen; Protein Engineering vol.15 no.8 pp.651–657, 2002
224	L	**Potentials for position i+2 of the type VIII Turn**	Harri Santa, Markku Ylisirnio, Tommi Hassinen; Protein Engineering vol.15 no.8 pp.651–657, 2002
225	L	**Potentials for position i+3 of the type VIII Turn**	Harri Santa, Markku Ylisirnio, Tommi Hassinen; Protein Engineering vol.15 no.8 pp.651–657, 2002
226	G	**Residue volumes**	Jerry Tsai, Robin Taylor, Cyrus Chothia and Mark Gerstein; J. Mol. Biol. (1999) 290, 253-266
227	G	**Residue volumes**	Jerry Tsai, Robin Taylor, Cyrus Chothia and Mark Gerstein; J. Mol. Biol. (1999) 290, 253-266
228	G	**Residue volumes**	Jerry Tsai, Robin Taylor, Cyrus Chothia and Mark Gerstein; J. Mol. Biol. (1999) 290, 253-266
229	L	**Enthalpies of Gly-X-Hyp with aa in position x**	Anton V. Persikov, John A. M. Ramshaw, Alan Kirkpatrick,and Barbara Brodsky; Biochemistry 2000, 39, 14960-14967
230	L	**Occurrences of aa in position x of Gly-X-Hyp**	Anton V. Persikov, John A. M. Ramshaw, Alan Kirkpatrick,and Barbara Brodsky; Biochemistry 2000, 39, 14960-14967
231	L	**Enthalpies of Gly-Pro-Y with aa in position y**	Anton V. Persikov, John A. M. Ramshaw, Alan Kirkpatrick,and Barbara Brodsky; Biochemistry 2000, 39, 14960-14967
232	L	**Occurrences of aa in position y of Gly-Pro-Y**	Anton V. Persikov, John A. M. Ramshaw, Alan Kirkpatrick,and Barbara Brodsky; Biochemistry 2000, 39, 14960-14967
233	G	**Principal property scale Z1 (Hydrophobicity) (QSAM methods: PLS and PCA)**	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. J Med Chem. 1998 Jul 2;41(14):2481-91.
234	G	**Principal property scale Z2 (molecular weight, van der Waals volume) (QSAM methods: PLS and PCA)**	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. J Med Chem. 1998 Jul 2;41(14):2481-91.
235	G	**Principal porperty scale Z3(electrophilicity, electronegativity) (QSAM methods: PLS and PCA)**	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. J Med Chem. 1998 Jul 2;41(14):2481-91.

Table IX.2 The number of randomly generated codes more robust than the standard genetic code for 226 Amino acid property scales (Pt, P1, P2 and P3). Genetic code robustness defined from the unbiased-weighted mean phenotypic change computed under the block-based model. Pt: The number of genetic codes more robust than the standard code for the whole block-based model. P1, P2 and P3: The number of genetic codes more robust than the standard code for the partial block-based models of edges between codon positions 1, 2 and 3, respectively. Pnumb: Numerical identifiers of the Amino acid property scales. aa: Amino acids. A: Alpha.B: beta, Asa: accessible Surface area.

Pnumb	Amino acid property names	P1	P2	P3	Pt
1	Scores for adenine-protein interaction	**1924977**	**1799147**	**990679**	**445281**
2	aa Helix Propensities in B/A proteins	**7889574**	**9467636**	**6529127**	**9413418**
3	aa Alpha Helical Propensities	**8730000**	**8853246**	**1238692**	**8462966**
4	aa Helix Propensities in alfa proteins	**6809377**	**6262395**	**2096990**	**8763340**
5	aa B-sheet Propensities in B+A proteins	**92495**	**8097226**	**997314**	**920819**
6	aa B-sheet Propensities in B proteins	**515434**	**7692498**	**83276**	**894800**
7	aa B-sheet Propensities in B/A proteins	**540133**	**6725196**	**1195493**	**1224272**
8	aa Alpha Helical Propensities	**8279072**	**8488203**	**473810**	**8417431**
9	Buried Alpha Helix solvent accessibilities	**5069743**	**9918349**	**341644**	**8021623**
10	aa Coil Propensities in B+A proteins	**1556324**	**9867968**	**2518686**	**6588562**
11	aa Coil Propensities in B/A proteins	**1242745**	**9738595**	**1340579**	**5491289**
12	Coil Propensities in B-sheet proteins jiang 1997	**495699**	**8325132**	**89826**	**1299296**
13	Helix Propensities in pept. whithout Helix-stabilizing schain interactions.	**8933750**	**9201058**	**416078**	**8507273**
14	Normalized frequency of Alpha-Helix	**1371506**	**5567746**	**2856733**	**2016153**
15	coil Accessible Surface area	**6581583**	**9811455**	**1859170**	**8637350**
16	aa Alpha Helix Propensities	**7119445**	**9299079**	**3938994**	**8595049**
17	aa Rotational Potentials in Alpha-Helix	**8623714**	**9536758**	**880679**	**8826639**
18	Exposed Alpha Helix solvent accessibilities	**353487**	**2096665**	**1972102**	**243252**
19	Helix Propensities in B+A proteins	**7247316**	**9444656**	**6348144**	**9181857**
20	conformational preference parameter for membrane-Buried helices	**45657**	**6404750**	**236421**	**57392**
21	Free energy for a-Helical conformation	**6954186**	**8794040**	**2495724**	**7853168**
22	thermodynamic scale for the aa Helix-forming tendencies	**8546526**	**8425752**	**1472612**	**8441056**
23	aa Rotational Potentials in Alpha Helix	**8448518**	**9260655**	**552990**	**8475045**
24	aa Rotational Potentials in Alpha Helix	**7229330**	**8326810**	**3491223**	**7787549**
25	aa Rotational Potentials in Alpha Helix	**6131808**	**7727083**	**2418808**	**6524634**
26	Helicity in water, 0222nm Circular dicroism (CD) spectra is used as a measure of Helicity (model peptides).	**8176816**	**9466091**	**1224835**	**8707127**
27	Helicity in n-butanol, 0222nm CD spectra is used as a measure of Helicity (model peptides).	**2295748**	**8215438**	**247**	**3043568**
28	statistical transmembrane Alpha-Helix Propensities in single-spanning proteins	**21906**	**9477014**	**324530**	**662700**
29	Free energy for a Helical region based on psi-phi matrices	**8409108**	**9356993**	**1469688**	**8617791**
30	Helix-coil stability constants	**7068129**	**7157722**	**1295430**	**6923280**
31	Helix-forming tendency in thermostable proteins	**6000901**	**8806880**	**4705404**	**7886743**
32	Helix propensity scale	**8526402**	**7689640**	**481975**	**7528482**
33	Helix propensity scale	**8639387**	**9308143**	**1959029**	**8909942**
34	Helix propagation Propensities	**9031317**	**9154745**	**717207**	**8530176**
35	transmembrane Alpha-Helix propensity	**9218**	**7298338**	**61939**	**23968**
36	Helix propensity scale	**4309673**	**9314768**	**3550941**	**7340990**
37	Asa in coil structures	**551412**	**5912867**	**4742413**	**1876342**

Pnumb	name	P1	P2	P3	Pt
38	Total Asa in folded beta s structures	**17446**	**2607333**	**312755**	**2181**
39	Hydrophilic Asa in folded beta s structures	**35323**	**5618438**	**1441403**	**168640**
40	Hydrophobic Asa in folded beta s structures	**259825**	**9829358**	**132552**	**1548712**
41	Total Asa in folded coil structures	**32370**	**3928490**	**64573**	**5105**
42	Hydrophilic Asa in folded coil structures	**37462**	**6155263**	**12956**	**19962**
43	Hydrophobic Asa in folded coil structures	**41089**	**4212336**	**70363**	**7607**
44	Hydrophilic Asa in folded proteins	**6998**	**4999102**	**58580**	**4266**
45	Hydrophobic Asa in folded proteins	**44037**	**4605839**	**62675**	**9885**
46	Total Asa in folded proteins	**21020**	**4243877**	**49401**	**3458**
47	Asa in Alpha Helix	**401872**	**5389053**	**1209952**	**559566**
48	Accessible Surface area	**22243**	**2728341**	**131094**	**3284**
49	Accessible Surface area in Beta strands	**1584771**	**5088808**	**5523351**	**2893860**
50	Total accessible Surface area	**4738399**	**4339589**	**5490501**	**4576471**
51	Hydrophilic accessible Surface area	**534952**	**9700672**	**817940**	**3432830**
52	Hydrophobic accessible Surface area	**140141**	**1482935**	**412026**	**8537**
53	Optimized beta-structure-coil equilibrium constant	**8992186**	**8510938**	**597101**	**8470503**
54	Buried Beta sheet solvent accessibility	**583761**	**4361765**	**167076**	**280715**
55	B sheet propensity	**277973**	**8677042**	**21319**	**1270462**
56	Propensity of Amino acid residues to occur in Isolated E-strand	**2497049**	**7485229**	**88683**	**2480691**
57	Propensity of Amino acid residues to occur in Edge β-strand	**8147176**	**7461063**	**121795**	**7277581**
58	Exposed Beta sheet solvent accessibility	**4418271**	**8627193**	**2792561**	**6416356**
59	Propensity of Amino acid residues to occur in Inner β-strand	**289708**	**8115925**	**27424**	**738851**
60	B sheet propensity	**3261796**	**9809546**	**1661868**	**7301929**
63	Free energy for B-strand region	**1113933**	**6845199**	**54071**	**1113625**
64	Free energy for B-strand region	**7727065**	**7249105**	**14077**	**7135887**
65	Free energy for B-strand conformation	**8425445**	**7182219**	**159848**	**7638946**
71	Cytosine_protein interaction	**445870**	**649081**	**8928751**	**334373**
72	Lipid accessibilities within the transmembrane Helix 1	**14214**	**6237910**	**316687**	**35720**
73	Buried coil solvent accessibility	**1448212**	**2847693**	**8969909**	**3092145**
74	Exposed coil solvent accessibility	**1329699**	**690902**	**2264976**	**2457166**
75	Random coil propensity	**2831186**	**9233626**	**2581698**	**6459179**
76	Position-Dependent Propensities for Polyproline II Helices	**5474497**	**6148001**	**87400**	**5237918**
77	Position-Dependent Propensities for Polyproline II Helices	**7097378**	**6913250**	**5273917**	**7541368**
78	Position-Dependent Propensities for Polyproline II Helices	**2558446**	**1381159**	**5239947**	**1416711**
79	Position-Dependent Propensities for Polyproline II Helices	**5970287**	**7619804**	**1085551**	**6071910**
80	Position-Dependent Propensities for Polyproline II Helices	**8068206**	**6250373**	**1012899**	**6818306**
81	Amino Acid Propensities in Polyproline II Helices L3	**6682899**	**6602984**	**416526**	**6306322**
82	Position-Dependent Propensities for Polyproline II Helices	**6373838**	**8542176**	**230559**	**7265264**

