Received: 9 September 2020 | Revised: 27 May 2021 | Accepted: 11 July 2021

DOI: 10.1002/prot.26178

RESEARCH ARTICLE

PROTEINS WILEY

Evolutionary models of amino acid substitutions based on the tertiary structure of their neighborhoods

Elias Primetis[1,2] | Spyridon Chavlis[2] | Pavlos Pavlidis[3]

[1]Department of Biology, University of Crete, Heraklion, Greece

[2]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Hellas, Heraklion, Greece

[3]Institute of Computer Science, Foundation for Research and Technology, Hellas, Heraklion, Greece

Correspondence
Pavlos Pavlidis, Institute of Computer Science, Foundation for Research and Technology, Hellas, Heraklion, Greece.
Email: pavlidis@ics.forth.gr

Abstract

Intra-protein residual vicinities depend on the involved amino acids. Energetically favorable vicinities (or interactions) have been preserved during evolution, while unfavorable vicinities have been eliminated. We describe, statistically, the interactions between amino acids using resolved protein structures. Based on the frequency of amino acid interactions, we have devised an amino acid substitution model that implements the following idea: amino acids that have similar neighbors in the protein tertiary structure can replace each other, while substitution is more difficult between amino acids that prefer different spatial neighbors. Using known tertiary structures for α-helical membrane (HM) proteins, we build evolutionary substitution matrices. We constructed maximum likelihood phylogenies using our amino acid substitution matrices and compared them to widely-used methods. Our results suggest that amino acid substitutions are associated with the spatial neighborhoods of amino acid residuals, providing, therefore, insights into the amino acid substitution process.

KEYWORDS
amino acid substitution matrix, protein evolution

1 | INTRODUCTION

Structure and functionality of a protein are closely related.[1] Its amino acid sequence as well as cofactors, ligands, and other parts of the same or other proteins form a complex network of interactions that is the basis of the unique physicochemical properties of each protein related to its function.[1] During the last decade, a multitude of tertiary protein structures have been determined by using techniques such as crystallography, NMR, electron microscopy, and hybrid methods.[2] Consequently, the number of the available tertiary structures in the RCSB-PDB database[3] amino acid is rapidly increasing.

Computational methods facilitate structural, functional, and evolutionary characterization of the proteins.[4] Protein function is tightly linked to its tertiary structure, and consequently to the vicinities, or interactions, amino acid residues have formed. For example, globular and membrane proteins are characterized by totally different physicochemical environments. Membrane proteins show a limited interaction with water molecules, and they are able to interact with the lipid bilayer. Thus, their transmembrane region adopts a single type of secondary structure—either a helix or a beta-sheet. Largely, the secondary structure is defined by non-neighboring (on the sequence level) amino acid residue interactions.[4]

Recently, we developed PrInS (freely available from http://pop-gen.eu/wordpress/software/prins-protein-residues-interaction-statistics (**Pr**otein **In**teraction **S**tatistics; Pavlidis et al. unpublished) an open-source software to score proteins based on the frequency of their residue interactions. PrInS uses protein structures stored in Protein Data Bank (PDB) to construct a statistical model of intraprotein amino acid residue interactions for a certain class of proteins (e.g., membrane proteins). PrInS scores every amino acid *a* proportionally to the number of "unexpected residues" that interact with it. The term "unexpected residues" means residues that they are rarely found to be in the vicinity of *a* if we consider all protein structures of a given dataset. Therefore, PrInS is able to pinpoint residues characterized by a large number of "less frequent" interactions, and therefore they may represent functional areas of the proteins or targets of natural selection based on the following assumption: even though a large number of unlikely interactions characterizes these amino acids, nature has preserved them.

Here, we use PrInS to describe statistically the intra-protein amino acid interactions and consequently to construct an amino acid substitution matrix endowed with the principle that amino acids with similar (residual) neighborhoods can substitute each other during evolution.

Protein evolution comprises two major principles. The first principle suggests that protein structure is more conserved than the sequence.[5] The second principle suggests that the physicochemical properties of amino acids constrain the structure, the function, and the evolution of proteins.[6] Protein evolutionary rate is strongly correlated with fractional residue burial.[5] This is due to the fact that the core of a protein is mostly formed by buried residues, which often play a crucial role in the stability of the folded structure.[7] The three-dimensional structure of the protein determines its evolutionary rate, since most mutations in the core of a protein tend to destabilize the protein.

Within protein families the backbone changes are infrequent, thus, preserving the folding properties over relatively long evolutionary distances, while substitutions are found often at the side chains. For example, for 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.[5]

In addition, the secondary structural elements of a protein evolve at different rates. Beta sheets evolve more slowly than helical regions and random coils evolve the fastest.[5] Secondary structure changes may eventually occur due to varying helix/sheet propensity. Some of these changes in secondary structural composition may be evolutionary neutral, whereas some structural transition may involve negative or positive selection. In the latter case, a new mutationally accessible fold may enable the development of a new favorable function that was not possible within the previous fold.[5]

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 destabilizing effect of the nonsynonymous mutations, which are often removed from the populations as a result of negative selection. The protein structure is important because it acts as a scaffold for properly orientating functional residues, such as a binding interface and a catalytic residue. As a result, the selective pressure for particular sequences (and not structures) over longer evolutionary periods is decreased, generating a neutral network of sequences interconnected via mutational changes.[5]

Choi and Kim,[8] based on the most common structural ancestor (CSA), showed that not all present-day proteins evolved from one single set of proteins in the last common ancestral organism, but new common ancestral proteins 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.[8]

Leelananda et al.[9] emphasized the need of structural alignments in understanding the nature of amino acid substitutions. It is known that there are some amino acid substitutions that occur more frequently in some topologies than in others and these are usually substitutions that do not affect the function of these proteins. Different protein topologies exhibit different amino acid substitution statistics.[9] Therefore, we created a protein family based substitution matrix derived from the tertiary structure of proteins and their amino acid neighborhoods. The results of our research can be used as an extra layer of structural information based on the proximity of amino acids in proteins of the same family. In Leelananda et al.[9] study, CATH structures were used to extract topological information about the proteins, while we used the PrInS algorithm, which allows us to statistically characterize interactions of proximal amino acids by using PDB files. The main advantage of our study is that input data come from a large database (PDB) with publicly available protein structures, even from the same organism.

In this study, we used the scoring matrices that are obtained from the PrInS algorithm to examine the evolution of proteins and we focused on the evolution of α-helical membrane proteins. The main hypothesis is that protein evolution is related to the three dimensional neighborhoods of amino acids. Specifically, amino acids that have similar amino acid neighborhoods can substitute each other during evolution.

2 | MATERIALS AND METHODS

2.1 | Dataset retrieval and name conversion

In eukaryotes, α-helical proteins exist mostly in the plasma membrane or sometimes in the outer cell membrane. In prokaryotes, they are present in their inner membranes. We used 82 α-helical membrane proteins, which were also scrutinized previously by Nath Jha et al.[4] to describe the statistical properties of amino acid interactions within α-helical membrane proteins.

