

**EVOLUTIONARY MODELS OF AMINO ACID SUBSTITUTIONS BASED ON
THEIR NEIGHBORHOOD TERTIARY STRUCTURE**

**MASTER OF SCIENCE IN MOLECULAR BIOLOGY AND BIOMEDICINE
MASTER THESIS**



**University of Crete
Department of Biology**

**Author: Elias Primetis
Supervisor: Dr Pavlos Pavlidis**

September 2017

Περίληψη

Οι ενδοπρωτεϊνικές αλληλεπιδράσεις βασίζονται στα αμινοξέα που παίρνουν μέρος σε αυτές. Υποτίθεται ότι οι ενεργειακά ευνοϊκές αλληλεπιδράσεις διατηρούνται, ενώ οι μη ευνοϊκές εξαλείφονται κατά την εξέλιξη. Εμείς χρησιμοποιήσαμε τον Protein Interaction Statistics (PrInS) αλγόριθμο για να περιγράψουμε στατιστικά τις αλληλεπιδράσεις των αμινοξέων, χρησιμοποιώντας πρωτεϊνικές δομές. Ο αλγόριθμος PrInS παράγει έναν βαθμολογικό πίνακα που περιγράφει την συχνότητα των αλληλεπιδράσεων μεταξύ των αμινοξέων στις πρωτεϊνικές δομές. Σε αυτή την έρευνα, χρησιμοποιήσαμε τις πρωτεϊνικές δομές των άλφα ελικοειδών πρωτεϊνών της μεμβράνης από την RCSB PDB βάση δεδομένων. Ο βαθμολογικός πίνακας που προέκυψε, μετατράπηκε σε έναν πίνακα αποστάσεων (Ευκλείδεια, Manhattan και Pearson) M των αμινοξέων, όπου η τιμή M_{ij} σημαίνει την απόσταση μεταξύ των γειτονιών των αμινοξέων i και j . Για να ελέγξουμε την εγκυρότητα της μεθοδολογίας μας, μετρήσαμε τις παρατηρημένες αμινοξικές υποκαταστάσεις στα 224 alignments ομόλογων πρωτεϊνών και τις συσχετίσαμε με τον πίνακα M . Υποθέτοντας της ανθρώπινες πρωτεϊνικές αλληλουχίες ως σημείο αναφοράς, υπολογίσαμε την απόσταση μεταξύ του ανθρώπου και των άλλων 19 ειδών (16 πρωτεύοντα και 3 άλλα θηλαστικά) για όλες τις 224 πρωτεΐνες των δεδομένων μας, χρησιμοποιώντας την προσέγγιση μας και τους πίνακες αμινοξικών υποκαταστάσεων BLOSUM62 και PAM120. Τα αποτελέσματα ήταν συγκρίσιμα, το οποίο υποδηλώνει ότι η δική μας προσέγγιση συλλαμβάνει πληροφορίες για την πρωτεϊνική εξέλιξη με παρόμοιο τρόπο με τους πίνακες αμινοξικών υποκαταστάσεων BLOSUM62 και PAM120. Τελικά, οι πίνακες αποστάσεων μετατράπηκαν σε πίνακες αναλογιών για τον υπολογισμό της πιθανοφάνειας των multiple alignments και των πρωτεϊνικών περιοχών κάθε multiple alignment.

Abstract

Intra-protein interactions depend on the involved amino acids. It is assumed that the energetically favourable interactions have been preserved, while the unfavourable have been eliminated during evolution. We have used the Protein Interaction Statistics (PrInS) algorithm to statistically describe interactions between amino acids using protein structures. PrInS produces a scoring matrix to describe the frequency of amino acid interactions in the protein structures. In this project, we used structures of alpha helical membrane proteins from the RCSB PDB database. The resulting scoring matrix was converted to an amino acids distance (Euclidean, Manhattan or Pearson) matrix M , where M_{ij} value denotes the distance between the neighbourhoods of amino acid i and j . To test the validity of our methodology, we counted the observed number of amino acid changes in 224 alignments of homologous proteins and we correlated them with the M matrix. Assuming human protein sequences as a reference, we calculated the distance between human and 19 other species (16 primates and 3 other mammals) for all 224 proteins of our dataset, using our approach, BLOSUM62 and PAM120 amino acid substitution matrices. Outcomes were comparable, suggesting that our approach captures information about protein evolution process in a similar fashion as BLOSUM62 and PAM120. Finally, distance matrices were converted to rate matrices to calculate the likelihood of multiple alignments and the likelihood of each site in alignments.

Acknowledgements

I would like to express my sincere thanks to Dr Pavlidis for his valuable help in the completion of this project.

Table of Contents

Abstract	I
Acknowledgements	II
1 Introduction	1
1.1 Protein Structure Significance and Computational Methods	1
1.2 Characteristics of Protein Evolution	1
1.3 Examples of Computational Methods in Protein Evolution	2
1.4 Summary	3
2 Methods	5
2.1 Downloading and Converting the Datasets	5
2.2 PrInS Software	5
2.3 Evaluation of PrInS Ability to Predict Amino Acid Substitutions	6
2.4 Multiple Alignment Scoring	6
2.5 Calculation of Overall and Site Likelihoods	7
3 Results	10
3.1 Initial Indications That PrInS Can Predict Amino Acid Substitutions	10
3.2 Heatmap Visualization of Multiple Alignment Scoring Reveals Clusters of Proteins	13
3.3 Optimization of Likelihoods Depends on the Rate Matrix	14
4 Discussion	17
4.1 PrInS Results Keep up with the Observed Amino Acid Substitutions	17
4.2 Similar Scoring Results and Functional Clustering of Proteins	18
4.3 Similar Likelihoods by Using Alignment and Structural Rate Matrices	19
4.4 Significance of the Results	20
4.5 Future Perspectives and Limitations	21
5 Conclusion	22
References	23
6 Supplementary Figures	25

1 Introduction

1.1 Protein Structure Significance and Computational Methods

The functionality and structure are closely related due to the three-dimensional structure of a protein. Its dynamics, amino acid sequence as well as various cofactors, ligands and other parts of the same or other proteins form a complex network of diverse interactions that form the basis of the unique physiochemical properties of each protein that are related to its function [1]. Moreover, during the last decade, the tertiary structures of many proteins have been determined by using various techniques, such as crystallography, NMR, electron microscopy and hybrid methods [2]. Consequently, the number of the available tertiary structures in the RCSB-PDB database is increasing rapidly.

Computational methods, such as modelling and dynamics can help in the structural, functional and evolutionary characterization of the proteins. Knowledge-based potentials that are derived on the basis of the frequency of the occurrence of interacting pairs of amino acids have been proved very useful in the modelling and the assessment of globular and not only proteins [3]. In the same fashion, an open-source software was developed by Dr Pavlidis and it is called Protein Interaction Statistics (PrInS). PrInS was developed to score residue by residue PDB files. As a result these interaction propensities are presented as interaction matrices. With these interaction matrices is possible to score every individual protein residue in order to identify the residues that are statistical outliers. The residues with the higher scores are involved in a few interactions while the residues with the lower scores are involved in many interactions. In this case interaction between two amino acids is defined when the distance between them is less than 6.5 Angstroms. As a result, the scoring matrices were used in order to evaluate the ability of PrInS to predict amino acid substitutions based on their tertiary structure.

1.2 Characteristics of Protein Evolution

There are two main principles in protein evolution. The first one is that it has been proved that protein structure is more conserved than the sequence, while the second one is that the physicochemical properties of amino acids constrain the structure, the function and the evolution of proteins [4], [5].

It has to be mentioned that evolutionary rate is strongly correlated with fractional residue burial. Within protein families the backbone changes are not so usual in order to preserve the folding profile over relatively long evolutionary distances, while single substitutions are found often at the side chains. For example, the proteins with binding function, the binding interface is under functional constraint and may evolve the slowest, with differences in rate between affinity-determining and specificity-determining residues. In addition, there is difference in the evolutionary rate of the secondary structure elements, with beta sheets to evolve more slowly than helical regions and random coils to evolve faster. Changes in secondary structure content after residue substitutions can occur due to varying helix/sheet propensity. Some of these changes in secondary structural composition are likely to be evolutionary neutral. In contrast, some structural transition involves positive selection. In this case, a new mutationally accessible fold may enable the development of a new function that was not possible within the previous fold [4].

In the context of folding the thermodynamic stability of the proteins with a stable unique tertiary structure is important. Thermodynamic stability is maintained throughout evolution despite the average destabilizing effect of the non-synonymous mutations. Also, the structure is important as a scaffold for properly orientating functional residues, such as a binding interface and a catalytic residue. As a result, there is little selective pressure for particular sequences (with functional properties in protein structure) within a structure over longer evolutionary periods, generating a neutral network of sequences connected by those accessible through the mutational process. Although, protein folds with excess thermodynamic stability are thought to possess more potential neofunctionalization. Consequently to this and previous knowledge, most of the mutations are either deleterious or neutral rather than adaptive both in terms of thermodynamic stability and fitness [4].