83	Position-Dependent Propensities for Polyproline II Helices	**4541942**	**6655113**	**2842128**	**4907366**
84	Position-Dependent Propensities for Polyproline II Helices	**8486424**	**4754660**	**4651303**	**7101135**
85	Amino Acid Propensities in Polyproline II Helices L+3	**6819500**	**6166696**	**1422182**	**6108347**
86	Position-Dependent Propensities for Polyproline II Helices	**7564883**	**6375653**	**1770173**	**6715413**
87	Position-Dependent Propensities for Polyproline II Helices	**6725994**	**4122457**	**1284211**	**4264444**
88	Position-Dependent Propensities for Polyproline II Helices	**7289649**	**6373628**	**2982414**	**6714235**
89	difference between Side-chain conformational entropies of Amino acids in the a-Helical and the coil states	**4603373**	**6839137**	**3928732**	**5508765**
90	Conformational entropy differences between free and Buried states of aa Side-chains	**3579483**	**4354921**	**5341142**	**3798707**
91	Absolute entropy	**7290912**	**9274785**	**4897379**	**4897379**
92	The Side-chain conformational entropies of Amino acids in the coil states	**8101860**	**5441874**	**484300**	**5388714**
93	Backbone Entropy of All Residues in the Coil Library	**7344823**	**7995020**	**4538**	**6724890**
94	Mean changes in Side-chain conformational entropy	**5245400**	**4424159**	**2561957**	**3898449**
95	The Side-chain conformational entropies of Amino acids in the a-Helical structures	**7750735**	**5495525**	**906521**	**5450036**
96	Side-chain conformational entropy	**333112**	**2878251**	**4748683**	**575899**
97	Side-chain conformational entropy				
98	Side-chain conformational entropy	**8613255**	**8756413**	**205198**	**8096713**
99	Sequence-dependence of backbone entropy OPLS-AA-01	**8322602**	**7917573**	**1615708**	**7984508**
100	Sequence-dependence of backbone entropy AMBER 94	**5172227**	**5493142**	**5938442**	**5593827**
101	Sequence-dependence of backbone entropy G-S-94	**7960326**	**5128673**	**584218**	**5174473**
102	Sequence-dependence of backbone entropy OPLS-UA	**6617863**	**8117723**	**1096385**	**6667360**
103	Scores for guanine-protein interaction	**5179699**	**7454362**	**151701**	**3980913**
104	High thermodynamic stability	**7533046**	**4955767**	**335379**	**4868366**
105	Alpha Helix 1 (p15) Propensities	**116781**	**3628998**	**7596702**	**1261992**
106	Alpha Helix 2 (p15) Propensities	**4456171**	**8789916**	**1527457**	**6082011**
107	Alpha Helix 3 (p15) Propensities	**6154446**	**9225965**	**6490426**	**8720573**
108	Hydrophobicity	**179975**	**8416664**	**128537**	**748453**
109	Lipid accessibilities within the transmembrane Helix 2	**2440559**	**1625350**	**4596927**	**1366721**
110	The free energy difference due to the main-chain conformational entropy between the a-Helical and the coil states	**7750735**	**5495525**	**906521**	**5450036**
111	Surface propensity scale	**95583**	**6863314**	**2014565**	**491304**
112	Metabolic costs of Amino acid biosynthesis, numbers of available hydrogen atoms in NADH, NADPH, and FADH2	**253301**	**354505**	**2521310**	**23563**
113	scaled Side Chain Hydrophobicity	**124358**	**7792909**	**27916**	**183161**
114	Hydrophobicity, free energies of transfer of aa from the solution to the Surface	**6116**	**8842175**	**197818**	**60069**
115	Hydrophobicity, for detecting amphipathic structures in proteins	**1117**	**7892854**	**190589**	**11390**
116	Hydrophobicity, (retention times on HPLC, ph 3)	**40275**	**6914710**	**70644**	**94563**
117	Hydrophobicity, (retention times on HPLC, ph 7.5)	**13191**	**7733114**	**4238**	**38425**
118	Consensus normalized Hydrophobicity scale	**1276799**	**9238166**	**1564296**	**4906093**
120	Hydrophobicity,Partition energy	**36520**	**6921425**	**666644**	**197640**
121	Hydrophilicity	**287388**	**3306331**	**683209**	**208967**
122	pure Hydrophobicity scale	**51545**	**5097606**	**3011013**	**186840**

123	Hydrophobicity	**214555**	**1230241**	**840738**	**20810**
124	Hydrophobicity	**39162**	**3736322**	**704501**	**77570**
125	Hydropathy index	**4140**	**9036906**	**141114**	**79875**
126	Hydrophobicity, free energy of transfer of aa from cyclohexylpyrrolidone to water	**3695701**	**5140584**	**32005**	**1752954**
127	Hydrophobicity, free energy of transfer of aa from etanol to water	**1284395**	**4372422**	**1030877**	**915566**
128	Hydrophobic parameter	**385139**	**4417404**	**837662**	**433838**
129	Hydrophobicity, ph 2.1 (retention times on HPLC)	**122928**	**3057784**	**872134**	**72620**
130	Hydrophobicity, ph 7.4 (retention times on HPLC)	**15533**	**919859**	**1258192**	**5236**
131	Hydrophobicity, partition coefficients (interior and Surface aa)	**38972**	**5374322**	**936474**	**145386**
132	Hydrophobicity, contact energies derived from protein structure	**2452**	**5686891**	**3389**	**903**
133	Hydrophobicity, Total free energy of hydration	**1850224**	**9969652**	**464943**	**7084216**
134	Helix propensity	**8760674**	**8549955**	**1060286**	**8457862**
135	Hydrophilicity, HPLC	**6417**	**5484519**	**124182**	**9882**
136	Average ratios between residue occurrences at the intra- and extracellular Sides of the membrane	**3922991**	**4685261**	**9363279**	**6069337**
137	Hydrophobicity,Accessibility reduction ratio	**25620**	**8306236**	**91019**	**111660**
138	Hydrophilicity of polar Amino acid Side-chains is markedly reduced by flanking peptide bonds	**492302**	**7962089**	**144024**	**954027**
139	distribution of Amino acid residues in transmembrane a-Helix bundles	**57377**	**4938990**	**3018017**	**440476**
141	Transmembrane central aa frequence	**52231**	**8691686**	**976515**	**445829**
142	Transmembrane extracelular aa frequence	**46158**	**5058868**	**1679970**	**194044**
143	Transmembrane intracelular aa frequence	**1231210**	**1753113**	**3864326**	**744985**
144	Transmembrane both termini aa frequence	**301725**	**5593613**	**3712822**	**1057753**
145	Transmembrane Total aa frequence	**1753505**	**2323384**	**4210348**	**1269574**
146	Transfer free energy to lipophilic phase	**2387228**	**9487692**	**2677701**	**6552722**
147	Hidrophobicity (HPLC)	**160160**	**1619095**	**27388**	**4536**
148	Hydrophobicity, Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water	**62823**	**77618**	**291333**	**325**
149	Polar requirement	**34153**	**2266517**	**998**	**2265**
150	Hydrophobicity, Hydration potential	**3245500**	**9399974**	**1046444**	**6704433**
151	Polarity	**1441678**	**7582381**	**710033**	**2052661**
152	Propensities of Amino acids for all protein-protein interfaces	**3006235**	**618933**	**1450316**	**448162**
153	Propensities of Amino acids for type I protein-protein interfaces	**1035224**	**7989**	**3477963**	**14190**
154	Propensities of Amino acids for type II protein-protein interfaces	**825245**	**5618523**	**3571343**	**1657399**
155	Propensities of Amino acids for type III protein-protein interfaces	**8760330**	**6567012**	**4310579**	**8080650**
156	aa isoelectric point	**5833210**	**4381307**	**7397764**	**5638653**
157	L1 (p15) Propensities	**7758934**	**2232066**	**4834520**	**5207232**
158	L2 (p15) Propensities	**9855716**	**7123734**	**893692**	**8789744**
159	L3 (p15) Propensities	**225994**	**3829890**	**7128463**	**1317715**
160	Average 11-20 Long-range contacts per residue	**403558**	**2068621**	**22515**	**48439**
161	Average 21-30 Long-range contacts per residue	**505563**	**3945109**	**2613838**	**863509**
162	Average 31-40 Long-range contacts per residue	**569085**	**1828559**	**2153612**	**361094**
163	Average 41-50 Long-range contacts per residue	**1478197**	**3740950**	**1751568**	**1196595**

164	Average p50 Long-range contacts per residue	**1549**	**5211155**	**102549**	**5188**
165	Average 4-10 Long-range contacts per residue	**58617**	**3295092**	**72985**	**25473**
166	LH (p15) Propensities	**6345018**	**7641084**	**1282205**	**6625884**
167	Medium Length linker propensity	**7406137**	**1753322**	**1892130**	**3475392**
168	Length 1 linker propensity	**8804812**	**1892583**	**1716712**	**5430194**
169	Length 2 linker propensity	**6791478**	**2415935**	**5424295**	**4601448**
170	Length 3 linker propensity	**5431494**	**1956150**	**1975137**	**2333304**
171	Total linker propensity	**8671865**	**1381509**	**992685**	**4345468**
172	Helical linker propensity	**8031137**	**8700418**	**5140224**	**8795068**
173	non-Helical linker propensity	**6795241**	**6955893**	**999229**	**6557131**
174	Small Length linker propensity	**1690**	**2414642**	**126271**	**1060**
175	Long Length linker propensity	**5148804**	**1086404**	**2580974**	**1698438**
176	Propensity of Amino acid residues to occur loops	**1306026**	**9355931**	**1509077**	**5021718**
177	Low thermodynamic stability	**2625379**	**2970315**	**3120171**	**1806119**
178	m1 (p15) Propensities	**21933**	**6315805**	**7147280**	**977469**
179	m2 (p15) Propensities	**1265629**	**9355387**	**328735**	**4289250**
180	m3(p15) Propensities	**9190572**	**7172046**	**426028**	**7930021**
181	The free energy difference due to the main-chain conformational entropy between the beta strand and the coil states	**1529871**	**8011611**	**1628956**	**3312602**
182	Average Medium-range contacts in globular proteins	**7105547**	**9578232**	**1984091**	**8509082**
183	Two rigid neighbors, Mean Location parameters (fit of the B-factors to a Gumbel distribution)	**101076**	**3445063**	**1258703**	**108422**
184	Two flexible neighbors, Mean Location parameters (fit of the B-factors to a Gumbel distribution)	**46361**	**1564039**	**548437**	**5540**
185	One rigid and one flexible neighbor, Mean Location parameters (fit of the B-factors to a Gumbel distribution)	**54856**	**1698797**	**640908**	**7414**
186	Mean Location parameters (fit of the B-factors to a Gumbel distribution)	**59039**	**2156701**	**878327**	**18139**
187	Amino acid melting point	**798202**	**6625835**	**2830792**	**2150690**
188	Medium thermodynamic stability	**104868**	**5883**	**3140009**	**1643**
189	Amino acid molecular weight	**2692101**	**2001593**	**4010199**	**1549599**
190	Free energy of residues at the N terminal position 1 in α-helices	**1233837**	**1203879**	**3455441**	**761192**
191	Free energy of residues at the N terminal position 2 in α-helices	**7735077**	**8636267**	**42733**	**6781336**
192	Free energy of residues at the N-cap position in α-helices	**528002**	**1161642**	**7887773**	**1141117**
193	Average partner number	**579993**	**1615198**	**1069427**	**146479**
194	Amino acid optical rotation	**6395794**	**6616239**	**2063102**	**5918243**
195	other (p15) Propensities	**2139463**	**8071802**	**955599**	**4241465**
196	Metabolic costs of Amino acid biosynthesis,numbers of High-energy phosphate bonds in ATP and GTP molecules	**505365**	**70542**	**5669636**	**102609**
197	phi (p15) Propensities	**6083155**	**7152755**	**770449**	**6103305**
198	Amino acid PK-C	**3217516**	**5512449**	**6013282**	**4624161**
199	Amino acid PK-N	**7105149**	**4428613**	**849223**	**4561508**
200	Propensity of Amino acid residues to occur polyproline type Helix	**6553621**	**8041859**	**472957**	**7035187**
201	Amino acid Propensities in polyproline II helices	**7354082**	**6096504**	**2444232**	**6360025**
202	r1 (p15) Propensities	**5516362**	**8998399**	**1434158**	**6734005**