We used α-helical membrane proteins as Nath Jha et al.[4] because α-helical membrane proteins are well-studied and they form a rather homogeneous set of proteins with restrictions posed by the physicochemical environment in which they are located and is drastically different from other protein families. α-helical proteins interact mainly with the lipid bilayer and with a limited number of water molecules inside the membrane. The transmembrane region of the membrane proteins generally adopts a single type of secondary structure—either a helix or a beta sheet. Thus, interactions of the set of 20 amino acids differ in their propensities. For example, even though Cys–Cys has the strongest pair propensity in globular proteins, it is extremely rare in membrane proteins. Furthermore, Nath Jha et al.[4] offered a deep understanding on the amino acid interaction properties of the α-helical membrane proteins and thus, their structures were used as a proof-of-concept for the initial development of the PrInS algorithm. First, for each of the protein in the dataset, we downloaded multiple sequence alignments of homologous protein sequences from the UCSC Genome Browser[10] for 16 primate and 3 nonprimate mammalian species. Second, three-dimensional structures were downloaded from the Protein Data Bank (PDB) database.

In addition to the α-helical membrane proteins, we tested the presented methodology in a set of 2298 mitochondrial protein

structures to generate a substitution matrix and compare it with the widely-used BLOSUM62 amino acid substitution matrix.

2.2 | The PrInS software

We applied PrInS on the three dimensional structures downloaded from PDB. PrInS constructs a scoring matrix M as follows: If the tertiary structure of a protein is represented by P and the total number of amino acid residues is l, then residues $1 \le k, m \le l$ interact if and only if:

$$A_{km} = \begin{cases} 1, & \text{if } d\left(C_\alpha^k - C_\alpha^m\right) \le 6.5\,\text{Å} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $d\left(C_\alpha^k - C_\alpha^m\right)$ denotes the distance between the C_α atoms of the k and m amino acids. In other words, the kth and the mth residues of a protein interact if and only if the Euclidean distance between their C_α atoms is less than 6.5 Å and they are not located on adjacent positions on the amino acid chain. The distance of 6.5 Å was used as a hard cut-off value in our study. However, it represents a widely accepted cut-off value to define amino acid interactions. For example, Jiao et al.[11] used also the same cut-off value to construct a network of amino acids on the protein backbone and explore its properties. If non-C_α atoms would be involved in the definition of the neighborhood, then a different distance value should be used as a cut-off. For example, Nath Jha et al.[12] defines the cut-off to be only 4.5 Å if any (non-hydrogen) atoms are employed in the distance definition. Clearly, other cut-off values will affect the results since they will re-define neighborhoods. However, based on the aforementioned studies the choice of 6.5 Å appears to be a less arbitrary decision.

2.2.1 | Defining the environment of each amino acid

According to Nath Jha et al.,[4] amino acids of a HM protein can be classified in three environments: The first environment comprises amino acids at different distances from the lipid bilayer. The second environment classifies the helices (and thus, their amino acids) on the basis of their interhelical interactions inside the membrane. Finally, the third environment classifies amino acid residues based on the number of interactions they make. In our study, we have used the third type of environments (residue-contact-based interaction) based on the results from Nath Jha et al.,[4] who demonstrated that the residue-contact-based environment description of residue interaction is more accurate for the α-helical proteins they studied. Thus, every amino acid residue P_i in protein P can be assigned to the pair (A, K), where A indexes the amino acid type of the kth residue (e.g., Alanine) and K represent the environment of the kth residue. Based on the results of Nath Jha et al.,[4] we used the number of noncovalent contacts each amino acid makes to define its environment. Thus, environment I comprises all amino acids with 1–5 contacts, environment II, amino acids with exactly 6 contacts, and environment III amino acids with more than 6 contacts. In our analysis, we employed solely noncovalent bonds, and we ignored the covalent bonds that form the amino acid backbone of the protein chain or they are the bonds found between the atoms of one residue. We neglect the covalent bonds because we do not focus on the protein chain (a peptide bond is an amide type of covalent chemical bond linking two consecutive alpha-amino acids defining the primary amino acid structure) but on the amino acid interactions that are present in a protein beyond the protein chain neighborhood. In other words, we focus on amino acids that are not direct neighbors along the protein chain. Furthermore, we should mention that we do not reject Cys–Cys bonds that are not direct neighbors along a protein chain. Cysteine is the sole amino acid whose side chain can form covalent bonds, yielding disulfide bridges with other cysteine side chains. If however, those cysteine amino acids are not consecutive amino acids, we consider them as a valid pair in our analysis.

Initially, PrInS was used to construct a matrix M, a 60×60 matrix (or equivalently nine 20×20 scoring matrices), that scores amino acid interactions in all environment pairs. For example, $M_{i,j}$, i, $j \le 20$ describe score interactions between amino acids of the environment I. In M the pairs of amino acids with the lower scores are those that interact frequently. In general, the interaction between amino acid A ($1 \le A \le 20$) and amino acid B ($1 \le B \le 20$) that belong to the environments K ($1 \le K \le 3$) and Q ($1 \le Q \le 3$), respectively, is given by:

$$M_{ij} = -ln\left(\frac{n_{A,K-B,Q}}{g \times (S_A/S) \times (S_B/S) \times E_{K,Q}}\right) \quad (2)$$

The coordinates i and j are given by $i = 20(K-1) + A$ and $j = 20(Q-1) + B$, respectively. The parameter $g = 2$ if the contacting pair comprises different amino acids in the same environment ($A \neq B$, $K = Q$), and $g = 1$, otherwise. S_A and S_B are the total number of amino acids A and B in the dataset, respectively. S is the total number of amino acids in the dataset and $E_{K,Q}$ equals to the number of interactions between environment K and Q. Finally, $n_{A,K-B,Q}$ is the number of interactions between amino acids A in environment K and amino acid B in environment Q. It is evident that $M_{i,j}$ assumes large (positive) values when the observed number of interactions $n_{A,K-B,Q}$, between A and B, is much lower than the interactions expected based on the total frequency of the amino acids A and B in the dataset and the number of interactions between the K and Q environments. In other words, the value of M_{ij} is large for rare interactions.

2.2.2 | Distance matrix

The M scoring matrix was split in nine 20×20 matrices and distance metrics (Euclidean, Manhattan, and Pearson Squared) between the elements of each matrix (amino acids) were calculated as follows: Let G_d represent the average 20×20 distance matrix for the distance metrics d ($d \in$ {Euclidean, Manhattan, Pearson Squared}). We name these matrices neighborhood Euclidean-distance based (*nEd*), *nMd*, and *nPd*, respectively, for Euclidean, Manhattan and Pearson Squared.

For comparison purposes, we created an additional 20 × 20 distance matrix based on the Hamming distance between a pair of codons showing the minimum number of nucleotide changes required to transform one amino acid to another. Smaller distance values indicate higher similarity between the respective amino acids. Distance matrices have been visualized as heatmaps (Figure S1a–d). Two amino acids that have similar neighborhoods, that is, similar amino acid neighbors in the tertiary structure will, thus, have a small distance (high similarity) between them. Consequently, according to our hypothesis, they will substitute each other during evolution at a higher rate.