1.3 Examples of Computational Methods in Protein Evolution

Due to the maturity of computational molecular biology and computational molecular evolution, it is now possible to integrate more the biophysical and the evolutionary attributes of proteins. The understanding of the energy landscape of a single sequence and homologous sequences linked through the mutational process plays a crucial role in protein evolution [4].

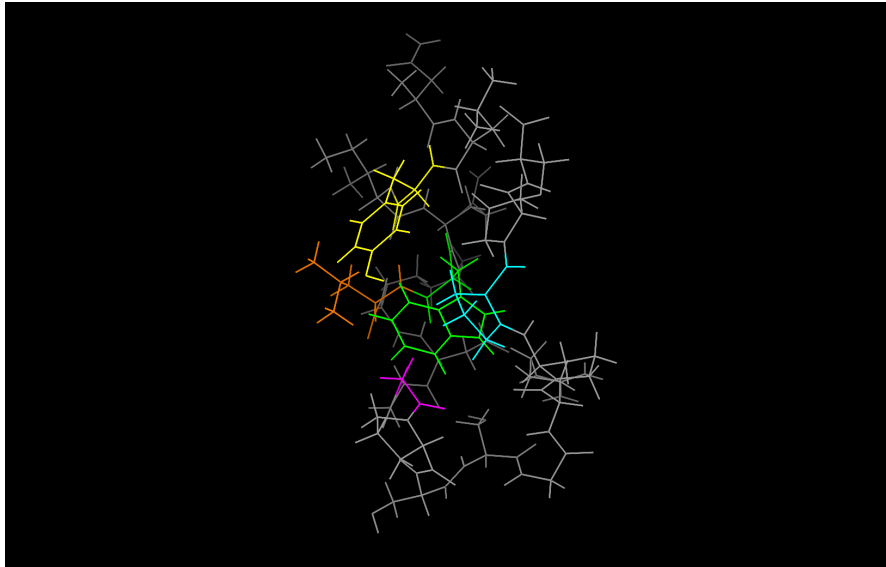
Modelling of sequence evolution in structurally ordered regions for evolutionary purposes has two main ongoing research topics. Retrospective analysis particularly in the construction of phylogenetic trees is one of the topics, where structural and biophysical considerations are viewed as an integral component of the protein evolution over long evolutionary distances. Also, attempts have been made to replace purely statistical

models that account for structure. In the second topic, the forward evolution of proteins or sequence simulation constrained by a fold that does not vary. In this topic, models to predict protein evolution were made by using evolutionary information about the proteins, which is already known. For both of the two paths informational and physical models are available. In informational models, average interaction propensities are used, but they suffer from a lack of folding specificity. On the contrary, physical models are based on inter-atomic or inter-residue interaction and they are considered as more specific than the informational models. Of course, both of them need improvements, such as better representation of side chains, which leads to a properly packed hydrophobic core [4]. In addition, there are more theoretical models, which describe the proteins evolution. It has been shown that conservation principles and their symmetries are fundamental in evolution and physical systems [5]. Also, the relationship between the local effects and the global constraint of conservation principles is well understood. In the same way, Halton and Warr showed that the protein evolution and other global properties of the system of proteins are constrained within structural bounds set by a conservation principle derived from information theory. At this point, it has to be mentioned that conservation of information does not operate at the local level, where the influences of natural selection are manifest in the variety of protein structure and function that is well understood. On the contrary, conservation of information is able to explain the global bounds within which the whole system of proteins is constrained. As a result, the evolution is constrained by conservation of information at different level from natural selection [5]. Another study in protein evolution by Choi and Kim based on the most common structural ancestor (CSA) and showed that not all present-day proteins evolved from one single set of proteins in the last common ancestral organism, but new common ancestral protein were born at different evolutionary times. These proteins are not traceable to one or two ancestral proteins, but they follow the rules of the "multiple birth model" for the evolution of protein sequence families [6].

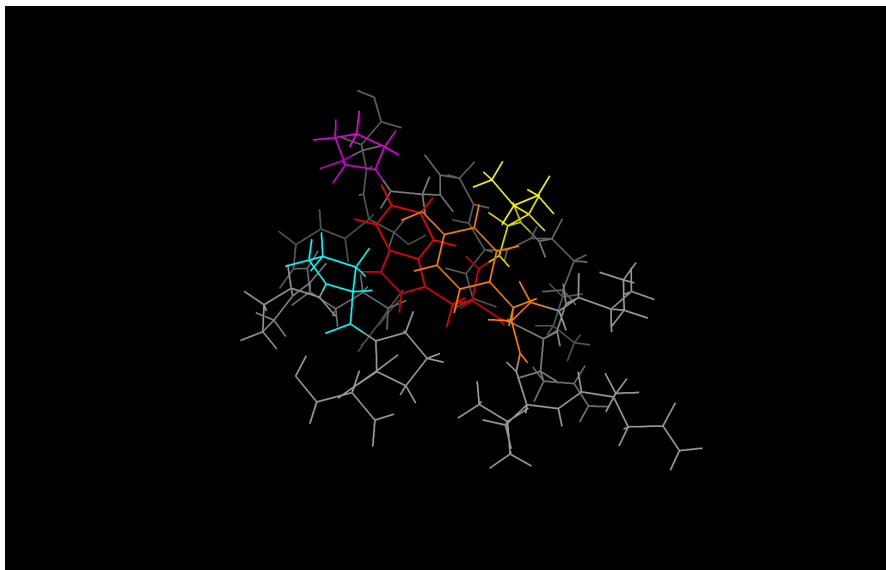
1.4 Summary

Summarizing, in this study we used the scoring matrices that are resulted from PrInS algorithm in order to examine the evolution of proteins and in this case the evolution of alpha helical membrane proteins. The central hypothesis of this project is that the evolution of proteins, in a specific protein family, can be described through amino acid substitutions by using protein structure data. Specifically, amino acids in different proteins that have the same amino acid neighbourhood can possibly substitute each other, as can be shown in Figure 1.1. This was possible by converting the scoring matrix into several different types of matrices in order to compare their results with the results of the model substitution matrices, such BLOSUM62 and PAM120, in the prediction of protein evolution [7], [8]. Finally, we chose this kind of proteins due to their biological significance in intracellular communication and coordination, as pores, ion channels and

receptors, as it has been mentioned in Jha publication [3].



(a) Protein A



(b) Protein B

Figure 1.1: Two amino acid neighbourhoods in two different proteins of the same protein family (e.g. alpha helical membrane proteins). (a) In the Protein A, green amino acid has specific neighbours, which are determined by their distance from the green amino acid. (b) In the Protein B, red amino acid has the same neighbours, which are also determined by their distance from the red amino acid, with the green amino acid in the protein A. According to our hypothesis green and red amino acids can substitute each other during evolution.

2 Methods

2.1 Downloading and Converting the Datasets

This research was started by selecting a protein category, in this case the alpha helical membrane proteins were used, whose proteins are going to be analysed. Then the names of these proteins were retrieved from UniProt Database [9]. Alongside, a multiple alignment file, which contains all the human amino acid sequences aligned with the same amino acid sequences in 19 different organisms (16 primates and 3 other mammals) was downloaded from the UCSC Genome Browser [10]. After the UniProt names were converted to Protein Data Bank (PDB) names in order to retrieve structural data for these proteins from the RCSB PDB [11]. Moreover, alongside with this conversion the UniProt names or the PDB names were converted in UCSC protein names in order to retrieve the alignments based on the UCSC protein name from the multiple alignment file. All these conversions were possible by the Biological Database Network [12]. Finally, the analysis of datasets was done by using several different scripts written in Python and R programming languages.

2.2 PrInS Software

Initially, the PrInS software was used in order to produce the a 60x60 scoring matrix from the PDB structural datasets of the alpha helical membrane proteins, which was used in Jha publication [3]. In this 60x60 scoring matrix, the pairs of amino acids with the lower scores are found often close (depends on the environment, see PrInS) in the tertiary structure of proteins. Afterwards, this 60x60 scoring matrix was divided in nine 20x20 matrices and different distances (Euclidean, Manhattan and Pearson) between the elements of these matrices were calculated. Then the nine 20x20 matrices of each category were added all together in order to create a 20x20 distance matrix for each calculated distance (Euclidean, Manhattan and Pearson). After, these distance matrices were visualised by using heatmaps. These heatmaps can be found in the Supplementary Figures 6.1a, 6.1c and 6.1d respectively. In these heatmaps, small distance values are depicted with dark grey colour, while high distance values with white. Lower distance value between two amino acids indicates similar "behaviour". If two amino acids has similar "behaviour" (low distance value in the distance matrices) they tend to have similar neighbours, while if two amino acids has no similar behaviour (high distance value in the distance matrices) they tend to have different neighbours. As a result, if two

amino acids have similar behaviour can possibly substitute each other according to our hypothesis. Finally, a 20x20 genetic code matrix was used as control and was generated by calculating the minimum number of base changes required to convert an amino acid to another amino acid. Genetic code matrix was the simplest way to show the amino acid substitutions and was used in all the analyses. Again this matrix was visualised by using a heatmap, which can be found in the Supplementary Figures 6.1b. In this heatmap, many codon changes are depicted with white, while a few codon changes with dark grey.