203	r2 (p15) Propensities	**5483769**	**6399091**	**4655908**	**5991878**
204	r3 (p15) Propensities	**7424334**	**7755301**	**282251**	**7427097**
205	Sidechain Radii	**5486645**	**4311436**	**4161890**	**4529815**
206	cis-trans prolyl isomerisation Rate constants	**3723672**	**5842676**	**965741**	**2954065**
207	trans-cis prolyl isomerisation Rate constants	**7581049**	**2081287**	**1930515**	**3752661**
208	Refractivity	**2502312**	**314134**	**928112**	**176225**
209	Two flexible neighbors, Mean Scale parameters (fit of the B-factors to a Gumbel distribution)	**20277**	**1089522**	**72640**	**236**
210	One rigid and one flexible neighbor, Mean Scale parameters (fit of the B-factors to a Gumbel distribution)	**111931**	**4319109**	**13293**	**18129**
211	Two rigid neighbors, Mean scale parameters (fit of the B-factors to a Gumbel distribution)	**16888**	**4774643**	**179237**	**20564**
212	Total Mean scale parameters (fit of the B-factors to a Gumbel distribution)	**49137**	**3064301**	**35682**	**4582**
213	attenuation of the non local electrostatic energies	**2630395**	**9469340**	**218354**	**5700217**
214	Scores for thymine-protein interaction	**5998623**	**1490661**	**24218**	**1260423**
215	Total metabolic costs of Amino acid biosynthesis	**389884**	**49543**	**4143881**	**19013**
216	Buried Turn propensity	**485692**	**9362108**	**46465**	**1621988**
217	Exposed Turn propensity	**3653704**	**7619070**	**2880992**	**5093060**
218	Averaged Turn Propensities in transmembrane Helices	**475510**	**4569912**	**228103**	**149199**
219	Normalized Turn potential in transmembrane Helices	**141535**	**8151384**	**57637**	**213148**
220	Turn propensity	**1033995**	**9155264**	**26860**	**2791472**
221	Turn propensity	**1647186**	**9498542**	**1934948**	**5380069**
222	Potentials for position i of the type VIII Turn	**5899286**	**7944322**	**922532**	**6222882**
223	Potentials for position i+1 of the type VIII Turn	**2943344**	**5671896**	**737542**	**2548079**
224	Potentials for position i+2 of the type VIII Turn	**984446**	**7028347**	**2868842**	**2619576**
225	Potentials for position i+3 of the type VIII Turn	**6724516**	**8227120**	**2056020**	**7524673**
226	Residue volumes	**2073074**	**1832223**	**5102480**	**1399870**
227	Residue volumes	**1993660**	**2372185**	**4694130**	**1535166**
228	Residue volumes	**2335125**	**1935311**	**5223061**	**1625665**
229	Enthalpies of Gly-X-Hyp with aa in position x	**1994458**	**4937706**	**5631645**	**3146093**
230	Occurrences of aa in position x of Gly-X-Hyp	**6456005**	**6368349**	**1661830**	**5903596**
231	Enthalpies of Gly-Pro-Y with aa in position y	**1511364**	**5088943**	**6448598**	**3130247**
232	Occurrences of aa in position y of Gly-Pro-Y	**6205366**	**6611046**	**655215**	**5820307**
233	Principal property scale Z1 (Hydrophobicity) (QSAM methods: PLS and PCA)	**26457**	**7832670**	**74321**	**46695**
234	Principal property scale Z2 (molecular weight, van der Waals volume) (QSAM methods: PLS and PCA)	**2441179**	**2116342**	**3493935**	**1357986**
235	Principal porperty scale Z3(electrophilicity, electronegativity) (QSAM methods: PLS and PCA)	**4825834**	**4227768**	**8899181**	**6143427**

Table IX.3 List of Non-thermophiles (N=418) and thermophiles (N=324). Non-thermophiles: mesophilic and psychrophilic prokaryotes. Thermophiles: thermophilic and hyperthermophilic prokaryotes. TS: Thermic status: NT: Non-thermophiles, H: Thermophiles, Taxid: NCBI Taxonomy identifier, Assembly: NCBI Refseq Assembly identifiers. CDS: The coding sequence sizes.

Procaryote species and strain names	TS	Assembly	Taxid	CDS
Absiella dolichum DSM 3991	NT	GCF_000154285.1	428127	2101
Acholeplasma laidlawii PG-8A	NT	GCF_000018785.1	441768	1374
Achromobacter insuavis AXX-A	NT	GCF_000219745.1	1003200	5920
Acidovorax avenae subsp, avenae	NT	GCF_003029785.1	80870	3960
Acinetobacter baumannii DU202	NT	GCF_000498375.2	1370126	3840
Acinetobacter haemolyticus TG19599	NT	GCF_000302315.1	1221297	3054
Acinetobacter johnsonii SH046	NT	GCF_000162055.1	575586	3251
Acinetobacter junii SH205,txt	NT	GCF_000162075.1	575587	3069
Acinetobacter lwoffii ATCC 9957 = CIP 70.31	NT	GCF_000369125.1	1311804	3203
Actinobacillus pleuropneumoniae serovar 12 str. 1096	NT	GCF_000178635.1	754261	1977
Actinobacillus succinogenes 130Z	NT	GCF_000017245.1	339671	2101
Actinomyces bovis	NT	GCF_900444995.1	1658	2062
Aequorivita antarctica,txt	NT	GCF_900489835.1	153266	3431
Aequorivita capsosiphonis DSM 23843	NT	GCF_000429125.1	1120951	3530
Aequorivita sublithincola DSM 14238	NT	GCF_000265385.1	746697	3134
Aeromonas hydrophila 145	NT	GCF_000586035.1	1273135	4256
Aeromonas molluscorum 848	NT	GCF_000388115.1	1268236	3596
Aeromonas salmonicida subsp. Salmonicida	NT	GCF_001643305.1	29491	4108
Aggregatibacter actinomycetemcomitans serotype d str. SA3033	NT	GCF_001596385.1	1434260	2031
Aggregatibacter actinomycetemcomitans serotype e str. ANH9776	NT	GCF_001596315.1	1434265	1678
Agrobacterium tumefaciens str. Cherry 2E-2-2	NT	GCF_000349865.1	1281779	4976
Agromyces cerinus subsp. cerinus	NT	GCF_900142065.1	232089	3788
Alcaligenes faecalis subsp. faecalis NCIB 8687	NT	GCF_000275465.1	1156918	3409
Aliivibrio logei ATCC 35077	NT	GCF_000390125.1	1269941	4631
Aliivibrio salmonicida LFI1238,	NT	GCF_000196495.1	316275	4004
Alteromonas macleodii str. 'Balearic Sea AD45'	NT	GCF_000300175.1	1004787	3926
Anaeroarcus burkinensis DSM 6283	NT	GCF_000430605.1	1120985	3060
Aquimarina latercula DSM 2041	NT	GCF_000430645.1	1121006	5376
Aquimarina muelleri DSM 19832	NT	GCF_000430665.1	1121007	3893
Arcobacter nitrofigilis DSM 7299	NT	GCF_000092245.1	572480	3086
Arcticibacter svalbardensis MN12-7	NT	GCF_000403135.1	1150600	3898
Arthrobacter globiformis NBRC 12137	NT	GCF_000238915.1	1077972	4245
Arthrobacter luteolus NBRC 107841	NT	GCF_001552075.1	1216973	3508
Arthrobacter psychrolactophilus	NT	GCF_003219795.1	92442	3396
Azonexus hydrophilus DSM 23864	NT	GCF_000429605.1	1121032	3100
Azospira oryzae PS	NT	GCF_000236665.1	640081	3425
Azotobacter vinelandii CA6	NT	GCF_000380365.1	1283331	4776
Bacillus psychrosaccharolyticus ATCC 23296	NT	GCF_000305495.1	1174504	4247
Bacillus subtilis subsp. subtilis str. 168	NT	GCF_000155325.1	224308	4142
Bacillus thuringiensis serovar canadensis	NT	GCF_002147125.1	180855	5843
Bacteroides ovatus ATCC 8483	NT	GCF_000154125.1	411476	4769
Bacteroides thetaiotaomicron VPI-5482	NT	GCF_000011065.1	226186	4816
Bartonella henselae str. Houston-1	NT	GCF_000046705.1	283166	1520
Bdellovibrio bacteriovorus str,	NT	GCF_000525675.1	765869	2760
Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM	NT	GCF_000269965.1	391904	2443
Bordetella bronchiseptica RB50	NT	GCF_000195675.1	257310	4993
Bordetella pertussis STO1-CHOC-0017	NT	GCF_000479755.1	1331277	3366
Borrelia recurrentis A1	NT	GCF_000019705.1	412418	1029
Borreliella garinii IPT120	NT	GCF_000501815.1	1408846	690

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Brachybacterium squillarum M-6-3	NT	GCF_000225825.1	1074488	2760
Brachyspira hyodysenteriae ATCC 27164	NT	GCF_000383255.1	1266923	2586
Bradyrhizobium japonicum USDA 135	NT	GCF_000472945.1	1038863	6704
Brevundimonas aveniformis DSM 17977	NT	GCF_000428765.1	1121123	2547
Brucella melitensis ATCC 23457	NT	GCF_000022625.1	546272	3152
Brucella pinnipedialis B2/94	NT	GCF_000221005.1	520461	3276
Burkholderia pseudomallei DL36	NT	GCF_002887895.1	1436041	6808
Caedibacter taeniospiralis,txt	NT	GCF_002803295.1	28907	1187
Campylobacter coli BIGS0010	NT	GCF_000314205.1	1247735	1606
Campylobacter hyointestinalis subsp. hyointestinalis LMG 9260	NT	GCF_001643955.1	1031746	1704
Campylobacter jejuni subsp. jejuni	NT	GCF_002804585.1	32022	1543
Candidatus Blochmannia floridanus	NT	GCF_000043285.1	203907	587
Candidatus Blochmannia vafer str. BVAF	NT	GCF_000185985.2	859654	585
Candidatus Pelagibacter ubique HTCC7217	NT	GCF_000702645.1	1400525	1473
Cardiobacterium hominis ATCC 15826	NT	GCF_000160655.1	638300	2322
Cardiobacterium valvarum F0432	NT	GCF_000239355.1	797473	2334
Carnobacterium funditum DSM 5970	NT	GCF_000744185.1	1449337	2035
Carnobacterium maltaromaticum ATCC 35586	NT	GCF_000238575.1	1087479	3295
Cellulophaga algicola DSM 14237	NT	GCF_000186265.1	688270	4156
Cellulophaga baltica NN016038	NT	GCF_000477035.2	1348585	3876
Cellulophaga lytica DSM 7489	NT	GCF_000190595.1	867900	3248
Chlamydia trachomatis B/TZ1A828/OT	NT	GCF_000026905.1	672161	905
Chryseobacterium glaciei	NT	GCF_001648155.1	1685010	4277
Chryseobacterium greenlandense	NT	GCF_001507325.1	345663	3599
Citrobacter freundii GTC 09629	NT	GCF_000388155.1	1297584	4792
Citrobacter koseri ATCC BAA-895	NT	GCF_000018045.1	290338	4321
Clavibacter michiganensis subsp. michiganensis NCPPB 382	NT	GCF_000063485.1	443906	3114
Clavibacter nebraskensis NCPPB 2581	NT	GCF_000355695.1	1097677	2801
Cloacibacillus evryensis DSM 19522	NT	GCF_000585335.1	866499	3009
Clostridium botulinum E3 str. Alaska E43	NT	GCF_000020285.1	508767	3132
Clostridium botulinum NCTC 2916	NT	GCF_000171055.1	445335	3571
Clostridium kluyveri DSM 555	NT	GCF_000016505.1	431943	3804
Clostridium perfringens str. 13	NT	GCF_000009685.1	195102	2593
Clostridium vincentii	NT	GCF_002995745.1	52704	3246
Colwellia piezophila ATCC BAA-637	NT	GCF_000378625.1	1265503	4377
Colwellia polaris	NT	GCF_002104515.1	326537	3688
Colwellia psychrerythraea 34H	NT	GCF_000012325.1	167879	4441
Conexibacter woesei DSM 14684	NT	GCF_000025265.1	469383	5868
Crocosphaera watsonii WH 0401	NT	GCF_001039615.1	555881	3640
Cryobacterium flavum	NT	GCF_900103805.1	1424659	3682
Cryobacterium levicorallinum	NT	GCF_900113585.1	995038	3290
Cryobacterium luteum	NT	GCF_900110125.1	1424661	3452
Cryobacterium psychrotolerans	NT	GCF_900101115.1	386301	2930
Cryobacterium roopkundense	NT	GCF_000764165.1	1001240	3849
Cryptobacterium curtum DSM 15641	NT	GCF_000023845.1	469378	1337
Cupriavidus oxalaticus NBRC 13593	NT	GCF_001592245.1	1349762	5824
Cystobacter fuscus DSM 2262	NT	GCF_000335475.2	1242864	9615
Cytophaga aurantiaca DSM 3654	NT	GCF_000379725.1	1121373	3726
Cytophaga xylanolytica	NT	GCF_003254275.1	990	3339
Deinococcus frigens DSM 12807	NT	GCF_000701425.1	1121380	3700
Deinococcus marmoris DSM 12784	NT	GCF_000701405.1	1121381	4297
Deinococcus radiodurans ATCC 13939	NT	GCF_000687895.1	1408434	3030
Delftia acidovorans CCUG 274B	NT	GCF_000411195.1	883101	6196