2.2.3 | Substitution rate matrix

Each of the four distance matrices were transformed into rate matrices D_r by using a similar procedure as in Dayhoff et al.[13] In particular, the following procedure was followed:

1. Find the maximum value of the distance matrix D_g, D_{max}.
2. Subtract each matrix element from the D_{max} and divide by D_{max}. This will transform D_g to an amino acid "similarity" symmetric matrix, where 1 denotes maximum "similarity" between two amino acids.
3. Similar to the transition rate matrices used in phylogenetics analyses, one element of the matrix is defined as a reference. All other elements are scaled proportionally to the reference. Substitution (transition) rate matrices are used in phylogenetics to model the transition process of amino acids along a branch of the phylogenetic tree. Therefore, the number of substitutions is a function of both the substitution rate and the branch length. Substitution rate matrices provide information related to the rate of substitution during an infinitesimal amount of time. We have converted all matrices to transition rate matrices to be able to compare them to the matrices used in phylogenetics-related software. BLOSUM62 and PAM120 transition rate matrices were obtained from the widely-used RAxML github repository (https://github.com/stamatak/standard-RAxML)[14]
4. The diagonal of the matrix is defined as the negative sum of all other elements of the respective row. This is because the sum of the elements of each row in a transition rate matrix should be equal to 0.

If an amino acid pair is characterized by a large (relative) substitution rate, then they can substitute each other during evolution at a high rate. The substitution rate matrix that is derived with the aforementioned algorithm is symmetric.

2.2.4 | Evaluation of PrInS ability to predict amino acid substitutions

For the whole set of multiple alignments, a substitution 20 × 20 matrix S was created as follows: Let B the set of sites from all sequence alignments that contain only two amino acids. Let $|B|$ the number of such sites. Then, if B_i, $0 < i < |B|$ is such a site, that contains only amino acids a and b, $B[a][b]$ is incremented by one. In other words, a cell in the S matrix that corresponds to amino acids a and b, counts how many times we observed a site from the multiple sequence alignments with the amino acids a and b. The assumption here is that during evolution there was at least one substitutions between a and b for the specific site of the alignment that contains these two amino acids. Also, the more sites with a and b, the more frequent the substitution between them.

To evaluate the relation of our neighborhood-based substitution distance matrix *nEd* and the multiple sequence alignments, we calculated the Pearson correlation coefficients between each amino acid in S and *nEd*. Negative values of Pearson correlation coefficient values suggest that the amino acid substitutions in the alignments are concordant with the neighborhood-based substitution rates distances (smaller distances in *nEd* suggest greater neighborhood similarity).

Moreover, we manually analyzed the amino acid pairs with the lowest distance values in the Euclidean, Manhattan and Pearson distance matrices with the tool "Common Substitution Tool" in the "Amino Acid Explorer" (https://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi) of NCBI.[15] Common Substitution Tool sorts an amino acid list from the most common to the least common substitutions, given an amino acid as input, based on the BLOSUM62 substitution matrix.[16]

2.3 | Overall and site likelihoods

Likelihood is a function of the parameters of a statistical model for given data. Since substitution rate matrices are part of the model, we compared the per site and the overall likelihood values obtained by our substitution matrices and those obtained by using BLOSUM62 and PAM120. Even though a multitude of transition rate matrices are available to model amino acid substitution, we focused on BLOSUM62 and PAM120 since they are among the most widely-used amino acid substitution matrices. Both the phylogeny and the equilibrium frequencies of the amino acids were considered known in this analysis. Thus, by keeping the remaining parameters the same for both models, we aimed to assess which rate matrix can better explain the observed multiple alignments. As overall likelihood, we defined the total likelihood of the alignment, while a site likelihood is the likelihood of a single position in the amino acid alignment. The phylogenetic tree was retrieved from the UCSC Genome Browser,[10] and the equilibrium frequencies of amino acids were obtained from the equilibrium frequencies in the BLOSUM62 model. We used the same phylogenetic tree and equilibrium frequencies for all comparisons.

3 | RESULTS

3.1 | Evaluation of PrInS ability to predict amino acid substitutions

To evaluate the performance of our approach, we inspected the amino acid states present in each polymorphic site of each one of

the 223 downloaded multiple sequence alignments. For example, if a site in a multiple sequence alignment is the following vector: `I,I,I,L,L,L,I` then, this implies that during evolution there was a substitution from I to L or vice versa. Most of the sites in a multiple sequence alignment contain a maximum of two amino acids, implying a minimum of a single substitution during evolution. If more than two states are included (implying more than a single substitution), then we considered only the two most frequent amino acid states. Thus, by considering all pairs present in the multiple sequence alignment sites, we construct a matrix which indicates which substitutions have occurred more frequently during evolution. This "observed" substitution matrix counts the amino acid substitutions that occurred in the 223 downloaded multiple alignments using the two most frequent amino acid residues at each alignment site. For each amino acid in this matrix, we evaluated its correlation (Pearson correlation coefficient) with the same amino acid in the matrices we generated following the neighborhood approach presented in this study. Thus, if, for a given amino acid, both matrices suggest similar substitution preferences, then neighborhood-based substitution models are concordant with the observed substitutions. Since we used *distance matrix* in this analysis, negative correlation coefficient indicate concordance. Results suggest that there is concordance between the substitution matrix and our neighborhood-based distance matrices for most of the amino acids. Figure 1 shows the Pearson correlation coefficients for amino acids using the substitution matrix and the *nEd*. The correlation plots between the substitution matrix and other distances are illustrated in the Figure S2a–c.

As shown in Figure 1, most of the amino acids are concordant when we compare the Euclidean distance matrix and the substitution matrix.

In the Figure S2b (substitution vs. Manhattan distance), the discordant amino acids were the five out of six amino acids of Figure 1.

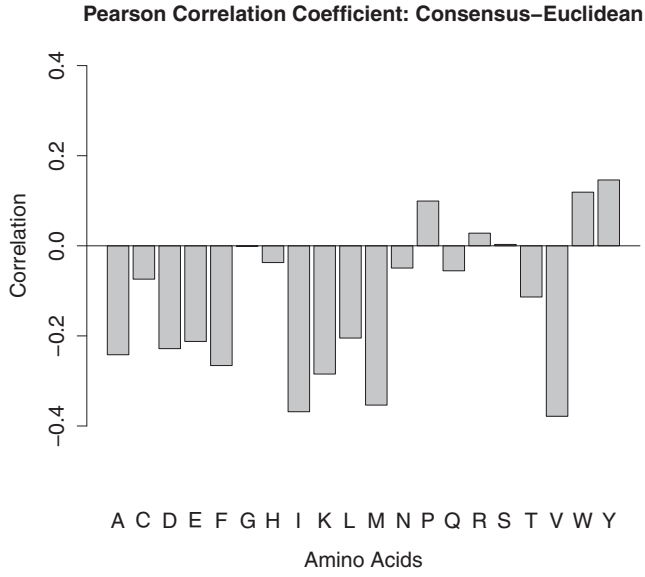

FIGURE 1 Pearson correlation coefficient between the substitution and *nEd*. Concordance is indicated with negative values, while discordance with positive values

Similar results are shown in Figure S2c, where the substitution matrix is compared against the squared Pearson distance matrix. Finally, there are no discordant amino acids in the correlation between the Consensus matrix and the Genetic Code matrix (Figure S2a).