2.3 Evaluation of PrInS Ability to Predict Amino Acid Substitutions

For all the multiple alignment, a consensus matrix was created. The dimensions of this matrix is 20 (amino acids) by the length all the multiple alignments in row. In the consensus matrix the amino acid occurrences are counted for every amino acid position of the multiple alignments. Then, the occurrences of substitution pairs of amino acids (only two amino acids per position) from the consensus matrix were counted and a 20x20 matrix was created. Consequently, Pearson Correlation Coefficients between the previous matrix and the 20x20 distance and genetic code matrices were calculated. Pearson Correlation Coefficient is a measure of the linear correlation between two variables X and Y and in this case between two matrices. These Pearson Correlation Coefficients were calculated in order to investigate if the amino acid substitutions in the alignments can be correlated with the low distance values of the amino acids in the 20x20 distance (Euclidean, Manhattan and Pearson Squared) and 20x20 genetic code matrices.

Moreover, the amino acid pairs, between the leucine and the other amino acids, with the lowest distance values in the Euclidean, Manhattan and Pearson distance matrices were checked with Common Substitution Tool in the Amino Acid Explorer of NCBI [13]. The function of Common Substitution Tool is to collocate a list of amino acids from the most to the less common substitutions by given an amino acid as input, based on the BLOSUM62 substitution matrix [7].

2.4 Multiple Alignment Scoring

Moreover, the multiple alignments of alpha helical membrane proteins, which were retrieved from the multiple alignment file were scored with the four 20x20 matrices. As it is known the three of them are distance matrices (Euclidean, Manhattan and Pearson Squared), while the fourth one is a genetic code matrix. Consequently, these multiple alignments were also scored with two substitution matrices (BLOSUM62 and PAM120) [7], [8]. The results of scoring were visualized with the help of heatmaps in order to evaluate the evolution of proteins in 20 different species. A result of heatmap

visualization was the identification of protein clusters, due to the default clustering (hierarchical clustering) in heatmaps that made in R. These protein clusters were intense in the heatmap that was resulted from the scoring of the multiple alignments with the Euclidean distance matrix. Then, the proteins in these specific clusters were identified and their functions were found via gProfiler [14]. gProfiler is a public web server for characterising and manipulating gene lists resulting from mining high-throughput genomic data. Finally, it was checked if these proteins clusters can be located in the other heatmaps, which were resulted from the scoring of the alignments with other distance, genetic code and substitution matrices.

2.5 Calculation of Overall and Site Likelihoods

The four 20x20 matrices (distance and genetic code) were transformed in rate matrices by using a similar procedure to the procedure that was used in the article of Dayhoff in which the PAM substitution matrices were firstly introduced [8]. This procedure in order to convert the Euclidean distance matrix into rate matrix is depicted in the algorithm below (Listing 2.1) written in Python programming language. The input of the algorithm is a symmetric matrix (distance or genetic code matrix), S . Given S , the following processes are performed:

1. Find the maximum value of the matrix, S_{max} .
2. Subtract each element from the S_{max} and divide by S_{max} . This will convert the S from a distance matrix to a amino acid "similarity" symmetric matrix, where the value 1 denotes maximum "similarity" between two amino acids.
3. Similar to other rate substitution matrices (BLOSUM62, PAM120), one element of the matrix is defined as a reference. All other elements are scaled proportionally to the reference.
4. The diagonal of the matrix is defined as the negative sum of all other elements of the respective row.

Likelihood

Likelihood is a function of the parameters of a statistical model for given data. Likelihood functions play a key role in statistical inference, especially methods of estimating a parameter from a set of statistics.

In this case, we wanted to estimate the likelihood of a rate matrix for a given multiple alignment. The phylogenetic tree and the equilibrium frequencies of the amino acids are known. In other words, we wanted to check which rate matrix can explain better the observed multiple alignments.

Then, the rate matrices, equilibrium frequencies and phylogenetic trees were used to calculate the overall likelihoods and the site likelihoods of multiple alignments. Overall likelihood is the likelihood of a protein, while a site likelihood is the likelihood of an amino acid position within a protein. The phylogenetic tree was retrieved from the UCSC Genome Browser [10], and the equilibrium frequencies of amino acids were identical as the equilibrium frequencies used in the BLOSUM62 model. After the previous calculations, PrInS rate matrices that were resulted from the distances matrices, were evaluated for their ability to calculate the two kinds of likelihoods.

Listing 2.1: Python Script to Convert Distance and Genetic Code Matrices to Rate Matrices (Euclidean Distance Matrix Example)

```

1 #command to import the libraries that are going to be used in this script
2 import numpy as np
3 import math
4
5 #function to return the absolute value of each element of a matrix (e.g Euclidean
   distance matrix)
6 def linearDistance(e1):
7     return abs(e1)
8
9 #function to convert a distance matrix to rate matrix
10 #matrix: a symmetric distance matrix (e.g Euclidean)
11 #minSimilarity: the minimum similarity between two amino acids
12 #distance: calculated by the function above
13 def R_matrix(matrix, minSimilarity, distance):
14     mtx=np.zeros((len(matrix), len(matrix)))
15     maxel = matrix.max()
16     for i in range(matrix.shape[0]):
17         for j in range(matrix.shape[1]):
18             w = distance((maxel - matrix[i,j])/maxel)
19             w = w + minSimilarity
20             if i==j:
21                 w=0
22             mtx[i,j]=w
23     refElement = mtx[mtx.shape[0]-1, mtx.shape[1]-2]
24     for i in range(matrix.shape[0]):
25         for j in range(matrix.shape[1]):
26             mtx[i,j] = mtx[i,j]/refElement
27             s = np.sum(mtx[i,])
28             mtx[i,i] = -s
29     return(mtx)
30
31 #function to take the lower trianglular matrix of a given square matrix (e.g rate
   matrix) and print it into a file
32 def pml_matrices(mat,name):
33     fileq=open(name, 'w')
34     print("", file=fileq)
35     for i in range(mat.shape[0]):
36         for j in range( i ):
37             print(mat[i,j], end='', file=fileq)
38             if(j < i-1):
39                 print(" ", end='', file=fileq)
40             else:
41                 print("\n", end='', file=fileq)
42     fileq.close()

```

In addition, to summarize the comparison between the Euclidean rate matrix and the BLOSUM62 rate matrix for the whole set of proteins, we assessed the number of times that the Euclidean rate matrix results in a higher likelihood value than the BLOSUM62 rate matrix for each amino acid pair. Assume a and b a given pair of amino acids and let k_{ab} being the sites in the whole set of proteins consisting solely by a and b . Also, let $L_E(i)$ and $L_B(i)$ represent the likelihoods using the Euclidean rate matrix and BLOSUM62 rate matrix for the site i , respectively.

Let,

$$I(i) = \begin{cases} 1, & \text{if } L_E(i) > L_B(i) \\ 0, & \text{otherwise} \end{cases} \quad (2.1)$$

Then,

$$P_{ab} = \frac{\sum_{i \in k_{ab}} I(i)}{k} \quad (2.2)$$

counts the proportion of sites consisting by a and b , which have a higher likelihood value for the Euclidean rate matrix. Results were stored in a 20x20 amino acid matrix P , where P_{ab} is provided by the Formula 2.2. This matrix was visualized as a heatmap.

3 Results

3.1 Initial Indications That PrInS Can Predict Amino Acid Substitutions

In this part of the analysis, a 20x20 matrix was constructed that contains all the amino acid substitutions pairs that occurred in these 224 multiple alignments. Consequently this matrix was checked if it is correlated with the distance matrices (Euclidean, Manhattan and Pearson) and the genetic code matrix. The result of the Pearson Correlation Coefficient showed that there is correlation between this consensus substitution matrix and all the other matrices (distance and genetic code matrices) for the most of the amino acids. In the Figure 3.1, the correlation between the consensus and the Euclidean distance matrices is shown. In this type of figures, the correlation is negative, while the disassociation is positive due to the distance matrices that were used in the Pearson Correlation Coefficient test. More specifically, in consensus substitution matrix the larger values show the amino acid substitutions with the more counts, while in distance matrices the amino acids with the most common "behaviour" in neighbourhood terms have lower values. The graphs of correlation between the consensus and other matrices (Genetic Code, Manhattan and Pearson) are found in Supplementary material Figures 6.2a- 6.2c.