Prokaryote species and strain names	TS	Assembly	Taxid	CDS
Demequina aestuarii	NT	GCF_000975095.1	327095	2554
Dermabacter hominis NBRC 106157	NT	GCF_001570785.1	1349750	1847
Dichelobacter nodosus VCS1703A	NT	GCF_000015345.1	246195	1277
Duganella zoogloeoides ATCC 25935	NT	GCF_000383895.1	1261617	5273
Eikenella corrodens CC92I	NT	GCF_000504685.1	1073362	2085
Enhydrobacter aerosaccus	NT	GCF_900167455.1	225324	6405
Enterococcus italicus DSM 15952	NT	GCF_001885995.1	888064	2154
Erwinia amylovora LA635	NT	GCF_000513415.1	1407062	3404
Erwinia pyrifoliae DSM 12163	NT	GCF_000026985.1	644651	3561
Erysipelothrix rhusiopathiae ATCC 19414	NT	GCF_000160815.2	525280	1612
Escherichia albertii NBRC 107761	NT	GCF_000759775.1	1115511	4125
Escherichia coli DSM 30083 = JCM 1649 = ATCC 11775	NT	GCF_000734955.1	866789	4859
Faecalibacterium prausnitzii M21/2	NT	GCF_000154385.1	411485	2868
Ferrimonas balearica DSM 9799	NT	GCF_000148645.1	550540	3746
Fibrobacter succinogenes subsp. elongatus	NT	GCF_003149165.1	706585	2816
Flavobacterium antarcticum DSM 19726	NT	GCF_000419685.1	1111730	2702
Flavobacterium flevense	NT	GCF_900142775.1	983	3446
Flavobacterium frigidarium DSM 17623	NT	GCF_000425505.1	1121890	3058
Flavobacterium frigoris PS1	NT	GCF_000252125.1	1086011	3311
Flavobacterium micromati	NT	GCF_900129585.1	229205	3072
Flavobacterium psychrophilum FPG3	NT	GCF_000754405.1	1452724	2311
Flavobacterium saccharophilum	NT	GCF_900142735.1	29534	4342
Flavobacterium segetis	NT	GCF_900129575.1	271157	2822
Flavobacterium urumqiense	NT	GCF_900108015.1	935224	2919
Flavobacterium xanthum	NT	GCF_900142695.1	69322	3157
Flavobacterium xueshanense	NT	GCF_900112975.1	935223	2959
Flectobacillus major DSM 103	NT	GCF_000427405.1	929703	4795
Flexibacter flexilis DSM 6793	NT	GCF_900112255.1	927664	3505
Francisella tularensis subsp. holarctica FSC200	NT	GCF_000168775.2	351581	1584
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	NT	GCF_000007325.1	190304	1983
Gardnerella vaginalis JCP8481B	NT	GCF_000414445.1	1261070	1183
Gemella haemolysans ATCC 10379	NT	GCF_000173915.1	546270	1628
Gemmatimonas aurantiaca T-27	NT	GCF_000010305.1	379066	3915
Giesbergeria anulus	NT	GCF_900111115.1	180197	3076
Glaciecola punicea ACAM 611	NT	GCF_000252165.1	1121923	2635
Glaciibacter superstes DSM 2113	NT	GCF_000421145.1	1121924	4414
Glaesserella parasuis ST4-1	NT	GCF_000690655.1	1399771	1998
Gloeobacter violaceus PCC 7421	NT	GCF_000011385.1	251221	4430
Gottschalkia acidurici 9a	NT	GCF_000299355.1	1128398	2838
Haemophilus influenzae Rd KW20	NT	GCF_000027305.1	71421	1610
Haemophilus parahaemolyticus HK385	NT	GCF_000262265.1	1095744	1881
Haemophilus parainfluenzae HK26	NT	GCF_000259485.1	1095745	1964
Halobiforma haloterrestris	NT	GCF_900112205.1	148448	4149
Halomonas alkaliantarctica	NT	GCF_000712975.1	232346	3356
Halomonas aquamarina	NT	GCF_900110265.1	77097	3128
Halomonas halodenitrificans DSM 735	NT	GCF_000620045.1	1121941	3040
Halomonas subglaciescola	NT	GCF_900142895.1	29571	2740
Halomonas titanicae BH1	NT	GCF_000336575.1	1204738	4753
Halorubrum distributum JCM 9100	NT	GCF_000337055.1	1227467	3073
Halorubrum lacusprofundi ATCC 49239	NT	GCF_000022205.1	416348	3456
Helicobacter bilis ATCC 43879	NT	GCF_000158435.2	613026	2160
Helicobacter pylori SA216A	NT	GCF_900005055.1	1345597	1447
Helicobacter rodentium ATCC 700285	NT	GCF_000687535.1	1449345	1764

Prokaryote species and strain names	TS	Assembly	Taxid	CDS
Holospora obtusa F1	NT	GCF_000469665.2	1399147	1064
Hymenobacter glacialis	NT	GCF_001816165.1	1908236	3375
Hymenobacter nivis	NT	GCF_003149515.1	1850093	4252
Hyphomonas johnsonii MHS-2	NT	GCF_000685275.1	1280950	3404
Hyphomonas oceanitis SCH89	NT	GCF_000685295.1	1280953	3982
Ignatzschineria larvae DSM 1322	NT	GCF_000510805.1	1111732	1958
Intrasporangium oryzae NRRL B-24470	NT	GCF_000576595.1	1386089	4076
Janibacter corallicola NBRC 107790	NT	GCF_001570965.1	1216969	2919
Janthinobacterium lividum PAMC 25724	NT	GCF_000242815.1	1112211	4152
Kandleria vitulina WCE2011	NT	GCF_000621925.1	1410659	1953
Kingella kingae KK274	NT	GCF_000470595.1	1305595	1697
Kingella oralis ATCC 51147	NT	GCF_000160435.1	629741	2351
Klebsiella aerogenes MGH 78	NT	GCF_000692215.1	1439323	4945
Klebsiella variicola At-22	NT	GCF_000025465.1	640131	5176
Kluyvera cryocrescens NBRC 102467	NT	GCF_001571285.1	1218112	4631
Komagataeibacter hansenii ATCC 23769	NT	GCF_000164395.1	714995	2978
Komagataeibacter xylinus NBRC 13693	NT	GCF_000964505.1	1234668	2891
Labedella gwakjiensis	NT	GCF_003014675.1	390269	3532
Lachnospira multipara ATCC 19207	NT	GCF_000424105.1	1282887	2470
Lactobacillus acidophi	NT	GCF_000011985.1	272621	1832
Legionella pneumophila subsp. pneumophila str.	NT	GCF_002002625.1	1206778	2953
Leifsonia rubra CMS 76R	NT	GCF_000477555.1	1348338	2468
Leifsonia xyli subsp. xyli str. CTCB07	NT	GCF_000007665.1	281090	2191
Leptospira interrogans serovar Hardjo str. Norma	NT	GCF_001293065.1	1279460	3893
Leptotrichia buccalis C-1013-b	NT	GCF_000023905.1	523794	2187
Leuconostoc citreum LBAE E16	NT	GCF_000239935.1	1127129	1718
Leuconostoc gelidum subsp. gasicomitatum	NT	GCF_001536305.1	1165892	1944
Listeria innocua Clip11262	NT	GCF_000195795.1	272626	3078
Listeria monocytogenes EGD	NT	GCF_000582845.1	1334565	2841
Lysobacter concretionis Ko07 = DSM 16239	NT	GCF_000768345.1	1122185	2575
Magnetospirillum magnetotacticum MS-1	NT	GCF_000829825.1	272627	4077
Malonomonas rubra DSM 5091	NT	GCF_900142125.1	1122189	3549
Mannheimia haemolytica serotype A1/A6 str.	NT	GCF_000584935.1	1450449	2043
Maribacter antarcticus DSM 21422	NT	GCF_000621125.1	1122191	3930
Marinobacterium georgiense DSM 11526	NT	GCF_900107855.1	1122198	3573
Marinomonas mediterranea MMB-1	NT	GCF_000192865.1	717774	4115
Marinomonas polaris DSM 16579	NT	GCF_900129155.1	1122206	4501
Mariprofundus ferrooxydans M34	NT	GCF_000379405.1	1188231	2586
Massilia glaciei	NT	GCF_003011895.2	1524097	5099
Melissococcus plutonius S1	NT	GCF_000747585.1	1385937	1595
Mesorhizobium loti R88b	NT	GCF_000517145.1	935548	6723
Methanobrevibacter smithii TS147C	NT	GCF_000189975.1	911125	1674
Methanococcoides burtonii DSM 6242	NT	GCF_000013725.1	259564	2406
Methanococcus aeolicus Nankai-3	NT	GCF_000017185.1	419665	1489
Methylobacterium platani JCM 14648	NT	GCF_001043885.1	1295136	5962
Methylococcus capsulatus str. Bath	NT	GCF_000008325.1	243233	2957
Micavibrio aeruginosavorus EPB	NT	GCF_000348745.1	349215	2284
Microbacterium oleivorans NBRC 103075	NT	GCF_001552475.1	1223528	2805
Microbulbifer marinus	NT	GCF_900107725.1	658218	3323
Microbulbifer yueqingensis	NT	GCF_900100355.1	658219	3145
Micrococcus luteus SK58	NT	GCF_000176875.1	596312	2279
Micromonospora aurantiaca ATCC 27029	NT	GCF_000145235.1	644283	6177
Microscilla marina ATCC 23134	NT	GCF_000169175.1	313606	7211
Moraxella catarrhalis 46P47B1	NT	GCF_000192945.1	857578	1624
Moraxella lacunata NBRC 102154	NT	GCF_001591245.1	1223506	2490