Furthermore, the amino acid pairs with the lowest values in the distance matrices are at the top of the amino acid list that the Common Substitution Tool of NCBI Amino Acid Explorer reports. This means that amino acids with similar three dimensional neighborhoods tend to substitute each other. For example, in Euclidean distance matrix, the Leucine–Isoleucine pair has the lowest distance value and according to the substitution list of Amino Acid Explorer, Isoleucine is the most common substitution of Leucine. Figure 2 shows the first five most common amino acid substitutions for Leucine according to "Common Substitution Tool" (Figure 2A) and according to the Euclidean distance matrix (Figure 2B). Evidently, the Euclidean distance matrix can predict the most common substitutions of Leucine, but in slightly different order. For example, Common Substitution Tool reports the following order of amino acids: Isoleucine, Methionine, Valine, Phenylalanine, and Alanine. According to the Euclidean distance matrix, the order is Isoleucine, Phenylalanine, Valine, Tryptophan, and Tyrosine. 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. Neglecting the order of the first five substitutions, drawing five amino acids and observing three common is marginally significant (p-value = .07, right tail of hypergeometric distribution). Differences between two reports are possibly due to the fact that Euclidean distance matrix is solely based on the structural information of a protein family, while the Common Substitution Tool is based on BLOSUM62 substitution matrix that was created using alignments of multiple protein families.

According to the previous analyses, we can rely on the structural information of the proteins, which is given by the Euclidean distance matrix, to model amino acid substitutions similarly as using sequence data.

3.2 | Pairwise comparison of proteins in different species

Using the Euclidean-based distance matrix, we scored the differences of each protein sequence to its human homolog. Results are presented as a heatmap and are clustered hierarchically based on their distance from the human homolog (Figure 3). Darker gray tone in Figure 3 denotes high similarity between human and other species homologs, whereas lighter tones suggest lower similarity. As expected, proteins from species that are evolutionarily distant from humans are more dissimilar to human homologs. This is especially true for a group of proteins clustered together using the Euclidean distance-based substitution matrix (Figure 3 proteins labeled with three dashes). The same analysis was repeated by using the Genetic Code, Manhattan,

(A)

1-letter code	3-letter code	Chemistry	Potential H-bonds	Molecular Weight	Isoelectric Point	Hydrophobicity
L	Leu	CH_2 -C-C-	0	113	6.0	0.918
I	Ile	CH_2 -C-C-	0	113	6.0	1.000
M	Met	S -C-C-	0	131	5.7	0.811
V	Val	CH_2 -C-C-	0	99	6.0	0.923
F	Phe	CH_2 ⬡	0	147	5.5	0.951
A	Ala	CH_2 -C-C-	0	71	6.0	0.806

(B)

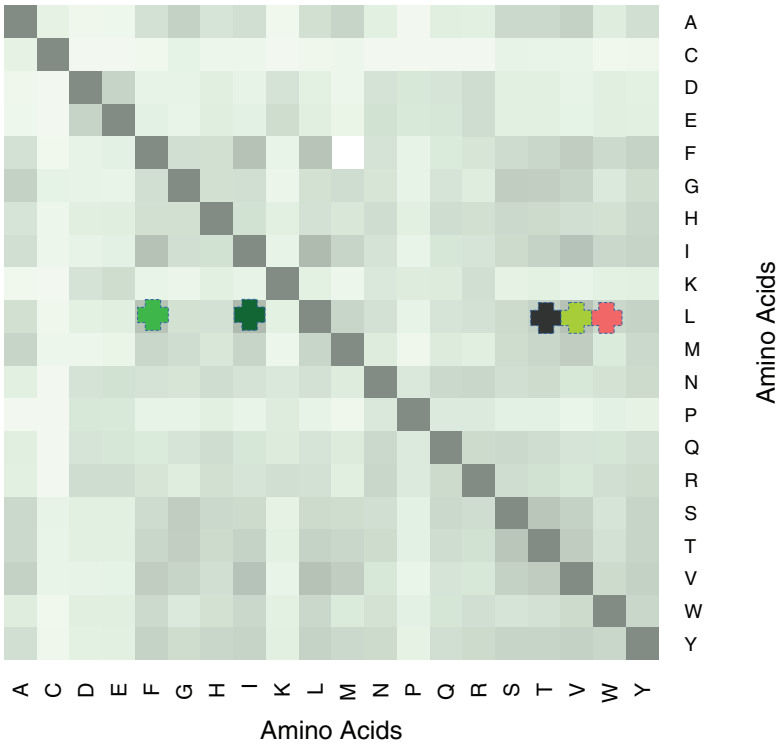

FIGURE 2 The first five most common substitutions of Leucine according to (A) Common Substitution Tool and (B) Euclidean Distance Matrix. The same color code is used in both figures. Dark green shows the first most common substitution, while the red shows the fifth most common substitution. Methionine and alanine are not found among the first five most common substitutions of Leucine in Euclidean distance matrix, while Tryptophan and Tyrosine are not found among the first most common substitutions in Common Substitution Tool

and Pearson distance matrices instead of Euclidean distance matrix. The results are shown in Figure S3a–c.

The group of proteins clustered together by the Euclidean-based matrix (marked in the lower part of Figure 3) was further scrutinized by Gene Ontology (GO) terms. We used gProfiler,[17] to obtain the related GO terms for this set of proteins. These proteins are characterized by a distinct function compared to proteins found in other clusters. On the one hand, the proteins that are located in this cluster play a crucial role in the intestinal absorption of phytosterol and in cholisterol and lipid transportation. On the other hand, the remaining proteins, which are not found in this discrete cluster, are responsible for the homeostatic and transducer/receptor mechanisms of the cells.

Hierarchical clustering of species based on the *nEd* distance matrix orders the species from human as expected, putting Chimpanzees, Gorilla, Orangutan, and other primates closer to human, while nonprimate mammals (e.g., mouse and dog) are located more distant (Figure 3). Thus, using the *nEd* distance matrix, the evolutionary relations between species behave as expected.

3.3 | Overall and site likelihoods

By converting PrInS distance matrices into rate matrices, it was possible to compute the overall likelihood and site likelihoods for every protein alignment. The resulting matrices from the *nEd*, *nMd*, and *nPSm* are named as *nEs*, *nMs*, and *nPSs*, respectively. For most of the proteins, 170 out of 223, the likelihoods computed by BLOSUM62 and PAM120 were better than our neighborhood Euclidean-based substitution matrices. This result is expected since BLOSUM62 and PAM120 substitution matrices were constructed based on alignment files, whereas the presented substitution matrices are alignment-unaware. For the remaining 54 protein alignments, our models and more specifically the Euclidean-based rate matrix resulted in higher likelihoods.