As it is obvious from the Figure 3.1, the most of the amino acids are correlated positively between the consensus matrix and the distance matrices. For example in Figure 3.1, the uncorrelated amino acids were glycine, proline, arginine, serine, tryptophan and tyrosine, while in the Figure 6.2b (Consensus VS Manhattan), the uncorrelated amino acids were proline, arginine, serine, tryptophan and tyrosine. Moreover, the profile of the uncorrelated amino acids was changed slightly in Figure 6.2c (Consensus VS Pearson), where these amino acids were glycine, histidine, asparagine, glutamine, arginine, tryptophan and tyrosine. The results of the previous correlation (Consensus VS Pearson) showed some different uncorrelated amino acids that had not been appeared in the previous correlations. In the other correlations the same amino acids were not strongly correlated. Finally, there are no uncorrelated amino acids in the correlation between the Consensus matrix and the genetic code matrix (Figure 6.2a).

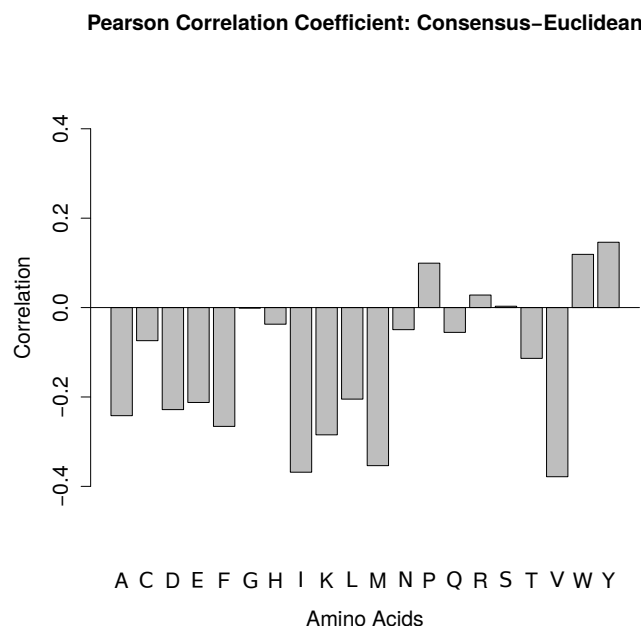



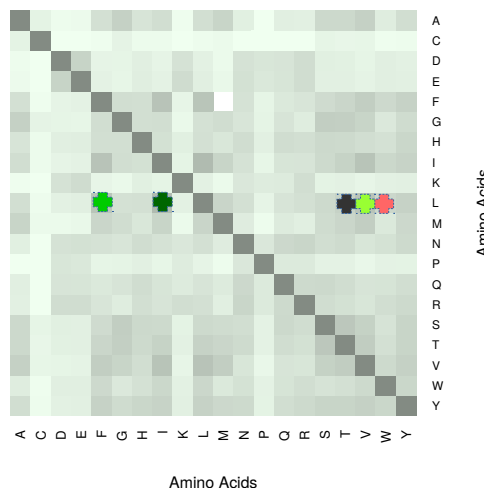
Figure 3.1: Pearson Correlation Coefficient between the consensus and Euclidean distance matrix. The correlation is represented with the negative values, while the disassociation with positive values, due to the different nature of these two matrices.

In addition in this step of the analysis, the distance matrices, were resulted from the initial analysis of PrInS matrix, are able to predict the amino acid substitutions. More specifically, the amino acid pairs with the lowest values in these distance matrices are found on the top of the amino acid list that the Common Substitution Tool of NCBI Amino Acid Explorer collocate [13]. That means that the amino acids with the most common "behaviour" tend to substitute each other. For example, in Euclidean distance matrix leucine-isoleucine pair has the lowest distance value and also according to the substitution list of Common Substitution Tool this substitution is the most common in both ways. This means that leucine substitutes isoleucine very easily and isoleucine substitutes leucine very easily too. Moreover, in the Figure 3.2, it is shown the first five most common amino acid substitutions of leucine in Common Substitution Tool (Figure 3.2a) and in Euclidean distance matrix (Figure 3.2b). It is clear that the Euclidean distance matrix can predict the most common substitutions of leucine, but in different order. For example, the order in Common Substitution Tool is isoleucine, methionine, valine, phenylalanine and alanine, while in Euclidean distance matrix is isoleucine, phenylalanine, valine, tryptophan and tyrosine. On the one hand, it has to be clarified that methionine is the sixth most common substitution of leucine, while alanine is the tenth most common substitution of leucine according to the Euclidean distance matrix. On the other hand, tryptophan and tyrosine are the seventh and the eighth most common substitutions of leucine according to the Common Substitution Tool. These most common substitution differences are possibly based on the fact that the generation

of the Euclidean distance matrix is based only on the structural information of a protein family, while the Common Substitution Tool is based on BLOSUM62 substitution matrix that was created from alignments from multiple protein families.

1-letter code	3-letter code	Chemistry	Potential H-bonds	Molecular Weight	Isoelectric Point	Hydrophobicity
L	Leu	CH ₂ -C-C-	1.0	113	6.0	0.918
I	Ile	CH ₂ -C-C-	1.0	113	6.0	1.000
M	Met	CH ₂ -C-C-	1.0	131	5.7	0.811
V	Val	CH ₂ -C-C-	1.0	99	6.0	0.923
F	Phe	CH ₂ 	1.0	147	5.5	0.951
A	Ala	CH ₂ -C-C-	1.0	71	6.0	0.806

(a) First Five Most Common Substitution of Leucine Based on Common Substitution Tool



(b) First Five Most Common Substitutions of Leucine Based on Euclidean Distance Matrix

Figure 3.2: The first five most common substitutions of leucine according to (a) Common Substitution Tool and (b) Euclidean Distance Matrix. The same colouration is used in both of the parts of this figure and was taken from the Common Substitution Tool. Dark green shows the first most common substitution, while the red shows the fifth most common substitution. Methionine and alanine are not found among the first five most common substitution of leucine in Euclidean distance matrix, while tryptophan and tyrosine are not found among the first most common substitutions in Common Substitution Tool. These are the amino acid differences between these two subfigures.

3.2 Heatmap Visualization of Multiple Alignment Scoring Reveals Clusters of Proteins

At the alignment scoring part of the analysis, heatmaps were created in order to depict the sequence difference of the proteins in 19 different species compared with the human. As it is clear from the heatmaps in Figure 3.3, many proteins do not significantly differ from the human homologue proteins (grey colour). The colour scale in these heatmaps varies from grey to white, which denote high and low sequential similarity between the proteins respectively. Of course, the species that are evolutionary close to the human have less protein sequence differences than the species whose common ancestors with human lived million years ago. These heatmaps were plotted due to the need to detect differences between the distance/genetic code matrices and model substitution matrices. As it is clear from Figure 3.3 there are no significant differences in the scoring of the multiple alignments with substitution and distance matrices.

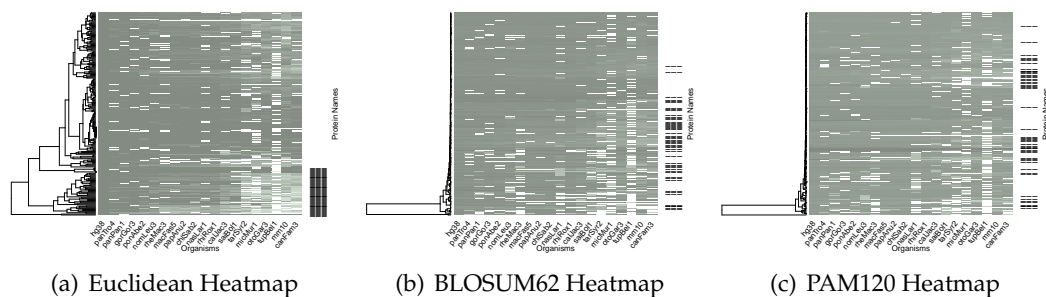


Figure 3.3: Heatmaps depict the scoring of multiple alignments with the (a) Euclidean distance matrix, (b) BLOSUM62 substitution matrix and (c) PAM120 substitution matrix. In this figure there is also shown the hierarchical clustering of proteins and the species, whose proteins were scored. Human is found in the left end of the heatmaps. The other species were put in evolutionary order. Finally, the proteins that were found on the protein cluster in Euclidean heatmap are represented in the two other heatmaps with three dashes.

Another interesting characteristic from these heatmaps was the clusters that were made due to the hierarchical clustering. As it clear from all these heatmaps in Figure 3.3 and more specifically the Euclidean heatmap Figure 3.3a, there are protein clusters that were made from the default hierarchical clustering of heatmap.2 function in R. First, the proteins in this protein cluster have many sequence differences in the evolutionary distant species from human. Also, the proteins in the protein cluster that were observed in the Euclidean heatmap, were identified and their function was checked by using gProfiler [14]. The result of this analysis showed that these proteins have a distinct function from the proteins that were found in other clusters. On the one hand, the proteins that are located in this discrete protein cluster play a crucial role in the intestinal

absorption of phytosterol and in cholesterol and lipid transportation. On the other hand, the rest of the proteins, which are not found in this discrete cluster are responsible for the homeostatic and transducer/receptor mechanisms of the cells. Consequently, it was checked if this specific protein cluster can be found in other heatmaps. The proteins that were found in the distinct cluster in Euclidean heatmap are depicted with three dashes in all the heatmaps. As a result there is a slight similarity of protein clusters between the Euclidean and the genetic code, Manhattan and Pearson heatmaps (Supplementary Figures 6.3a- 6.3c). In addition, the proteins that were found in the protein cluster in Euclidean heatmap are more dispersed in the model substitution matrices heatmaps, BLOSUM62 (Figure 3.3b) and PAM120 (Figure 3.3c).