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Morganella morganii SC01	NT	GCF_000307755.2	1239989	3898
Moritella dasanensis ArB 0140	NT	GCF_000276805.1	1201293	4154
Moritella yayanosii	NT	GCF_900465055.1	69539	3701
Mucispirillum schaedleri ASF457	NT	GCF_900157005.1	1379858	2119
Mycobacterium leprae Br4923	NT	GCF_000026685.1	561304	2131
Mycobacterium tuberculosis D 4155	NT	GCF_000658535.1	1438824	4097
Myxococcus fulvus 124B02	NT	GCF_000988565.1	1334629	8460
Myxococcus xanthus DK 1622	NT	GCF_000012685.1	246197	7181
Neisseria lactamica Y92-1009	NT	GCF_000180595.1	869214	1927
Neisseria meningitidis ATCC 13091	NT	GCF_000146655.1	862513	2168
Neisseria polysaccharea ATCC 43768	NT	GCF_000176735.1	546267	1922
Neorickettsia sennetsu str. Miyayama	NT	GCF_000013165.1	222891	759
Nitriliruptor alkaliphilus DSM 45188	NT	GCF_000969705.1	1069448	5079
Nitrosomonas cryotolerans ATCC 49181	NT	GCF_900143275.1	1131553	2448
Nocardia farcinica IFM 10152	NT	GCF_000009805.1	247156	5747
Nocardiopsis xinjiangensis YIM 90004	NT	GCF_000341145.1	1246474	4383
Nosocomiicoccus ampullae	NT	GCF_001696685.1	489910	1544
Nostoc commune NIES-4072	NT	GCF_003113895.1	2005467	6724
Oceanicoccus sagamiensis	NT	GCF_002117105.1	716816	3900
Octadecabacter arcticus 238	NT	GCF_000155735.2	391616	5047
Oenococcus oeni DSM 20252 = AWRIB129	NT	GCF_000309445.1	1122618	1614
Oleispira antarctica RB-8	NT	GCF_000967895.1	698738	3866
Olsenella urininfantis	NT	GCF_900155635.1	1871033	1515
Orientia tsutsugamushi str. TA716	NT	GCF_000964855.1	1359175	1372
Paeniclostridium sordellii VPI 9048	NT	GCF_000444095.1	1292035	3335
Paeniglutamicibacter antarcticus	NT	GCF_900010755.1	494023	3865
Pantoea agglomerans 299R	NT	GCF_000330765.1	1261128	4163
Paraburkholderia bannensis NBRC 103871	NT	GCF_000685015.1	1218075	7442
Paraglaciecola arctica BSs20135	NT	GCF_000314995.1	493475	5024
Paraglaciecola hydrolytica	NT	GCF_001565895.1	1799789	4402
Paraglaciecola polaris LMG 21857	NT	GCF_000315055.1	1129793	4338
Parvimonas micra A293	NT	GCF_000493795.1	1408286	1498
Pasteurella multocida subsp. multocida str. HB03	NT	GCF_000512395.1	1147130	2050
Bacillus cereus ATCC 14579	NT	GCF_000007825.1	226900	5231
Pectobacterium carotovorum subsp. carotovorum	NT	GCF_001039055.1	1267545	4194
Pediococcus claussenii ATCC BAA	NT	GCF_900203775.1	487	2062
Pedobacter antarcticus 4BY	NT	GCF_000722885.1	1358423	3889
Pelosinus fermentans DSM 17108	NT	GCF_000271485.2	1122947	4485
Peptostreptococcus anaerobius 653-L	NT	GCF_000178095.1	596329	1787
Phenylobacterium immobile (ATCC 35973)	NT	GCF_001375595.1	31967	3148
Photobacterium frigidiphilum	NT	GCF_003025615.1	264736	5622
Photobacterium profundum SS9	NT	GCF_000196255.1	298386	5354
Photorhabdus luminescens NBAII H75HRPL105	NT	GCF_000826725.2	1429883	4139
Pirellula staleyi DSM 6068	NT	GCF_000025185.1	530564	4597
Piscirickettsia salmonis AUSTRAL-005	NT	GCF_000576045.2	1398558	3111
Planococcus donghaensis MPA1U2	NT	GCF_000189395.1	933115	3149
Planococcus halocryophilus Or1	NT	GCF_000342445.1	1005941	2737
Planococcus kocurii	NT	GCF_001465835.2	1374	3252
Planomicrobium glaciei CHR43	NT	GCF_000513535.1	1273538	3716
Plantibacter cousiniae	NT	GCF_900167175.1	199709	3667
Plesiomonas shigelloides 302-73	NT	GCF_000392595.1	1315976	3227
Polaribacter butkevichii	NT	GCF_002954605.1	218490	3308
Polaribacter filamentus	NT	GCF_002943715.1	53483	3511
Polaribacter glomeratus	NT	GCF_002954665.1	102	3352
Polaribacter irgensii 23-P	NT	GCF_000153225.1	313594	2400

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Polaribacter sp. MED152	NT	GCF_000152945.2	313598	2615
Polaribacter sp. SA4-10	NT	GCF_002163835.1	754397	2926
Polaromonas glacialis	NT	GCF_000709345.1	866564	4731
Polaromonas jejuensis NBRC 106434	NT	GCF_001598235.1	1321608	4706
Polaromonas naphthalenivorans CJ2	NT	GCF_000015505.1	365044	4879
Polaromonas sp. EUR3 1.2.1	NT	GCF_000688115.1	1305734	3950
Porphyromonas catoniae F0037	NT	GCF_000318215.2	1127696	1591
Porphyromonas gingivicanis JCM 15907	NT	GCF_000614585.1	1236526	1529
Prevotella intermedia ATCC 25611 = DSM 20706	NT	GCF_000439065.1	1122984	2141
Prochlorococcus marinus str. MIT 9301	NT	GCF_000015965.1	167546	1786
Propionibacterium freudenreichi	NT	GCF_900000085.1	1744	2163
Propionivibrio dicarboxylicus	NT	GCF_900099695.1	83767	4034
Proteus mirabilis PR03	NT	GCF_000372565.1	1279010	3402
Proteus mirabilis WGLW6	NT	GCF_000297815.1	1125694	3649
Pseudoalteromonas agarivorans DSM 14585	NT	GCF_002310855.1	1312369	3902
Pseudoalteromonas amylolytica	NT	GCF_001854605.1	1859457	4051
Pseudoalteromonas arctica A 37-1-2	NT	GCF_000238395.3	1117313	4061
Pseudoalteromonas atlantica T6c	NT	GCF_000014225.1	342610	4319
Pseudoalteromonas citrea DSM 8771	NT	GCF_000238375.2	1117314	4319
Pseudoalteromonas distincta	NT	GCF_000814675.1	77608	3888
Pseudoalteromonas espejiana DSM 9414	NT	GCF_002221525.1	1314869	3894
Pseudoalteromonas haloplanktis ATCC 14393	NT	GCF_000238355.1	1117315	4239
Pseudoalteromonas luteoviolacea S4054	NT	GCF_000974945.1	1129367	4868
Pseudoalteromonas sp. Bsw20308	NT	GCF_000310105.2	283699	3990
Pseudoalteromonas translucida KMM 520	NT	GCF_001465295.1	1315283	3569
Pseudoalteromonas undina DSM 6065	NT	GCF_000238275.2	1117320	3532
Pseudomonas aeruginosa ATCC 15442	NT	GCF_000504485.1	1424337	6338
Pseudomonas fluorescens LMG 5329	NT	GCF_000411675.1	1324332	6216
Pseudomonas fragi NBRC 3458	NT	GCF_002091615.1	1215101	4484
Pseudomonas grimontii	NT	GCF_900101085.1	129847	6286
Pseudomonas mendocina ZWU0006	NT	GCF_000798915.1	1339237	4659
Pseudomonas mucidolens NBRC 103159	NT	GCF_002091735.1	1215111	5084
Pseudomonas putida W15Oct28	NT	GCF_000708715.2	1449986	5425
Pseudomonas synxantha NBRC 3913	NT	GCF_002091795.1	1215118	6024
Pseudomonas syringae pv. actinidiae ICMP 18744	NT	GCF_000342185.1	1104680	5476
Pseudomonas viridiflava UASWS0038	NT	GCF_000307715.1	450396	5174
Pseudophaeobacter arcticus DSM 23566	NT	GCF_000473205.1	999550	4660
Psychrobacillus insolitus	NT	GCF_003254155.1	1461	3176
Psychrobacter arcticus 273-4	NT	GCF_000012305.1	259536	2113
Psychrobacter cryohalolentis K5	NT	GCF_000013905.1	335284	2505
Psychrobacter fozii	NT	GCF_003217155.1	198480	2822
Psychrobacter glacincola	NT	GCF_001411745.2	56810	2862
Psychrobacter piscatorii	NT	GCF_001444505.1	554343	2541
Psychrobacter urativorans	NT	GCF_001298525.1	45610	2385
Psychroflexus torquis ATCC 700755	NT	GCF_000153485.2	313595	3574
Psychromonas aquimarina ATCC BAA-1526	NT	GCF_000428725.1	1278312	4624
Psychromonas arctica DSM 14288	NT	GCF_000482725.1	1123036	3867
Psychromonas hadalis ATCC BAA-638	NT	GCF_000420245.1	1278302	3424
Psychromonas ingrahamii 37	NT	GCF_000015285.1	357804	3683
Psychromonas ossibalaenae ATCC BAA-1528	NT	GCF_000381745.1	1278307	4330
Psychroserpens burtonensis DSM 12212	NT	GCF_000425305.1	1123037	3381
Ralstonia pickettii NBRC 102503	NT	GCF_001544155.1	1218114	4296
Raoultella ornithinolytica NBRC 105727 = ATCC 31898	NT	GCF_001598295.1	1349784	5099
Rathayibacter caricis DSM 15933	NT	GCF_003044275.1	1328867	3736
Renibacterium salmoninarum ATCC 33209	NT	GCF_000018885.1	288705	2633
Rheinheimera baltica DSM 14885	NT	GCF_000425345.1	1123053	3828
Rhizobium leguminosarum bv. viciae GB30	NT	GCF_000419745.1	1041142	6985