For every protein, we calculated the individual site likelihood for each amino acid site in the multiple sequence alignment using either the BLOSUM62 (or PAM120) or the *nEs* matrix. Results are illustrated in Figure 4) for the protein NCKX1, where the likelihood difference

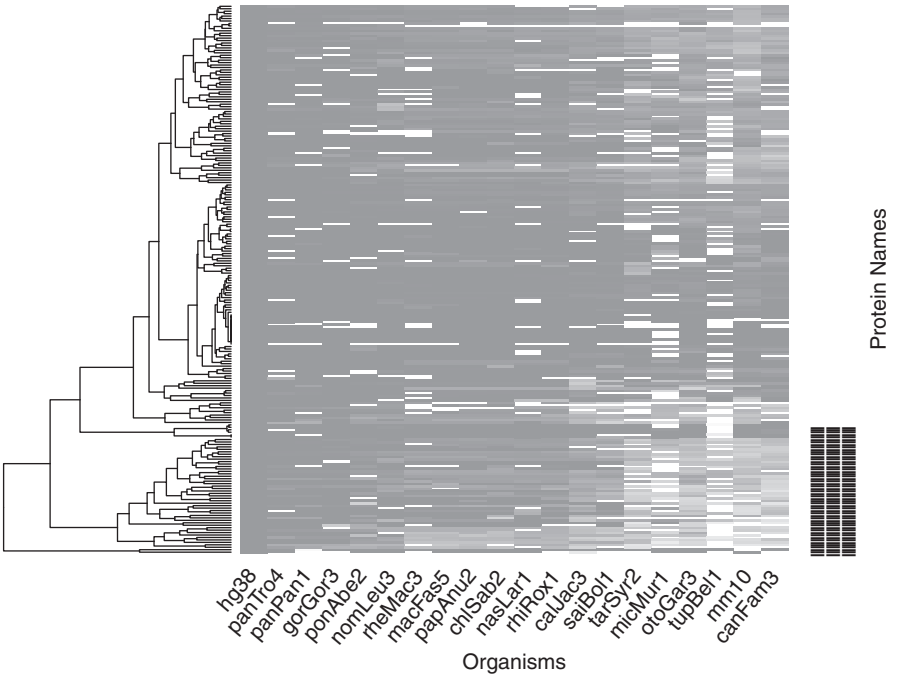

FIGURE 3 An illustration of the scoring of multiple alignments with the Euclidean distance matrix as a heatmap. Each protein (rows) from human (first column) is compared against the homologous proteins from another species (columns) using pairwise amino acid sequence alignments. We used only the alignment sites with amino acids *A* and *B*, where $A \neq B$ and $A \neq -$ and $B \neq -$. The total score (a cell in the heatmap) is calculated as the sum of all site scores. Proteins form two clusters based on the hierarchical tree on the left side of the figure. Proteins of the bottom cluster (also denoted with three dashes are further analyzed using Gene Ontology terms

between BLOSUM68 and *nEs* is plotted for every amino acid. Thus, negative values are related to sites where the likelihood is greater for the *nEs* matrix, whereas positive values show sites where BLOSUM62 results in greater likelihoods. For the protein NCKX1, 240 sites are scored with a higher likelihood using the Euclidean substitution matrix derived from PrInS (negative values), whereas 390 sites are scored with a higher likelihood using BLOSUM62 (positive values). The overall likelihood difference for this protein is positive (264.4) indicating that overall BLOSUM62 results in higher likelihood scores. NCKX1 was randomly selected to illustrate the site likelihood differences between the two approaches. It is a critical component of the visual transduction cascade, controlling the calcium concentration of outer segments during light and darkness.[18]

In the Figure 4a–c, we show the site likelihood differences between the BLOSUM62 rate matrix and (i) the genetic code rate matrix **4a**, (ii) the Manhattan rate matrix **4b**, and (iii) the Pearson rate matrix **4c**.

In calculating overall and site likelihoods, the sequence-based methods, such as BLOSUM62 and PAM120 rate matrices, are better than the structure-based methods (Euclidean matrix); however, this is expected because BLOSUM62 and PAM120 rate matrices are based on sequence alignments and here are used to score sequence alignments, whereas the structure-based methods we present are alignment-unaware.

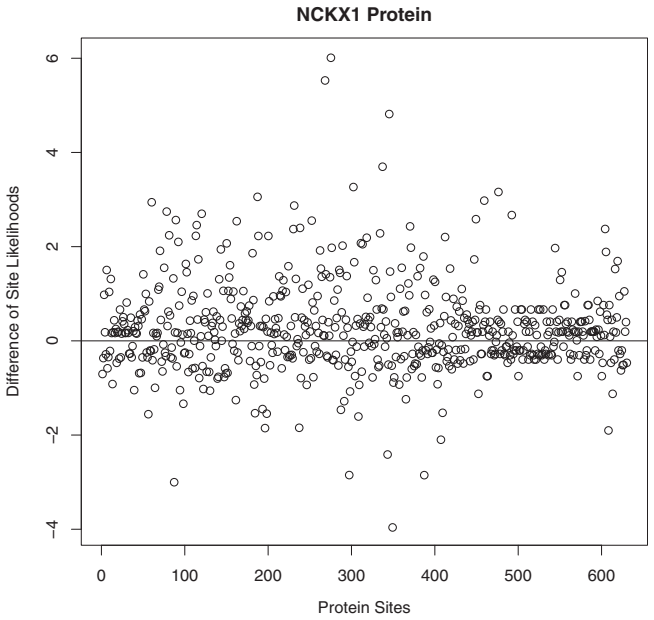

FIGURE 4 In this graph the differences between the BLOSUM62-based and the Euclidean-based likelihoods for each site of the NCKX1 protein is shown. On the *x* axis, the protein sites are shown, while on the *y* axis the differences between the likelihoods are depicted. Positive differences indicate higher likelihoods for the BLOSUM62 rate matrix, whereas negative differences indicate higher likelihoods for the Euclidean rate matrix

3.4 | Comparison between the average substitution likelihoods for different rate matrices

The 20 amino acids form $\binom{20}{2} = 190$ (unordered) pairs. For each of them, we have calculated the average differences between the BLOSUM62-based and the *nEs*-based likelihoods. For the calculations, we considered only the sites that consist of two amino acid states. Since all the parameters of the evolutionary model, but the substitution rate matrix are fixed (phylogenetic tree, equilibrium frequencies), then the matrix that results in the highest average likelihood for a certain amino acid pair describes the preferential model for this amino