3.3 Optimization of Likelihoods Depends on the Rate Matrix

By converting the PrInS distance matrices and the genetic code matrix into rate matrices, it was possible to calculate the likelihood and the site likelihoods for every protein. For the most of the proteins, 170 out of 223, the likelihoods were higher by using the rate matrices that were constructed from the default BLOSUM62 and PAM120 substitution matrices [7], [8]. For the rest of them, 53 out of 223, the Euclidean Distance Matrix had higher likelihoods.

In site likelihood terms, the things are not so clear and mainly depends on each protein. More specifically, after the subtraction of site likelihoods (default site likelihoods (e.g BLOSUM62) - PrInS site likelihoods (e.g Euclidean)) for every protein, the negative difference of site likelihoods indicated that the PrInS rate matrices are more reliable in the calculation of these specific site likelihoods. In the Figure 3.4, the differences between the BLOSUM62 and the Euclidean site likelihoods for the NCKX1 protein are depicted. This protein was selected randomly in order to show an example of site likelihoods. NCKX1 is a critical component of the visual transduction cascade, controlling the calcium concentration of outer segments during light and darkness [15]. Finally, in the Supplementary Figures 6.4a- 6.4c are shown the differences of the site likelihoods between the BLOSUM62 rate matrix and the genetic code rate matrix 6.4a, the Manhattan rate matrix 6.4b and the Pearson rate matrix 6.4c.

As it was mentioned before, matrix P was visualized as a heatmap (Figure 3.5), where black boxes represent the amino acid pairs, for which Euclidean rate matrix always calculates higher site likelihoods. Also, the different shades of grey from the dark grey to the lighter grey in the other boxes show the decrease of the ability of the Euclidean rate matrix to calculate always higher site likelihoods for specific amino acid pairs. Moreover, the not a number (NaN) results are represented with white boxes and with a red cross in the centre. NaN results are found when specific kinds of amino acid pairs are not found in the multiple alignments (divide zero with zero). Finally, the boxes in the diagonal are

always white with red crosses, because amino acid pairs of the same amino acid are not designated.

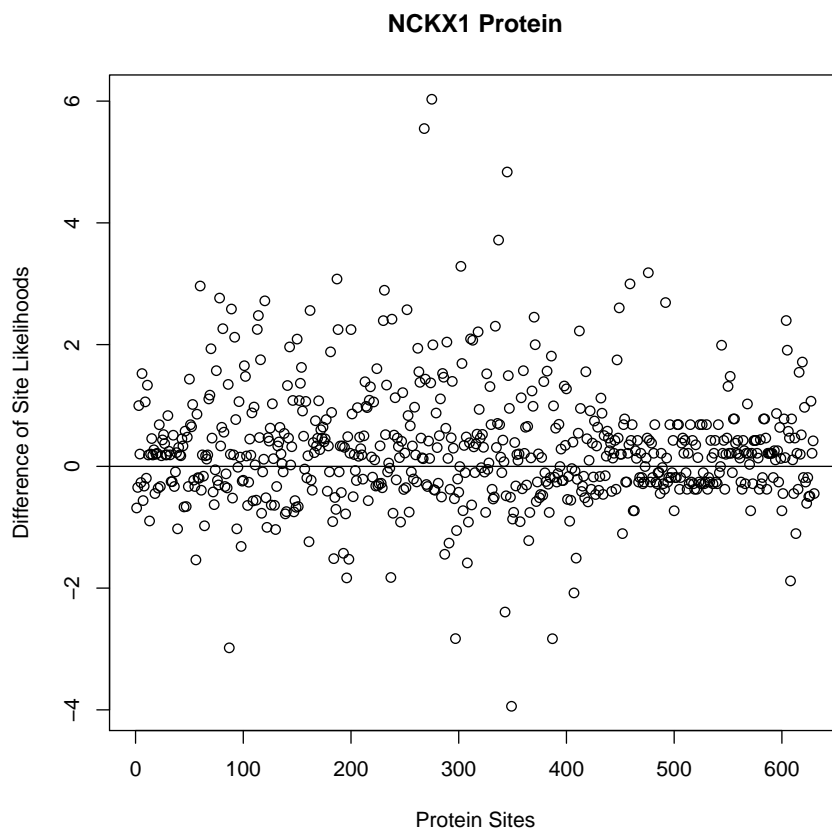


Figure 3.4: In this graph the differences between the BLOSUM62 and the Euclidean site likelihoods for the NCKX1 protein is shown. On the x axis, the protein sites are shown, while on the y axis the differences between the likelihoods are depicted. On the one hand, positive differences indicate that the BLOSUM62 rate matrix is more reliable in the calculation of these specific site likelihoods. On the other hand, negative differences indicate that the Euclidean rate matrix is more reliable in the calculation of these specific site likelihoods.

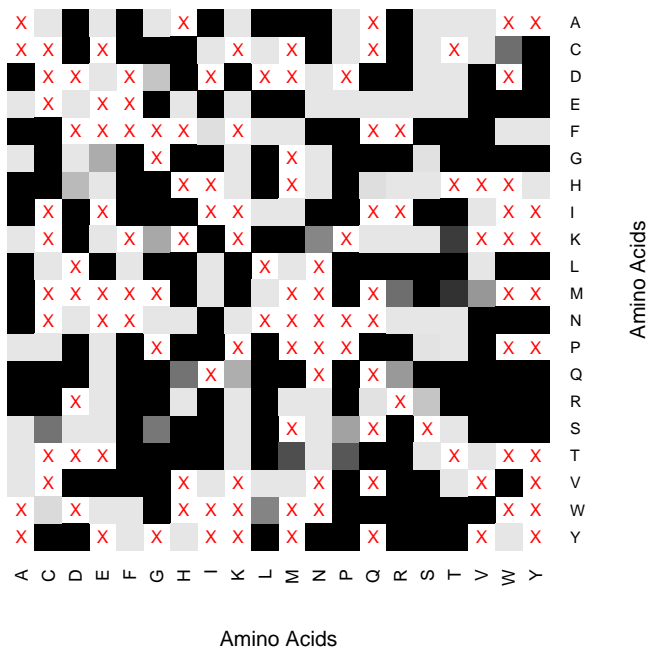


Figure 3.5: In this graph the ability of Euclidean rate matrix to calculate higher site likelihoods than the BLOSUM62 rate matrix is depicted. Black boxes show that for specific amino acid pairs Euclidean rate matrix always calculates a higher site likelihood. Also, the different shades of grey in the boxes depict the reduction of the ability of Euclidean rate matrix to calculate always higher site likelihoods for some amino acid pairs. Finally, white boxes with red crosses in the centre depict the amino acid pairs that were not found in the multiple alignments.

4 Discussion

In this project we checked if the evolutionary substitutions of amino acids can be determined from their neighbourhood tertiary structure. In order to achieve this, we used PrInS software to find how many times an amino acid is close to other in the space in the proteins of a protein family. Then, we converted the PrInS scoring matrices in distance and rate matrices in order to examine if these matrices can model evolution, via evolutionary tests.

4.1 PrInS Results Keep up with the Observed Amino Acid Substitutions

The results of Pearson Correlation Coefficient were a first indication that PrInS distance matrices can model evolution. In other words, the substitutions that happened in multiple alignments and are found on consensus matrix are kept up with the similar "behaviour" of amino acids. In distance matrices, the lowest distance between two amino acids denotes high "behavioural" similarity of these amino acids (tend to have similar neighbours). More specifically, from the three distance matrices that were correlated with the consensus matrix, only five amino acids were uncorrelated in at least two correlations. These amino acids were proline, arginine, serine, tryptophan and tyrosine. From consensus matrix is obvious that there are a few substitutions that involve tryptophan and tyrosine, while there are a lot of substitutions involving proline, arginine and serine. In contrast from the distance matrices is clear that in the columns of tryptophan and tyrosine there are small distance values, while in the proline, arginine and serine columns there are large values. These two facts can explain the dissociation of these amino acids in these two kinds of matrices, because a big number of substitutions is associated with small distance values for an amino acid. Moreover, in the correlation between the consensus and the Pearson matrices, uncorrelated amino acids were observed that had not been observed in the other correlations. It has to be mentioned that in other correlations, these amino acids were not strongly correlated. A possible explanation for this is that the Pearson matrix has lower values than the Euclidean and Manhattan matrices, which are not so close to the amino acid substitution counts that are observed in the consensus matrix.