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Rhodococcus erythropolis CCM2595	NT	GCF_000454045.1	1136179	5776
Rhodococcus rhodnii LMG 5362	NT	GCF_000389715.1	1273125	3993
Rhodoferax antarcticus ANT.BR	NT	GCF_001938565.1	1111071	3669
Rhodoferax ferrireducens T118	NT	GCF_000013605.1	338969	4479
Rickettsia prowazekii str. Dachau	NT	GCF_000277225.1	1105097	851
Robiginitomaculum antarcticum DSM 21748	NT	GCF_000365025.1	1123059	2584
Roseicitreum antarcticum	NT	GCF_900107025.1	564137	3809
Roseisalinus antarcticus	NT	GCF_900172355.1	254357	4439
Roseobacter denitrificans OCh 114	NT	GCF_900113215.1	375451	3986
Rothia mucilaginosa ATCC 25296	NT	GCF_000175615.1	553201	1715
Rubrobacter radiotolerans DSM 5868	NT	GCF_900175965.1	643560	3183
Ruminobacter amylophilus	NT	GCF_900115655.1	867	2211
Runella slithyformis DSM 19594	NT	GCF_000218895.1	761193	5744
Saccharibacter floricola DSM 15669	NT	GCF_000378165.1	1123227	2133
Salinibacterium xinjiangense	NT	GCF_900230175.1	386302	2787
Salipiger mucosus DSM 16094	NT	GCF_000442255.1	1123237	5245
Salmonella enterica subsp. enterica serovar	NT	GCF_000009505.1	550537	4502
Sandarakinorhabdus limnophila DSM 17366	NT	GCF_000420765.1	1123240	2456
Sanguibacter antarcticus	NT	GCF_002564005.1	372484	3091
Serratia liquefaciens ATCC 27592	NT	GCF_000422085.1	1346614	4811
Serratia odorifera DSM 4582	NT	GCF_000163595.1	667129	4596
Serratia proteamaculans 568	NT	GCF_000018085.1	399741	4999
Serratia symbiotica SCt-VLC	NT	GCF_900002265.1	1347341	1695
Shewanella algae JCM 14758	NT	GCF_000614935.1	1236541	3403
Shewanella baltica OS183	NT	GCF_000179535.2	693971	4213
Shewanella benthica KT99	NT	GCF_000172075.1	314608	3381
Shewanella frigidimarina NCIMB	NT	GCF_000014705.1	318167	4024
Shewanella halifaxensis HAW-EB4	NT	GCF_000019185.1	458817	4309
Shewanella loihica PV-4	NT	GCF_000016065.1	323850	3902
Shewanella marina JCM 15074	NT	GCF_000614975.1	1236542	3444
Shewanella oneidensis MR-1	NT	GCF_000146165.2	211586	4214
Shewanella piezotolerans WP3	NT	GCF_000014885.1	225849	4395
Shewanella psychrophila	NT	GCF_002005305.1	225848	5276
Shewanella putrefaciens JCM 201	NT	GCF_000018025.1	425104	3696
Shewanella sediminis HAW-EB3	NT	GCF_000018025.1	425104	4564
Shewanella violacea DSS12	NT	GCF_000091325.1	637905	3908
Shigella flexneri Shi06AH130	NT	GCF_000565825.1	1434144	4247
Simplicispira psychrophila DSM 11588	NT	GCF_000688255.1	1123255	3224
Sinorhizobium fredii USDA 257	NT	GCF_000265205.3	1185652	6292
Slackia exigua ATCC 700122	NT	GCF_000162875.1	649764	1719
Snodgrassella alvi wkB2	NT	GCF_000600005.1	1196094	2217
Solirubrobacter soli DSM 22325	NT	GCF_000600005.1	1196094	2217
Sphingomonas adhaesiva NBRC 150	NT	GCF_001569165.1	28182	3698
Sphingomonas aerolata	NT	GCF_003046295.1	185951	3343
Sphingomonas aurantiaca	NT	GCF_003050705.1	185949	3853
Sphingomonas jatrophae	NT	GCF_900113315.1	1166337	3732
Sphingopyxis granuli NBRC 100800	NT	GCF_001591045.1	1219060	3739
Spiroplasma mirum ATCC 29335	NT	GCF_000517365.1	838561	1099
Sporolactobacillus laevolacticus DSM 442	NT	GCF_000497245.1	1395513	3409
Sporosarcina globispora	NT	GCF_001274725.1	1459	5251
Sporosarcina psychrophila	NT	GCF_001590685.1	1476	4269
Staphylococcus aureus subsp. aureus N315	NT	GCF_000009645.1	158879	2776
Staphylococcus epidermidis ATCC 12228	NT	GCF_000007645.1	176280	2482
Staphylococcus hominis SK119	NT	GCF_000174735.1	629742	2106
Starkeya novella DSM 506	NT	GCF_000092925.1	639283	4401
Stenotrophomonas maltophilia stmalt0435	NT	GCF_000455625.1	1347913	4213

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Stenotrophomonas pictorum JCM 9942	NT	GCF_001310775.1	1236960	2346
Sterolibacterium denitrificans	NT	GCF_001586935.1	157592	814
Streptococcus anginosus subsp. whileyi MAS624	NT	GCF_000478925.1	1353243	1938
Streptococcus cristatus ATCC 51100	NT	GCF_900475445.1	889201	1880
Streptococcus mitis bv. 2 str. F0392	NT	GCF_000221165.1	768726	1798
Streptococcus pyogenes ATCC 10782	NT	GCF_000146715.1	864568	1765
Streptococcus suis 11538	NT	GCF_000440275.1	1214180	1997
Streptomyces alboniger	NT	GCF_001507305.1	132473	1840
Sulfurimonas autotrophica DSM 16294	NT	GCF_000147355.1	563040	2140
Synechococcus elongatus PCC 6301	NT	GCF_000010065.1	269084	2602
Tenacibaculum ovolyticum DSM 18103	NT	GCF_000430545.1	1123347	3555
Thermomonas fusca DSM 15424	NT	GCF_000423885.1	1123377	2800
Thiobacillus thioparus DSM 505	NT	GCF_000373385.1	1123393	3071
Thiocapsa marina 5811	NT	GCF_000223985.1	768671	4701
Tistrella mobilis KA081020-065	NT	GCF_000264455.2	1110502	5729
Treponema pallidum subsp. pallidum str. Nichols	NT	GCF_000410535.2	243276	1004
Tumebacillus permanentifrigoris	NT	GCF_003148565.1	378543	4234
Undibacterium pigrum	NT	GCF_003201815.1	401470	5619
Variovorax paradoxus NBRC 15149	NT	GCF_001591365.1	1321610	6173
Veillonella parvula DSM 2008	NT	GCF_000024945.1	479436	1824
Vibrio cholerae PhVC-311	NT	GCF_001027505.1	1399575	3578
Vibrio natriegens NBRC 15636 = ATCC 14048 = DSM 759	NT	GCF_001456255.1	1219067	4509
Vibrio parahaemolyticus 1911C	NT	GCF_002775145.1	1287654	4554
Vibrio vulnificus JY1701	NT	GCF_000269765.1	1035117	3911
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	NT	GCF_000008885.1	36870	636
Wolbachia endosymbiont of Cimex lectularius	NT	GCF_000829315.1	246273	981
Wolbachia pipientis wAlbB	NT	GCF_000242415.2	1116230	955
Xanthomonas arboricola pv. arracaciae	NT	GCF_002940565.1	487851	4193
Xanthomonas axonopodis pv. melhusii	NT	GCF_002019215.1	487834	4207
Xanthomonas campestris pv. campestris str. CN18	NT	GCF_900034305.1	1358019	4078
Xanthomonas campestris pv. vesicatoria str. 85-10	NT	GCF_001854165.1	316273	4465
Xanthomonas citri pv. phaseoli var. fuscans	NT	GCF_002759275.1	473423	4274
Xanthomonas oryzae ATCC 35933	NT	GCF_000482445.1	1313303	3425
Xenorhabdus budapestensis	NT	GCF_002632465.1	290110	3458
Xylella fastidiosa subsp. fastidiosa GB514	NT	GCF_000148405.1	788929	2074
Yersinia enterocolitica subsp. enterocolitica WA-314	NT	GCF_000297175.1	1194086	4023
Yersinia pestis CO92	NT	GCF_000009065.1	214092	3979
Yokenella regensburgei ATCC 49455	NT	GCF_000735455.1	911023	4479
Zymobacter palmae DSM 10491	NT	GCF_000620025.1	1123510	2510
Zymomonas mobilis subsp. mobilis ZM4 = ATCC 31821	NT	GCF_000007105.1	264203	1748
Acetomicrobium mobile DSM 13181	H	GCF_000266925.1	891968	2010
Acetomicrobium thermoterrenum DSM 13490	H	GCF_900107215.1	1120987	1840
Acidianus brierleyi	H	GCF_003201835.1	41673	2859
Acidianus hospitalis W1	H	GCF_000213215.1	933801	2332
Acidianus sulfidivorans JP7	H	GCF_003201765.1	619593	2270
Acidilobus saccharovorans 345-15	H	GCF_000144915.1	666510	1478
Acidimicrobium ferrooxidans DSM 10331	H	GCF_000023265.1	525909	2034
Acidothermus cellulolyticus 11B	H	GCF_000015025.1	351607	2152
Aciduliprofundum boonei T469	H	GCF_000025665.1	439481	1521
Aciduliprofundum sp. MAR08-339	H	GCF_000327505.1	673860	1491
Aeropyrum camini SY1 = JCM 12091	H	GCF_001316065.1	1198449	1628
Alicyclobacillus acidiphilus NBRC 100859	H	GCF_001544355.1	1255277	3549
Alicyclobacillus acidocaldarius LAA1	H	GCF_000173835.1	543302	2740
Alicyclobacillus contaminans DSM 17975	H	GCF_000429525.1	1120971	3152
Alicyclobacillus herbarius DSM 13609	H	GCF_000430585.1	1120972	3011
Alicyclobacillus hesperidum URH17-3-68	H	GCF_000294675.1	1200346	2757
Alicyclobacillus pomorum DSM 14955	H	GCF_000472905.1	1111479	3171

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Alicyclobacillus sendaiensis NBRC 100866	H	GCF_001552675.1	1220572	2606
Alicyclobacillus vulcanalis	H	GCF_900156755.1	252246	2782
Ammonifex degensii KC4	H	GCF_000024605.1	429009	2121
Anaerobranca californiensis DSM 14826	H	GCF_900142275.1	1120989	1978
Anaerobranca gottschalkii DSM 13577	H	GCF_900111575.1	1120990	2191
Anaerolinea thermophila UNI-1	H	GCF_000199675.1	926569	3105
Anoxybacillus flavithermus subsp. yunnanensis str. E13	H	GCF_000753835.1	1380408	2515
Anoxybacillus kamchatkensis G10	H	GCF_000283415.1	1212546	2910
Anoxybacillus tepidamans PS2	H	GCF_000620165.1	1382358	3292
Anoxybacillus thermarum	H	GCF_000836725.1	404937	2681
Aquifex aeolicus VF5	H	GCF_000008625.1	224324	1526
Archaeoglobus profundus DSM 5631	H	GCF_000025285.1	572546	1785
Archaeoglobus sulfaticallidus PM70-1	H	GCF_000385565.1	387631	2180
Archaeoglobus veneficus SNP6	H	GCF_000194625.1	693661	2072
Archaeoglobus fulgidus DSM 4304	H	GCF_000008665.1	224325	2369
Bacillus amyloliquefaciens EBL11	H	GCF_000559145.1	1457158	3719
Bacillus licheniformis DSM 13 = ATCC 14580	H	GCF_000011645.1	279010	4219
Bifidobacterium thermacidophilum subsp. porcinum DSM 17755	H	GCF_000771045.1	1435463	1504
Bifidobacterium thermophilum DSM 20212	H	GCF_000687575.1	1410648	1626
Caldibacillus debilis DSM 16016	H	GCF_000383875.1	1121917	3188
Caldicellulosiruptor acetigenus DSM 7040	H	GCF_000421725.1	1121259	2328
Caldicellulosiruptor bescii DSM 6725	H	GCF_000022325.1	521460	2599
Caldicellulosiruptor hydrothermalis 108	H	GCF_000166355.1	632292	2542
Caldicellulosiruptor kristjanssonii I77R1B	H	GCF_000166695.1	632335	2475
Caldicellulosiruptor lactoaceticus 6A	H	GCF_000193435.2	632516	2256
Caldicellulosiruptor naganoensis NA10	H	GCF_000955735.1	1387569	1991
Caldicellulosiruptor obsidiansis OB47	H	GCF_000145215.1	608506	2167
Caldicellulosiruptor owensensis OL	H	GCF_000166335.1	632518	2131
Caldicoprobacter faecalis	H	GCF_900115765.1	937334	2274
Caldicoprobacter oshimai DSM 21659	H	GCF_000526435.1	1304880	2398
Caldilinea aerophila DSM 14535 = NBRC 104270	H	GCF_000281175.1	926550	4038
Caldimicrobium thiodismutans	H	GCF_001548275.1	1653476	1765
Caldimonas manganoxidans ATCC BAA-369	H	GCF_000381125.1	1265502	3222
Caldimonas taiwanensis NBRC 104434	H	GCF_001592165.1	1349753	3143
Caldisalinibacter kiritimatiensis	H	GCF_000387765.1	1304284	2557
Caldisericum exile AZM16c01	H	GCF_000284335.1	511051	1496
Caldisphaera lagunensis DSM 15908	H	GCF_000317795.1	1056495	1491
Calditerricola satsumensis JCM 14719	H	GCF_001311905.1	1294024	1672
Calditerrivibrio nitroreducens DSM 19672	H	GCF_000183405.1	768670	2099
Caldivirga maquilingensis IC-167	H	GCF_000018305.1	397948	2005
Caldivirga sp. MU80	H	GCF_001663375.1	1650354	2161
Caloramator fervidus	H	GCF_900108045.1	29344	1735
Caloramator proteoclasticus DSM 10124	H	GCF_900129265.1	1121262	2437
Caminicella sporogenes DSM 14501	H	GCF_900142285.1	1121266	2263
Candidatus Desulforudis audaxviator MP104C	H	GCF_000018425.1	477974	2220
Carboxydothermus ferrireducens DSM 11255	H	GCF_000427565.1	1119529	2462
Carboxydothermus hydrogenoformans Z-2901	H	GCF_000012865.1	246194	2417
Carboxydothermus islandicus	H	GCF_001950325.1	661089	2369
Carboxydothermus pertinax	H	GCF_001950255.1	870242	2476
Chlorobaculum tepidum TLS	H	GCF_000006985.1	194439	2245
Chloroflexus aggregans DSM 9485	H	GCF_000021945.1	326427	3707
Chloroflexus aurantiacus Y-400-fl	H	GCF_000022185.1	480224	4056
Chloroflexus islandicus	H	GCF_001650695.1	1707952	3861
Chthonomonas calidirosea T49	H	GCF_000427095.1	1303518	2837