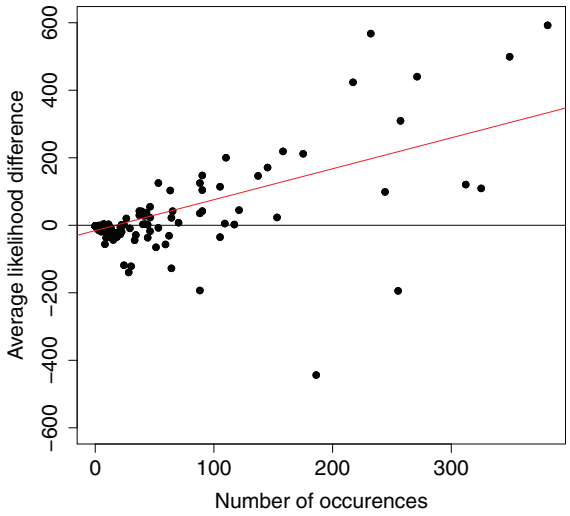

FIGURE 5 The average likelihood difference between the BLOSUM62-based and the Euclidean-based approaches that have been used to calculate likelihoods. As the figure indicates, for the majority of amino acid pairs the likelihood difference is close to 0, that is, both the Euclidean-based calculations and the BLOSUM62-based calculations result in similar outcomes. Positive values are fewer than negative values (see text); however, the magnitude is much greater for positive values than for negative, indicating that BLOSUM62 results in much greater likelihoods than the Euclidean-based approach. Furthermore, as the occurrence frequency of the amino acid increases the difference of the likelihood increases as well, suggesting that BLOSUM62 outperforms the Euclidean-based approach for the amino acid pairs that occur frequently, either within the same protein or in different proteins [Color figure can be viewed at wileyonlinelibrary.com]

acid pair. Surprisingly, even though most of the protein alignments are scored with a higher likelihood using BLOSUM62, 111 (out of 190) amino acid pairs are characterized by negative average likelihood difference. This indicates that the *nEs*-based likelihood is greater than the BLOSUM62-based likelihood. The average likelihood is greater for the BLOSUM62 for only 50 pairs, whereas it cannot be assessed for 29 pairs because no occurrences of these 29 pairs were present in the alignment datasets. To scrutinize further this result, we plotted the average likelihood difference as a function of the occurrence frequency of the amino acid pair in the alignment (Figure 5). In Figure 5, it is apparent that the average likelihood difference is positively correlated with the frequency of occurrences of the amino acid pairs ($r = 0.621$, r is the correlation coefficient, CI: [0.55, 0.67], CI denotes the 95% confidence intervals). In other words, the more frequent a substitution is in the alignments, the higher the difference of likelihoods, favoring the BLOSUM62 versus the *nEs*.

The heatmap in Figure 6 shows all amino acid pairs and the likelihood differences between the BLOSUM62 and the *nEs* matrix for all amino acid pairs. Cells denoted by a "−" are characterized by a higher likelihood for the Euclidean-based matrix, whereas a "+" denotes cells with higher BLOSUM62-based likelihood. The darker the color of the cell the higher the difference between the two likelihoods. Cells with an "o" illustrate pairs where no site with this pair of amino acids was found in any of the alignments. The greatest value for a "−" cell is represented for the "Arginine–Glycine" pair. This means that the *nEs* matrix is a preferential model for the specific amino acid pair. On the other hand, the pair with the highest "+" value is the "Isoleucine–Valine" pair.

In contrast to the overall and site likelihoods, the average substitution likelihoods are calculated better by using structural (Euclidean rate matrix) than sequence (BLOSUM62 and PAM120 matrices) based methods.

3.5 | Comparison to Grantham's and Sneath's distance matrices

Grantham's distance[19] between two amino acids depends on three properties: composition (defined as the atomic weight ratio of noncarbon elements in end groups or rings to carbons in the side chain), polarity, and molecular volume. Based on these three properties, distance $D_{G[i,j]}$, between amino acids i and j, is defined as:

$$D_{G[i,j]} = \left[\alpha\left(c_i - c_j\right)^2 + \beta\left(p_i - p_j\right)^2 + \gamma\left(v_i - v_j\right)^2\right]^{1/2} \quad (3)$$

where c is the composition, p the polarity, and v the molecular volume. The three components of the distance (c, p, v) are not independent. Thus, the constants α, β, γ serve as normalizing factors and they can be calculated as a function of c, v, p.[19] Grantham (1974) provides the distances of Equation (3) for all amino acid pairs, thus a distance matrix D. Furthermore, Grantham demonstrated that the relative substitution frequency of amino acid pairs and their distances $D_{G[i,j]}$'s are correlated, underlying the physicochemical basis of amino acid substitution.

Similarly to Grantham's amino acid distances, Sneath's index[20] takes into account 134 categories of activity and structure. The dissimilarity index $D_{S[i,j]}$ is the percentage of the sum of all properties not shared between two amino acids.

The amino acid neighborhood-based distance is correlated with both the Grantham's and Sneath's distances Figure 7, highlighting the fact that amino acid neighborhoods capture information related to the physicochemical properties of amino acids and also their relative substitution frequency.

The correlations between the Euclidean and Sneath's and the Euclidean and Grantham's distance matrices comprise additional evidence for the ability of protein structural information to predict the amino acid substitutions in a protein family.

3.6 | Construction of transition rate matrices for mitochondrial proteins

We tested our neighborhood-based amino acid transition rate matrix construction pipeline with human mitochondrial protein structures from the Protein Data Bank to examine whether the proposed