In addition, the Common Substitution Tool of NCBI Amino Acid Explorer strengthens the previous observation [13]. The amino acid pairs with the lower distance values in the distance matrices and especially in Euclidean are found to substitute each other more easily. This was proved by using the Common Substitution Tool and giving as input an amino acid. The result of this analysis returned a list of the amino acids from the most common to the the less common substitution for the amino acid we put as input. Of course, there are some differences in the order of the amino acids between the Common Substitution Tool and the distance matrices (e.g Euclidean). These differences in order are possibly based on the fact that the generation of the Euclidean distance matrix is based only on the structural information of a protein family (alpha helical membrane proteins), while the Common Substitution Tool is based on BLOSUM62 substitution matrix that was created from alignments from multiple protein families. This possibly means that some specific amino acid substitutions are observed more frequently in the alpha helical membrane proteins, which are not found so often in the other protein families. These substitutions may play a crucial role in the distinct function of the alpha helical membrane proteins. After these outcomes we were able to check the distance matrices in other evolutionary tests in order to evaluate further the ability of PrInS to model evolution.

4.2 Similar Scoring Results and Functional Clustering of Proteins

After the evaluation of the ability of PrInS to predict amino acid substitutions, the scoring of multiple alignments with the several matrices was followed. The matrices that were used in the scoring of multiple alignments were distance, genetic and model substitutions matrices (BLOSUM62 and PAM120) [7], [8]. This was the second step in our analysis. Generally, the most of the proteins did not significantly differ from the human homologues, which are found in the first column of all the heatmaps. Obviously, homologous proteins in evolutionary distant species from human differ more than the homologues in more related species.

The phenomenon of protein differences was more obvious in Euclidean heatmap and the proteins in this cluster were identified and their functions were checked. It is possible that these sequence differences in the proteins of evolutionary distant species from human are related with the function of these proteins. For this reason, the functions of these proteins was found by using the gProfiler [14]. The proteins in the discrete protein cluster play a crucial role in the intestinal absorption of phytosterol and in cholesterol and lipid transportation, while the proteins of the other clusters are responsible for the homeostatic and transducer/receptor mechanisms of the cells. In addition, the proteins of the distinct protein cluster in the Euclidean heatmap, were found dispersed in the model substitution heatmaps (BLOSUM62 and PAM120), while they were slightly more

clustered in the other heatmaps (Genetic, Manhattan and Pearson), which are found in Supplementary Figures 6.3a- 6.3c [7], [8]. This partially proves the common origin of the distance matrices.

4.3 Similar Likelihoods by Using Alignment and Structural Rate Matrices

At this part of the analysis the overall likelihoods and the site likelihoods of the proteins were calculated by using common phylogenetic tree, multiple alignments and equilibrium frequencies and different rate matrices. As it was mentioned before the origins of these rate matrices vary. The result of this analysis showed that for 170 out of 223 proteins the overall likelihoods of the proteins were higher by using the BLOSUM62 and PAM120 rate matrices, while for the rest of proteins the overall likelihoods were higher by using the PrInS rate matrices and mainly with the Euclidean rate matrix [7], [8]. In addition, the calculation of the likelihoods by using the rate matrix that was made from the genetic code matrix was not so accurate. It is expected that for most of the alignments BLOSUM62 and PAM120 rate matrices would result in better likelihoods since their generation was based on alignments. In contrast, our approach is solely based on the local tertiary structures of proteins and is unaware of the alignments. Although it has to be mentioned that in both of these cases the values of overall likelihoods did not significantly differ, which means that our rate matrices can be used in order to calculate the likelihoods of multiple alignments. These results support that rate matrices that derived from protein structural data can be used in the calculations of likelihoods. Consequently, we can assume that protein evolution is affected or driven by the local tertiary structure of the proteins.

The site likelihoods of the proteins were calculated and the results of these calculations keep up with the previous outcomes. For some protein sites the BLOSUM62 and PAM120 rate matrices have higher likelihoods, while for other protein sites the site likelihoods were higher with the PrInS rate matrices [7], [8]. The ability of our rate matrices and especially the Euclidean rate matrix to calculate always higher site likelihoods for specific amino acid substitutions is an interesting result of our analysis. For example, the leucine-alanine pair has always higher site likelihood with the Euclidean rate matrices, such as the proline-glutamine pair. In contrast, amino acid pairs that have usually higher site likelihoods with the BLOSUM62 rate matrix and less often with the Euclidean rate matrix are the histidine-tyrosine and the phenylalanine-tryptophan pairs. These results are not always correlate with the favourable or the unfavourable amino acid substitutions results that were found in the previous steps of this research (Consensus Matrix, Distance/Genetic Code Matrices and Common Substitution Tool). Euclidean rate matrix calculate higher site likelihoods for amino acid pairs independent of the substitution patterns of the amino acids. Also, some amino acid pairs were not found

in the multiple alignments of alpha helical membrane proteins. This result it has to be evaluated further in the future in other protein families in order to gain a greater insight in the significance of this outcome. In addition, different areas with many positive or negative differences of site likelihoods can be observed on the proteins. These areas may have a key role in the function of protein. This result has to be investigated further in the future. Summarizing PrInS rate matrices can be characterised as a valuable tool in order to calculate the overall and site likelihoods of multiple alignments.

4.4 Significance of the Results

Our results clearly denote that protein evolution can be described by using protein structure data. The first indication, which support this statement is the correlation between the consensus matrix and the distance/genetic code matrices. In all the matrices, the most amino acids show similar substitution preferences. In the consensus matrix, the substitutions were observed in multiple alignments, while in the the distance matrices the substitutions can be predicted by using protein structural data and more specifically amino acid neighbourhoods in the space. However, all the amino acids were correlated in the correlation between the consensus and the genetic matrices. In this project, genetic matrix was used as control and was generated by calculating the minimum number of base changes required to convert an amino acid to another amino acid. Genetic code matrix was the simplest way to show the amino acid substitutions and was used in all the analyses. Moreover, the most common substitutions of amino acids were found by using the Common Substitution Tool of NCBI Amino Acid Explorer [13]. Then it was checked if these substitutions can be predicted by using the distance and genetic matrices. The results showed that distance and genetic matrices can predict these substitutions, but sometimes with a slightly different order, as it was explained before. This may be based on the fact that we used only a protein family, while Common Substitution Tool of NCBI Amino Acid Explorer is based on BLOSUM62 substitution matrix [13], [7]. Also, the results of scoring of the multiple alignments, in order to find sequence differences, with substitution, distance and genetic code matrices did not significantly different. Also, protein clusters were made at the visualization of these results using heatmaps. It was shown that proteins in different clusters have distinct functions, which has to be checked further in the future. In addition, the results of the calculations of overall and site likelihoods did not significantly differ by using the rate substitution or the rate distance matrices. Especially, the rate matrix that was made from the Euclidean distance matrix was able to calculate in similar fashion with the rate substitution matrices (BLOSUM62 and PAM120) the likelihoods [7], [8]. At this point it has to be mentioned that we used only structural data in order to produce all the matrices and not sequence and physicochemical data.

4.5 Future Perspectives and Limitations

One of the future perspectives of this project is to apply this procedure to more protein families in order to evaluate further the ability of PrInS to model evolution. In this case, it can be checked further if the distance and genetic code matrices can detect distinct protein functions as it was happened with the alpha helical membrane proteins via the protein clusters in the Euclidean distance matrix heatmap. Also, it would be interesting to apply this procedure to a group of proteins from several different protein families. This could be useful in order to check if there is a mutual path to determine protein evolution from protein structural data without depending on protein families. A limitation we faced in this project was that we could not find trustworthy tools in order to associate the protein site likelihoods, for example these which were higher with our rate matrices, with the tertiary structure of the proteins. We wanted to find association between these two protein characteristics in order to find the distribution of these sites on the tertiary structure of the proteins. It is possible that these sites can be found close to functionally important residues of proteins and our model matrices would play a crucial role in order to locate them.

5 Conclusion

In this project we have shown that evolutionary substitutions of amino acids can be determined from their neighbourhood tertiary structure. This was possible by statistically describing the neighbourhood of amino acids in the structures of proteins of a specific protein family. The results of this research showed that it is possible to consider the local structure of proteins when studying their evolution.