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Coprothermobacter platensis DSM 11748	H	GCF_000378005.1	1259795	1401
Coprothermobacter proteolyticus DSM 5265	H	GCF_000020945.1	309798	1409
Deinococcus geothermalis DSM 11300	H	GCF_000196275.1	319795	3003
Desulfacinum hydrothermale DSM 13146	H	GCF_900176285.1	1121390	3175
Desulfotomaculum acetoxidans DSM 771	H	GCF_000024205.1	485916	4007
Desulfotomaculum aeronauticum DSM 10349	H	GCF_900142375.1	1121421	3723
Desulfotomaculum alcoholivorax DSM 16058	H	GCF_000430885.1	1121422	3326
Desulfotomaculum alkaliphilum DSM 12257	H	GCF_000711975.1	1121423	2574
Desulfotomaculum arcticum	H	GCF_900113335.1	341036	4865
Desulfotomaculum australicum DSM 11792	H	GCF_900129285.1	1121425	2797
Desulfotomaculum geothermicum	H	GCF_900115975.1	39060	3536
Desulfotomaculum intricatum	H	GCF_001592105.1	1285191	2850
Desulfotomaculum profundi	H	GCF_002607855.1	1383067	2530
Desulfotomaculum thermocisternum DSM 10259	H	GCF_000686645.1	1121430	2562
Desulfotomaculum thermosubterraneum DSM 16057	H	GCF_900142025.1	1121432	3262
Desulfovirgula thermocuniculi DSM 16036	H	GCF_000429345.1	1121468	3031
Desulfurella acetivorans A63	H	GCF_000517565.1	694431	1814
Desulfurella amilsii	H	GCF_002119425.1	1562698	1982
Desulfurella multipotens	H	GCF_900101285.1	79269	1770
Desulfurobacterium atlanticum	H	GCF_900188395.1	240169	1728
Desulfurobacterium indicum	H	GCF_001968985.1	1914305	1625
Desulfurobacterium sp. TC5-1	H	GCF_000421485.1	1158318	1661
Desulfurobacterium thermolithotrophum DSM 11699	H	GCF_000191045.1	868864	1508
Desulfurococcus amylolyticus 1221n	H	GCF_000020905.1	490899	1386
Desulfurococcus amylolyticus DSM 16532	H	GCF_000231015.2	768672	1422
Desulfurococcus amylolyticus Z-533	H	GCF_000513855.1	1150674	1330
Desulfurococcus mucosus DSM 2161	H	GCF_001006085.1	675631	1206
Desulfurococcus mucosus DSM 2162	H	GCF_000186365.1	765177	1345
Dictyoglomus thermophilum H-6-12	H	GCF_000020965.1	309799	1862
Dictyoglomus turgidum DSM 6724	H	GCF_000021645.1	515635	1742
Dissulfuribacter thermophilus	H	GCF_001687335.1	1156395	2232
Ferroglobus placidus DSM 10642	H	GCF_000025505.1	589924	2479
Fervidicella metallireducens AeB	H	GCF_000601455.1	1403537	2585
Fervidicola ferrireducens	H	GCF_001562425.1	520764	2305
Fervidobacterium changbaicum	H	GCF_900100515.1	310769	1904
Fervidobacterium gondwanense DSM 13020	H	GCF_900143265.1	1121883	1936
Fervidobacterium islandicum	H	GCF_000767275.2	2423	1897
Fervidobacterium nodosum Rt17-B1	H	GCF_000017545.1	381764	1755
Fervidobacterium pennivorans DSM 9078	H	GCF_000235405.2	771875	1930
Fervidobacterium thailandense	H	GCF_001719065.1	1008305	1861
Geobacillus kaustophilus NBRC 102445	H	GCF_000739955.1	1220595	3284
Geobacillus stearothermophilus ATCC 7953	H	GCF_000705495.1	937593	2654
Geobacillus thermocatenulatus GS-1	H	GCF_000612265.1	1444308	3359
Geobacillus thermodenitrificans subsp. thermodenitrificans DSM 465	H	GCF_000496575.1	1413215	3263
Geobacillus thermoleovorans B23	H	GCF_000474195.1	1406857	3220
Geobacillus uzenensis	H	GCF_002217665.1	129339	3103
Geobacillus zalihae NBRC 101842	H	GCF_001544135.1	1220596	3364
Geoglobus acetivorans	H	GCF_000789255.1	565033	2159
Geoglobus ahangari	H	GCF_001006045.1	113653	1985
Geothermobacter sp. EPR-M	H	GCF_002093115.1	1969733	3251

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Geotoga petraea	H	GCF_900102615.1	28234	2022
Haloarcula californiae ATCC 33799	H	GCF_000337755.1	662475	4182
Haloferax elongans ATCC BAA-1513	H	GCF_000336755.1	1230453	3776
Halothermothrix orenii H 168	H	GCF_000020485.1	373903	2362
Hippea alviniae EP5-r	H	GCF_000420385.1	944480	1738
Hippea maritima DSM 10411	H	GCF_000194135.1	760142	1703
Hydrogenobacter hydrogenophilus	H	GCF_900215655.1	35835	1674
Hydrogenobacter thermophilus TK-6	H	GCF_000010785.1	608538	1870
Hydrogenobaculum sp. 3684	H	GCF_000213785.1	547143	1578
Hydrogenobaculum sp. SN	H	GCF_000348765.2	547146	1578
Hydrogenobaculum sp. Y04AAS1	H	GCF_000020785.1	380749	1597
Hyperthermus butylicus DSM 5456	H	GCF_000015145.1	415426	1675
Ignicoccus hospitalis KIN4/I	H	GCF_000017945.1	453591	1448
Ignicoccus islandicus DSM 13165	H	GCF_001481685.1	940295	1478
Isosphaera pallida ATCC 43644	H	GCF_000186345.1	575540	3893
Marinithermus hydrothermalis DSM 14884	H	GCF_000195335.1	869210	2174
Marinitoga hydrogenitolerans DSM 16785	H	GCF_900129175.1	1122195	2095
Marinitoga piezophila KA3	H	GCF_000255135.1	443254	2033
Meiothermus ruber H328	H	GCF_000346125.2	1297799	2857
Metallosphaera cuprina Ar-4	H	GCF_000204925.1	1006006	1894
Metallosphaera hakonensis	H	GCF_003201675.1	79601	2393
Metallosphaera sedula DSM 5348	H	GCF_000016605.1	399549	2298
Metallosphaera yellowstonensis MK1	H	GCF_000243315.1	671065	2680
Methanocaldococcus bathoardescens	H	GCF_000739065.1	1301915	1614
Methanocaldococcus fervens AG86	H	GCF_000023985.1	573064	1554
Methanocaldococcus infernus ME	H	GCF_000092305.1	573063	1437
Methanocaldococcus jannaschii DSM 2661	H	GCF_000091665.1	243232	1762
Methanocaldococcus sp. FS406-22	H	GCF_000025525.1	644281	1790
Methanocaldococcus villosus KIN24-T80	H	GCF_000371805.1	1069083	1346
Methanocaldococcus vulcanius M7	H	GCF_000024625.1	579137	1695
Methanoculleus thermophilus	H	GCF_001571405.1	2200	2171
Methanohalobium evestigatum Z-7303	H	GCF_000196655.1	644295	2267
Methanosarcina thermophila TM-1	H	GCF_000969885.1	523844	2597
Methanothermobacter marburgensis str. Marburg	H	GCF_000145295.1	79929	1701
Methanothermobacter tenebrarum	H	GCF_003264935.1	680118	1543
Methanothermobacter thermautotrophicus str. Delta H	H	GCF_000008645.1	187420	1756
Methanothermobacter wolfeii	H	GCF_900095815.1	145261	1655
Methanothermococcus okinawensis IH1	H	GCF_000179575.2	647113	1576
Methanothermococcus thermolithotrophicus DSM 2095	H	GCF_000376965.1	523845	1624
Methanothermus fervidus DSM 2088	H	GCF_000166095.1	523846	1296
Methanotorris igneus Kol 5	H	GCF_000214415.1	880724	1751
Methylacidiphilum infernorum V4	H	GCF_000019665.1	481448	2108
Moorella glycerini	H	GCF_001373375.1	55779	3432
Moorella humiferrea	H	GCF_002995755.1	676965	2583
Moorella mulderi DSM 14980	H	GCF_001594015.1	1122241	2919
Palaeococcus ferrophilus DSM 13482	H	GCF_000966265.1	588319	2246
Palaeococcus pacificus DY20341	H	GCF_000725425.1	1343739	1950
Parageobacillus caldoxylosilyticus NBRC 107762	H	GCF_000632715.1	1220594	3574
Parageobacillus thermantarcticus	H	GCF_900111865.1	186116	3263
Parageobacillus thermoglucosidasius NBRC 107763	H	GCF_000648295.1	1223501	3674
Persephonella hydrogeniphila	H	GCF_900215515.1	198703	2050
Persephonella marina EX-H1	H	GCF_000021565.1	123214	2033
Persephonella sp. IF05-L8	H	GCF_000703045.1	1158338	1876