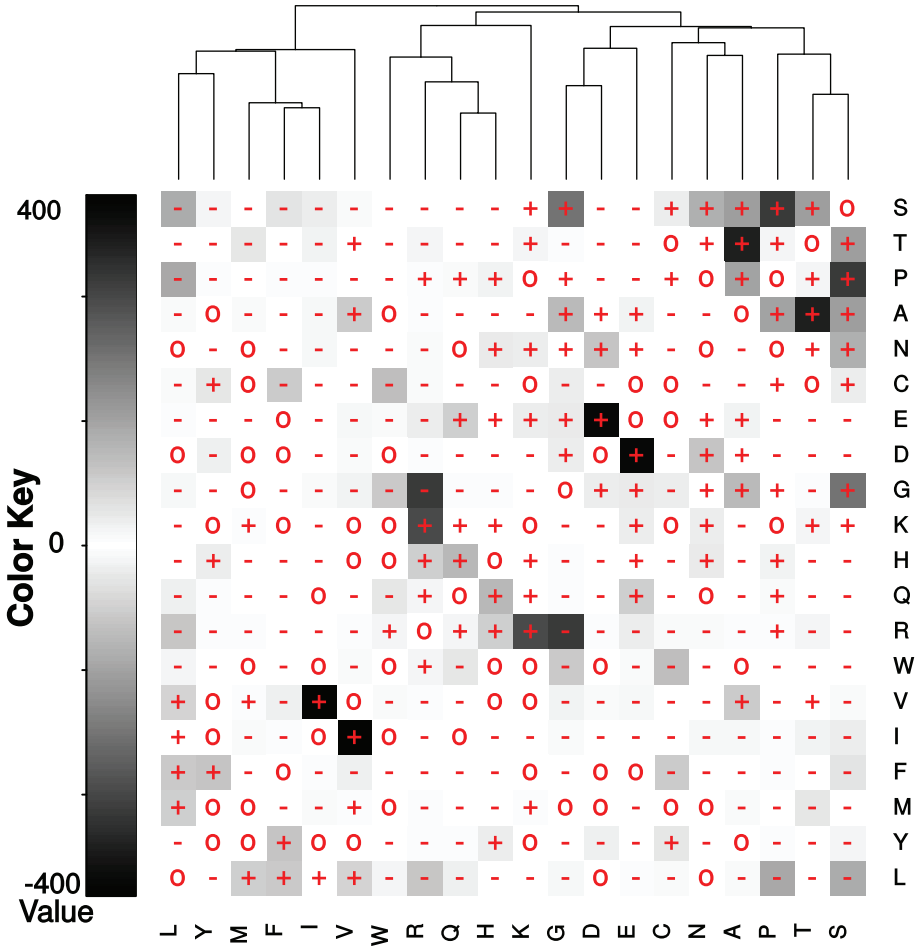

FIGURE 6 Comparison of likelihoods between BLOSUM62 and *nEs* for all amino acid pairs found in the protein alignments. Boxes with a "−" pinpoint to the specific amino acid pairs where *nEs* calculates a higher site likelihood. Contrarily, a "+" indicates amino acid pairs that the BLOSUM62 approach results in greater likelihood. Finally, "o" cells depict the amino acid pairs that were not found in the multiple alignments, thus no comparison was possible. The darker the tone of the cell the greater the difference in the likelihood between the BLOSUM62 and the *nEs* approach [Color figure can be viewed at wileyonlinelibrary.com]

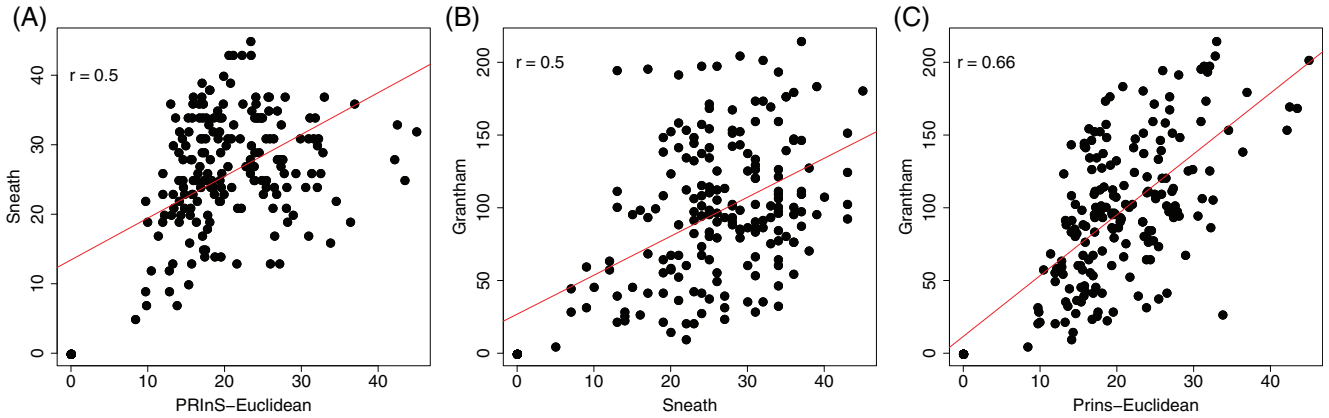

FIGURE 7 Scatterplots between amino acid distances for different distance methods. (A) Euclidean-based distance versus Sneath's index ($r \approx 0.5$). (B) Sneath's index versus Grantham's distance ($r \approx 0.5$). (C) Euclidean-based distance versus Grantham's distance ($r \approx 0.66$) [Color figure can be viewed at wileyonlinelibrary.com]

methodology can be extended to other protein groups beyond the α-helical membrane proteins. Since mitochondrial proteins do not represent a certain functional group of proteins (e.g., transmembranes) we expect less concordance to BLOSUM62 and PAM120 than the α-helical membrane proteins. Indeed, Figure 8 shows that the Pearson Correlation coefficient between the elements of the BLOSUM62 transition rate matrix and the mitochondrial Euclidean-based transition rate matrix (Supporting Information S1) is 0.228, whereas the Pearson correlation coefficient between the BLOSUM62 and the α-helical Euclidean-based transition rate matrix is 0.362. Thus, the agreement between BLOSUM62 and α-helical matrix is greater than the mitochondrial matrix (Figure 8).

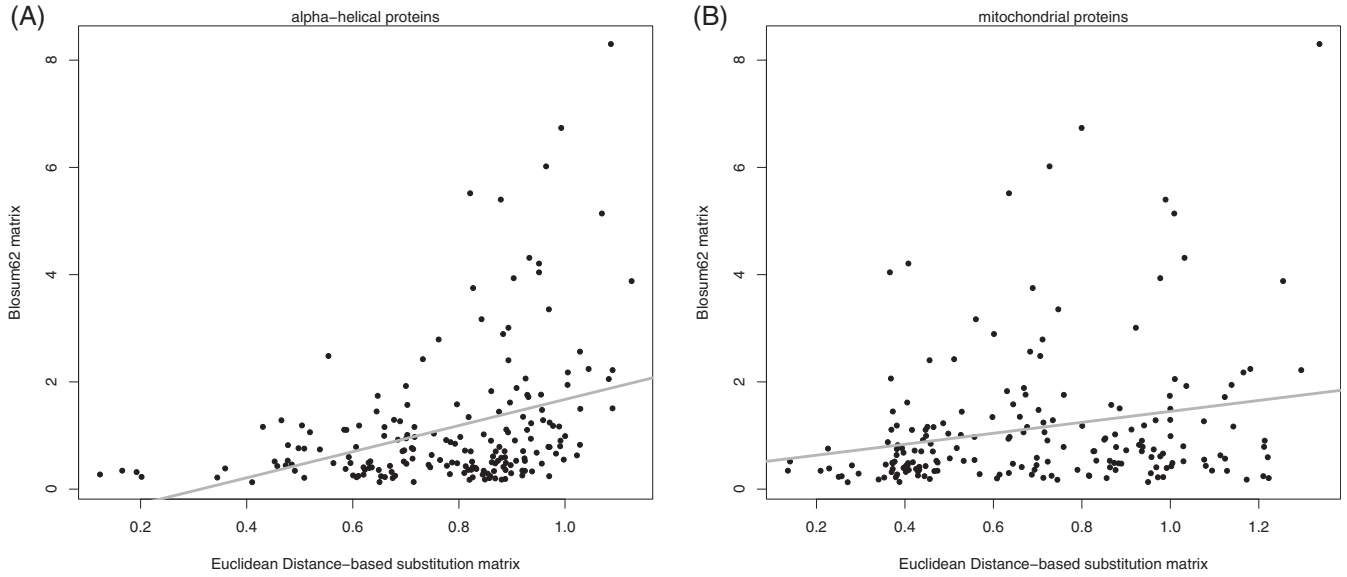

FIGURE 8 Pearson correlation coefficient between (A) BLOSUM62 and α-helical transmembrane protein transition rate matrices (Euclidean distance-based) and (B) BLOSUM62 and human mitochondrial transition rate matrices (Euclidean distance-based). The correlation coefficient in (A) is 0.362 and in (B) is 0.228. For both datasets, the positive correlation coefficient suggests that transition rate matrices based on amino acid neighborhood are at least partially concordant with the BLOSUM62 matrix, even though there is a greater matrix with the α-helical based matrix

4 | DISCUSSION

4.1 | Evaluation of PrInS ability to predict amino acid substitutions

In this study, we focused on the amino acid substitutions. We investigated if they can be determined from the properties of their neighborhood tertiary structure. We described statistically the amino acid residual neighborhoods using the software PrInS. Then, we converted PrInS output files to amino acid distance matrices and substitution rate matrices and evaluated their ability to model evolutionary changes.