References

- [1] C. L. Worth, S. Gong, and T. L. Blundell, "Structural and functional constraints in the evolution of protein families", *Nature Reviews Molecular Cell Biology*, December 2, 2009.
- [2] M. Egli, "Diffraction techniques in structural biology", in *Current Protocols in Nucleic Acid Chemistry*, S. L. Beaucage, D. E. Bergstrom, P. Herdewijn, and A. Matsuda, Eds., DOI: 10.1002/0471142700.nc0713s41, Hoboken, NJ, USA: John Wiley & Sons, Inc., Jun. 2010, pp. 7.13.1–7.13.35.
- [3] A. Nath Jha, S. Vishveshwara, and J. R. Banavar, "Amino acid interaction preferences in helical membrane proteins", *Protein Engineering Design and Selection*, vol. 24, no. 8, pp. 579–588, August 1, 2011.
- [4] J. Siltberg-Liberles, J. A. Grahnen, and D. A. Liberles, "The evolution of protein structures and structural ensembles under functional constraint", *Genes*, vol. 2, no. 4, pp. 748–762, October 28, 2011.
- [5] L. Hatton and G. Warr, "Protein structure and evolution: Are they constrained globally by a principle derived from information theory?", *PLOS one*, vol. 10, no. 5, e0125663, 2015.
- [6] I.-G. Choi and S.-H. Kim, "Evolution of protein structural classes and protein sequence families", *Proceedings of the National Academy of Sciences*, vol. 103, no. 38, pp. 14 056–14 061, 2006.
- [7] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks", *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.
- [8] M. Dayhoff, R. Schwartz, and B. Orcutt, "22 a model of evolutionary change in proteins", in *Atlas of protein sequence and structure*, vol. 5, National Biomedical Research Foundation Silver Spring, MD, 1978, pp. 345–352.
- [9] The UniProt Consortium, "UniProt: The universal protein knowledgebase", *Nucleic Acids Research*, vol. 45, pp. D158–D169, D1 January 4, 2017.
- [10] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC", *Genome research*, vol. 12, no. 6, pp. 996–1006, 2002.

-
- [11] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The protein data bank", *The FEBS Journal*, vol. 80, no. 2, pp. 319–324, 1977.
- [12] U. Mudunuri, A. Che, M. Yi, and R. M. Stephens, "bioDBnet: The biological database network", *Bioinformatics*, vol. 25, no. 4, pp. 555–556, February 15, 2009.
- [13] B. Bulka, S. J. Freeland, *et al.*, "An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices", *BMC bioinformatics*, vol. 7, no. 1, p. 329, 2006.
- [14] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo, "G:profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments", *Nucleic Acids Research*, vol. 35, W193–W200, suppl_2 Jul. 2007.
- [15] C. J. McKiernan and M. Friedlander, "The retinal rod $\text{Na}^+/\text{Ca}^{2+}$, K^+ exchanger contains a noncleaved signal sequence required for translocation of the n terminus", *Journal of Biological Chemistry*, vol. 274, no. 53, pp. 38 177–38 182, 1999.

6 Supplementary Figures

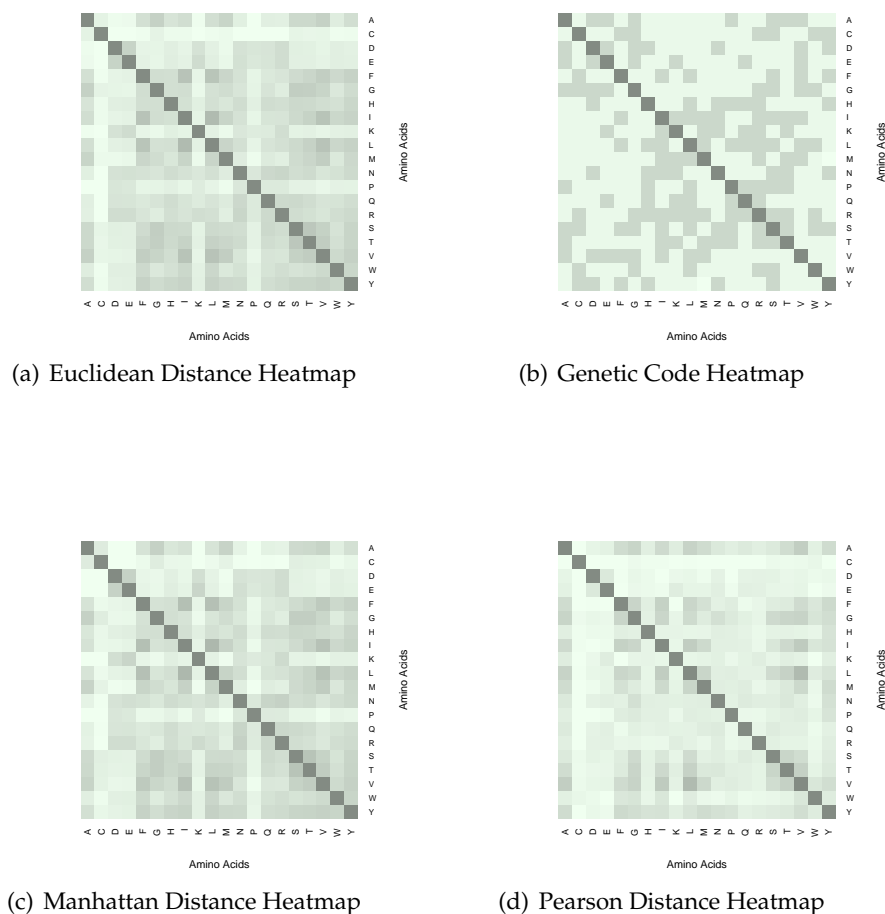


Figure 6.1: Heatmaps that depict (a) the Euclidean Distance Matrix, (b) the Genetic Code Matrix, (c) the Manhattan Distance Matrix and (d) the Pearson Distance Matrix. (a) In this figure, such as in figures 6.1c and 6.1d (Distance Matrices) the lower distance values are depicted with dark grey, while the larger distance values with the with white. Again the lower distance values in amino acid pairs denote "behavioural" similarity. (b) In the genetic code heatmap, the fewer base changes are depicted with the grey colour, while the many base changes with white. Finally, this heatmap shows how many base changes have to be occurred in order to change an amino acid into another. (c-d) As it is obvious, there are fewer behavioural amino acid similarities in the Pearson heatmap than in Manhattan distance heatmap.

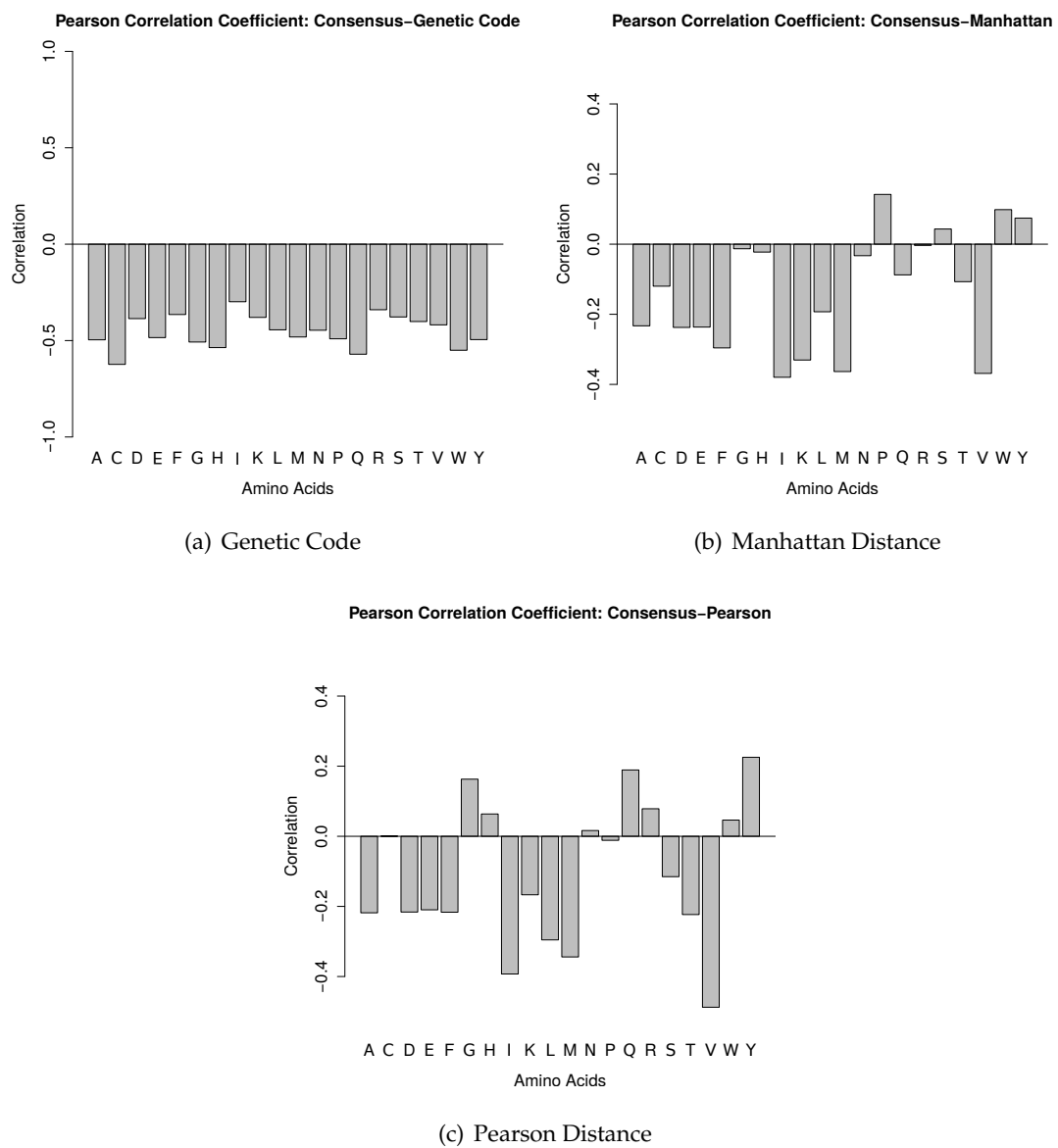


Figure 6.2: Pearson Correlation Coefficient between the consensus and (a) Genetic Code Matrix, (b) Manhattan Distance Matrix and (c) Pearson Distance Matrix. The correlation is represented with the negative values, while the disassociation with positive values, due to the different nature of consensus and genetic code/distance matrices.