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Persephonella sp. KM09-Lau-8	H	GCF_000703085.1	1158345	2160
Petrotoga miotherma DSM 10691	H	GCF_002895605.1	1434326	1876
Petrotoga mobilis SJ95	H	GCF_000018605.1	403833	1938
Picrophilus torridus DSM 9790	H	GCF_000711815.1	1123384	2058
Pseudothermotoga hypogea DSM 11164 = NBRC 106472	H	GCF_000711815.1	1123384	2058
Pseudothermotoga lettingae TMO	H	GCF_000017865.1	416591	2052
Pseudothermotoga thermarum DSM 5069	H	GCF_000217815.1	688269	1966
Pyrobaculum islandicum DSM 4184	H	GCF_000015205.1	384616	1966
Pyrococcus abyssi GE5	H	GCF_000195935.2	272844	1862
Pyrococcus furiosus DSM 3638	H	GCF_000007305.1	186497	1979
Pyrococcus horikoshii OT3	H	GCF_000011105.1	70601	1801
Pyrococcus kukulkanii	H	GCF_001577775.1	1609559	2064
Pyrococcus sp. ST04	H	GCF_000263735.1	1183377	1789
Pyrococcus yayanosii CH1	H	GCF_000215995.1	529709	1786
Pyrodictium occultum	H	GCF_001462395.1	2309	1617
Pyrolobus fumarii 1A	H	GCF_000223395.1	694429	1906
Rhodothermus marinus DSM 4252	H	GCF_000024845.1	518766	2865
Rhodothermus marinus SG0.5JP17-171	H	GCF_000565305.1	762569	2862
Rhodothermus profundi	H	GCF_900142415.1	633813	2605
Rubrobacter radiotolerans DSM 5868	H	GCF_900175965.1	643560	3183
Ruminiclostridium thermocellum BC1	H	GCF_000493655.1	1349417	2882
Sphaerobacter thermophilus DSM 20745	H	GCF_000024985.1	479434	3445
Spirochaeta thermophila DSM 6578	H	GCF_000184345.1	869211	2239
Staphylothermus hellenicus DSM 12710	H	GCF_000092465.1	591019	1586
Staphylothermus marinus F1	H	GCF_000015945.1	399550	1598
Streptococcus thermophilus TH1435	H	GCF_000521285.1	1415776	1608
Sulfobacillus thermosulfidooxidans str. Cutipay	H	GCF_000294425.1	1214914	3648
Sulfolobus acidocaldarius DSM 639	H	GCF_000012285.1	330779	2243
Sulfolobus islandicus REY15A	H	GCF_000189555.1	930945	2540
Sulfolobus metallicus DSM 6482 = JCM 9184	H	GCF_001316045.1	523847	1776
Sulfolobus solfataricus P2	H	GCF_000007005.1	273057	2829
Sulfurisphaera tokodaii str. 7	H	GCF_000011205.1	273063	2770
Sulfurivirga caldicuralii	H	GCF_900141795.1	364032	1620
Tepidibacter formicigenes DSM 15518	H	GCF_900142235.1	1123349	2464
Tepidiphilus margaritifer DSM 15129	H	GCF_000425565.1	1123354	2067
Thermaerobacter marianensis DSM 12885	H	GCF_000184705.1	644966	2302
Thermanaerovibrio acidaminovorans DSM 6589	H	GCF_000024905.1	525903	1730
Thermanaerovibrio velox DSM 12556	H	GCF_000237825.1	926567	1672
Thermoactinomyces daqus	H	GCF_000763315.1	1329516	3444
Thermoactinomyces vulgaris	H	GCF_001294365.1	2026	3232
Moorella thermoacetica ATCC 39073	H	GCF_000013105.1	264732	2460
Nautilia profundicola AmH	H	GCF_000021725.1	598659	1704
Thermoanaerobacter ethanolicus CCSD1	H	GCF_000175815.1	589861	1935
Thermoanaerobacter indiensis BSB-33	H	GCF_000373165.1	1125975	2371
Thermoanaerobacter italicus Ab9	H	GCF_000025645.1	580331	2265
Thermoanaerobacter kivui	H	GCF_000763575.1	2325	2173
Thermoanaerobacter mathranii subsp. mathranii str. A3	H	GCF_000092965.1	583358	2139
Thermoanaerobacter sp. YS13	H	GCF_000806225.2	1511746	2565
Thermoanaerobacter thermohydrosulfuricus WC1	H	GCF_000353265.2	1198630	2437
Thermoanaerobacter wiegelii Rt8.B1	H	GCF_000147695.2	697303	2424
Thermoanaerobacterium aotearoense SCUT27	H	GCF_000512105.1	1421016	2687
Thermoanaerobacterium sp. PSU-2	H	GCF_002102475.1	1930849	2475
Thermoanaerobacterium sp. RBIITD	H	GCF_900205865.1	1550240	3019
Thermoanaerobacterium thermosaccharolyticum DSM 571	H	GCF_000145615.1	580327	2566
Thermoanaerobaculum aquaticum	H	GCF_000687145.1	1312852	2281

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Thermobrachium celere DSM 8682	H	GCF_000430995.1	941824	2243
Thermococcus barophilus MP	H	GCF_000151105.2	391623	2182
Thermococcus celericrescens	H	GCF_001484195.1	227598	2331
Thermococcus cleftensis	H	GCF_000265525.1	163003	2033
Thermococcus eurythermalis	H	GCF_000769655.1	1505907	2180
Thermococcus gammatolerans EJ3	H	GCF_000022365.1	593117	2117
Thermococcus gorgonarius	H	GCF_002214385.1	71997	1774
Thermococcus guaymasensis DSM 11113	H	GCF_000816105.1	1432656	2029
Thermococcus kodakarensis KOD1	H	GCF_000009965.1	69014	2237
Thermococcus litoralis DSM 5473	H	GCF_000246985.2	523849	2306
Thermococcus nautili	H	GCF_000585495.1	195522	2132
Thermococcus onnurineus NA1	H	GCF_000018365.1	523850	1934
Thermococcus pacificus	H	GCF_002214485.1	71998	1868
Thermococcus peptonophilus	H	GCF_001592435.1	53952	1974
Thermococcus piezophilus	H	GCF_001647085.1	1712654	1856
Thermococcus profundus	H	GCF_002214585.1	49899	2075
Thermococcus radiotolerans	H	GCF_002214565.1	187880	1984
Thermococcus sibiricus MM 739	H	GCF_000022545.1	604354	1913
Thermococcus siculi	H	GCF_002214505.1	72803	2099
Thermococcus sp. 2319x1	H	GCF_001484685.1	1674923	2017
Thermococcus sp. 4557	H	GCF_000221185.1	1042877	2085
Thermococcus sp. 5-4	H	GCF_002197185.1	2008440	1961
Thermococcus sp. AM4	H	GCF_000151205.2	246969	2231
Thermococcus sp. EP1	H	GCF_001317345.1	1591054	1909
Thermococcus sp. PK	H	GCF_000430485.1	913025	2175
Thermococcus zilligii AN1	H	GCF_000258515.1	1151117	1789
Thermocrinis albus DSM 14484	H	GCF_000025605.1	638303	1580
Thermocrinis minervae	H	GCF_900142435.1	381751	1460
Thermocrinis ruber	H	GCF_000512735.1	75906	1594
Thermocrinis sp. GBS	H	GCF_000702425.1	1313265	1385
Thermocrispum agreste DSM 44070	H	GCF_000427905.1	1111738	3619
Thermodesulfatator atlanticus DSM 21156	H	GCF_000421585.1	1123371	2184
Thermodesulfatator autotrophicus	H	GCF_001642325.1	1795632	2128
Thermodesulfatator indicus DSM 15286	H	GCF_000217795.1	667014	2226
Thermodesulfobacterium commune DSM 2178	H	GCF_000734015.1	289377	1702
Thermodesulfobacterium geofontis OPF15	H	GCF_000215975.1	795359	1617
Thermodesulfobacterium hveragerdense DSM 12571	H	GCF_000423845.1	1123372	1689
Thermodesulfobacterium hydrogeniphilum	H	GCF_000746255.1	161156	1624
Thermodesulfobacterium thermophilum DSM 1276	H	GCF_000421605.1	1123373	1709
Thermodesulfobium acidiphilum	H	GCF_003057965.1	1794699	1721
Thermodesulfobium narugense DSM 14796	H	GCF_000212395.1	747365	1825
Thermodesulforhabdus norvegica	H	GCF_900114975.1	39841	2631
Thermodesulfovibrio aggregans	H	GCF_001514535.1	86166	1955
Thermodesulfovibrio islandicus DSM 12570	H	GCF_000482825.1	1123375	2042
Thermodesulfovibrio sp. N1	H	GCF_001707915.1	1871110	1905
Thermodesulfovibrio thiophilus DSM 17215	H	GCF_000423865.1	1123376	1867
Thermodesulfovibrio yellowstonii DSM 11347	H	GCF_000020985.1	289376	2028
Thermofilum adornatus	H	GCF_000446015.1	1365176	1825
Thermofilum carboxyditrophus 1505	H	GCF_000813245.1	697581	1849
Thermofilum pendens Hrk 5	H	GCF_000015225.1	368408	1866
Thermofilum uzonense	H	GCF_000993805.1	1550241	1641
Thermoflavifilum aggregans	H	CF_002797735.1	454188	2362
Thermogladius calderae 1633	H	GCF_000264495.1	1184251	1400
Thermoleophilum album	H	GCF_900108055.1	29539	2047
Thermomicrobium roseum DSM 5159	H	GCF_000021685.1	309801	2644
Thermomonas hydrothermalis	H	GCF_900129205.1	213588	2172
Thermoplasma acidophilum DSM 1728	H	GCF_000195915.1	273075	1521

Prokaryote species and strain names	TS	Assembly	taxid	CDS
Thermoplasma volcanium GSS1	H	GCF_000011185.1	273116	1545
Thermoproteus sp. CP80	H	GCF_002077075.2	1650659	1567
Thermoproteus tenax Kra 1	H	GCF_000253055.1	768679	1959
Thermoproteus uzoniensis 768-20	H	GCF_000193375.1	999630	2114
Thermosediminibacter oceani DSM 16646	H	GCF_000144645.1	555079	2195
Thermosipho affectus	H	GCF_001990485.1	660294	1740
Thermosipho africanus TCF52B	H	GCF_000021285.1	484019	1878
Thermosipho atlanticus DSM 15807	H	GCF_900129985.1	1123380	1566
Thermosipho melanesiensis BI429	H	GCF_000016905.1	391009	1877
Thermosipho sp. 1070	H	GCF_001682135.1	1437364	1735
Thermosipho sp. 1074	H	GCF_001999655.1	1643331	1777
Thermosipho sp. 1223	H	GCF_001999705.1	1643332	1742
Thermosphaera aggregans DSM 11486	H	GCF_000092185.1	633148	1368
Thermosulfidibacter takaii ABI70S6	H	GCF_001547735.1	1298851	1814
Thermosynechococcus elongatus BP-1	H	GCF_000011345.1	197221	2476
Thermotoga caldifontis AZM44c09	H	GCF_000828655.1	1408159	1937
Thermotoga naphthophila RKU-10	H	GCF_000025105.1	590168	1778
Thermotoga neapolitana DSM 4359	H	GCF_000018945.1	309803	1833
Thermotoga petrophila RKU-1	H	GCF_000016785.1	390874	1782
Thermotoga profunda AZM34c06	H	GCF_000828675.1	1408160	2072
Thermotoga sp. 2812B	H	GCF_000789335.1	1157948	1805
Thermotoga sp. Cell2	H	GCF_000789375.1	1157947	1631
Thermotoga sp. EMP	H	GCF_000294555.1	1157949	1799
Thermotoga sp. KOL6	H	GCF_002866025.1	126741	1689
Thermotoga sp. Mc24	H	GCF_000784835.1	1231241	1764
Thermotoga sp. RQ2	H	GCF_000019625.1	126740	1836
Thermotoga sp. RQ7	H	GCF_000832145.1	126738	1817
Thermotoga sp. SG1	H	GCF_002865985.1	126739	1799
Thermotoga sp. TBGT1765	H	GCF_000784795.1	1263836	1679
Thermotoga sp. TBGT1766	H	GCF_000784825.1	1230478	1680
Thermotoga sp. Xyl54	H	GCF_000784785.1	1235863	1695
Thermovibrio ammonificans HB-1	H	GCF_000185805.1	648996	1801
Thermus aquaticus Y51MC23	H	GCF_000173055.1	498848	2325
Thermus tengchongensis YIM 77401	H	GCF_000744175.1	1449357	2506
Vulcanisaeta distributa DSM 14429	H	GCF_000148385.1	572478	2420
Vulcanisaeta moutnovskia 768-28	H	GCF_000190315.1	985053	2357
Vulcanisaeta sp. EB80	H	GCF_002078205.2	1650660	2284
Vulcanisaeta thermophila	H	GCF_001748385.1	867917	2039