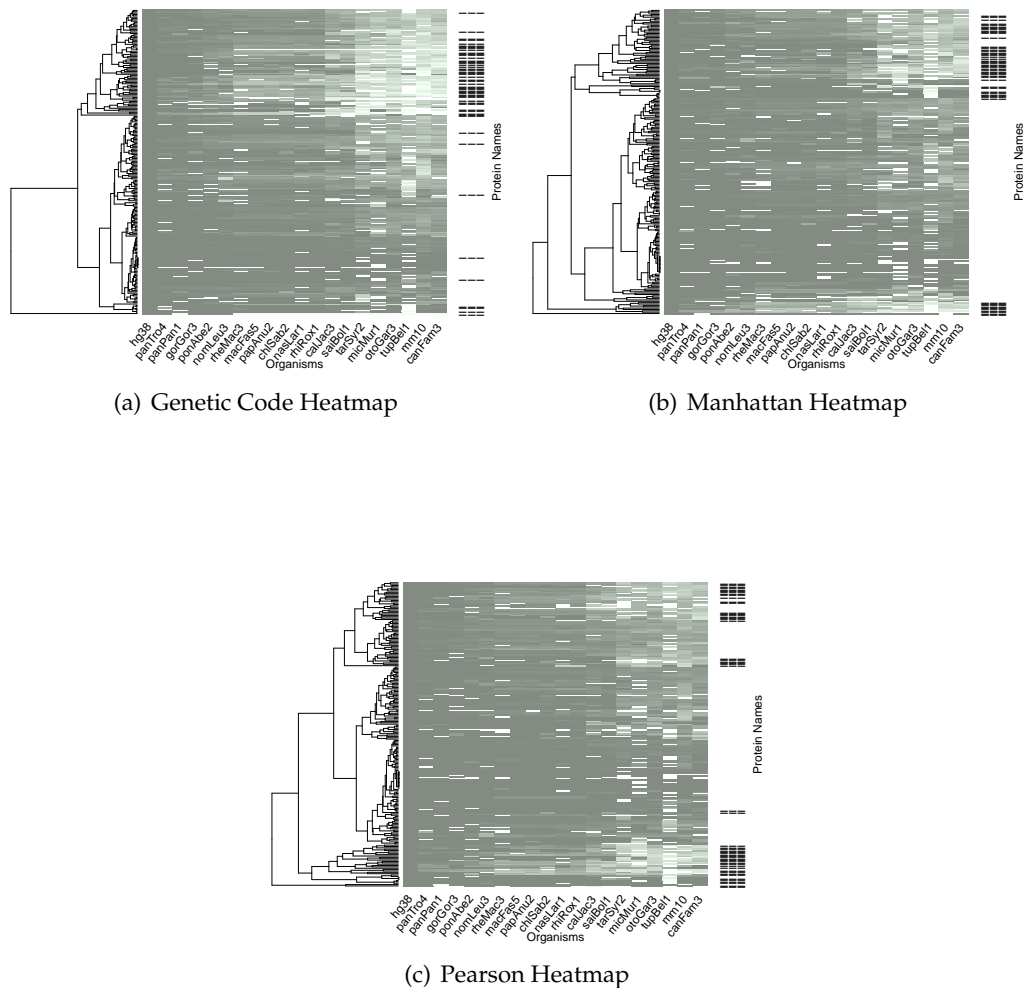


Figure 6.3: Heatmaps that depict the scoring of multiple alignments with the (a) Genetic Code Matrix, (b) Manhattan Distance Matrix and (c) Pearson Distance Matrix. In this figure there are also shown the hierarchical clustering of proteins and the species, whose proteins were scored. Human is found in the left end of the heatmaps. The other species were put in evolutionary order. Finally, the proteins that were found on the protein cluster in Euclidean heatmap (Figure 3.3a in the main part of the thesis) are represented in these heatmaps with three dashes. In these heatmaps the clustering of the proteins that were found in Euclidean heatmap is more obvious than in BLOSUM62 and PAM120 heatmaps (Figure 3.3b and 3.3c in the main part of the thesis).

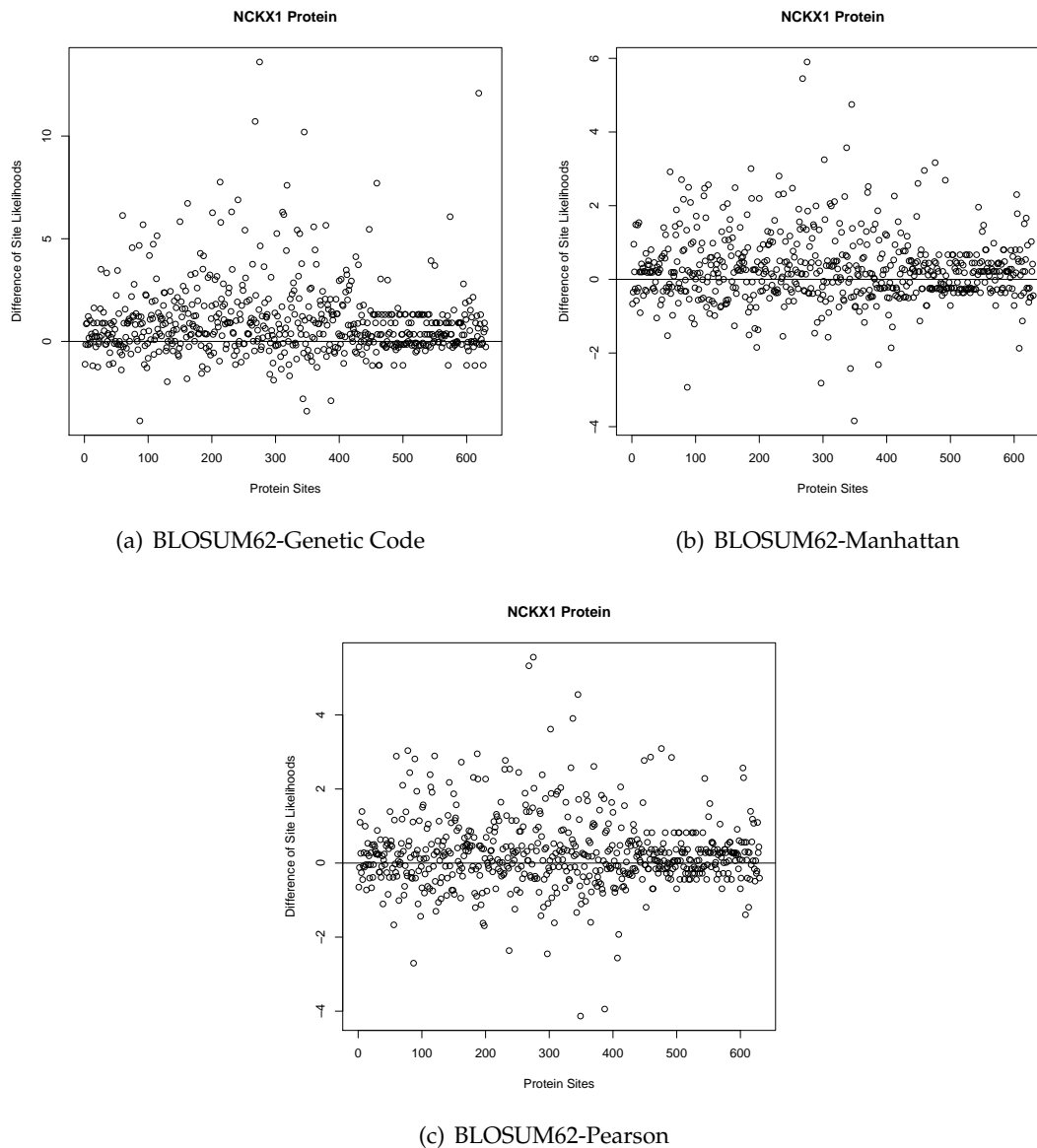


Figure 6.4: In these graphs the differences between the BLOSUM62 and the (a) Genetic , (b) Manhattan and the (c) Pearson Site Likelihoods for the NCKX1 protein are shown. On the x axis, the protein sites are shown, while on the y axis the differences between the likelihoods are depicted. On the one hand, positive differences indicate that the BLOSUM62 rate matrix is more reliable in the calculation of these specific site likelihoods. On the other hand, negative differences indicate that the (b) Genetic Code, (c) Manhattan and (d) Pearson rate matrices more reliable in the calculation of these specific site likelihoods.