Correlation analysis between the observed substitution frequencies (from multiple sequence alignments) and the distance matrices indicated that residual neighborhoods capture evolutionary information and thus they can be useful in modeling evolution. In other words, substitutions in multiple sequence alignments can be predicted from the amino acid neighborhoods: residuals with similar neighbors can substitute each other, with a higher rate, during evolution. More specifically, from the three distance matrices (Euclidean, Manhattan, and Pearson Squared) that were compared against the substitution matrix, only five amino acids were discordant in at least two comparisons. These amino acids were proline, arginine, serine, tryptophan, and tyrosine. From the substitution matrix, only a few substitutions involve tryptophan and tyrosine, whereas there is a multitude of substitutions involving proline, arginine and serine. In contrast, based on distance matrices, in the columns of tryptophan and tyrosine, there are small distance values, while in the proline, arginine, and serine columns distances are large. This can explain the discordance of these amino acids in these two matrices, because a large number of substitutions are associated with small distance values for an amino acid.

In addition, the Common Substitution Tool of NCBI Amino Acid Explorer[15] enhances our main result: Amino acid pairs with the lower distance values in the distance matrices, especially using Euclidean distances, are found to substitute each other more frequently. Amino Acid Explorer returned a list of the amino acids from the most common to the less common substitution for an amino acid we put as input. 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 are possibly based on the fact that the generation of the Euclidean distance matrix is based solely on the structural information of a protein family (α-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 α-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 α-helical membrane proteins.

4.2 | Multiple alignment scoring and clustering of proteins

After the evaluation of the ability of PrInS to predict amino acid substitutions, the scoring of multiple alignments with the matrices generated by our methodology was followed. The matrices that were used in the scoring of multiple alignments were distance-based, genetic-code-based and model substitutions matrices (BLOSUM62 and PAM120).[13,16] This was the second step in our analysis. Generally, most of the proteins did not significantly differ from the human homologs, which are found in the first column of all the heatmaps.

Obviously, homologous proteins in evolutionary distant species from human differ more than the homologs in more related species. Based on the substitution rate matrices, we calculated the likelihood for every protein pairwise alignment (human vs. nonhuman species). Using the Euclidean-based matrix, a cluster of proteins was appeared (hierarchical clustering) and its functions were studied using gProfiler.[17] 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.

4.3 | Limitations

Even though there are specific substitution matrices for transmembrane proteins (e.g., PHAT),[21] in this study we employed only general substitution matrices such as the BLOSUM62 and the PAM120. This is because our goal was not to provide a more suitable substitution matrix for a specific protein family but to provide insights into a plausible mechanism guiding the amino acid substitution during evolution. A thorough comparison with substitution matrices specialized on certain protein families (e.g., transmembrane proteins, globular proteins) will be the goal of a future study. In addition, a future goal is to construct substitution matrices for all protein families present in Protein Data Bank and compare them to obtain insights into their evolution.

5 | CONCLUSION

By using only publicly available structural data (PDB files), *without multiple sequence alignment files*, we were able to create transition rate matrices concordant with the widely-used in phylogenetics BLOSUM62 and PAM120 transition rate matrices. With these matrices we scored multiple-sequence alignments of proteins from closely-related species including human. It is important that neighborhood-based transition rate matrices can be easily built from publicly available data (PDB files) that exist in databases such as the Protein Data Bank database. The pipeline is not restricted to a specific species and it can be applied to any species.

The main result of this project is that, by solely using protein structural data, we are able to predict the amino acid substitutions in a protein family. The proposed algorithm can be generalized for any protein family to (i) build family specific substitution rate matrices and (ii) based on these matrices to highlight the differences observed between protein families. Furthermore, the pipeline is straightforward comprising the following steps: (i) download PDB files from a publicly available database, (ii) run the PrInS algorithm to generate a statistical description of amino acid neighborhoods, and finally (iii) generate transition rate matrices in which amino acids that have similar neighbors will be able to substitute each other at a higher rate. In many cases, these results are similar with the results obtained by multiple sequence alignment data. Finally, based on the outcomes of the study, we can support our initial hypothesis that the amino acids with similar neighbors in the tertiary structure of the protein can substitute each other during evolution.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26178.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Orientations of Proteins in Membranes (OPM) database. Protein IDs used in this study are mentioned in table 1 of the Nath Jha et al.4 These data were derived from the following resources available in the public domain: OPM (https://opm.phar.umich.edu/) and UCSC (https://genome.ucsc.edu/). Software PrInS is freely available from http://popgen.eu/wordpress/software/prins-protein-residues-interaction-statistics. Also, in the same address, we provide the protein data used in this study. We also provide PrInS and the protein data in the github repository https://github.com/idaios/prins. Substitution rate matrices produced by our study are available in the Supporting Information and in addition in https://github.com/idaios/prins/transition_matrices/. Also, the list of PDB files used for the analysis of mitochondrial data are available in the aforementioned github repository (mitochondrionpdbs.list).

ORCID

Pavlos Pavlidis https://orcid.org/0000-0002-8359-7257

REFERENCES

1. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*. 2009;10(10):709-720.
2. Egli M. Diffraction techniques in structural biology. In: Beaucage SL, Bergstrom DE, Herdewijn P, Matsuda A, eds. *Current Protocols in Nucleic Acid Chemistry*. Vol 41. Hoboken, NJ: John Wiley & Sons; 2010:7.13.1-7.13.35. https://doi.org/10.1002/0471142700.nc0713s41
3. Bernstein FC, Koetzle TF, Williams GJ, et al. The protein data bank. *FEBS J*. 1977;80(2):319-324.
4. Nath Jha A, Vishveshwara S, Banavar J. Amino acid interaction preferences in helical membrane proteins. *Protein Eng Des Sel*. 2011;24(8):579-588.
5. Siltberg-Liberles J, Grahnen JA, Liberles DA. The evolution of protein structures and structural ensembles under functional constraint. *Genes*. 2011;2(4):748-762.
6. Hatton L, Warr G. Protein structure and evolution: are they constrained globally by a principle derived from information theory? *PLoS One*. 2015;10(5):e0125663.
7. Franzosa E, Xia Y. Structural perspectives on protein evolution. *Annu Rep Comput Chem*. 2008;4(1):3-21.
8. Choi I-G, Kim S-H. Evolution of protein structural classes and protein sequence families. *Proc Natl Acad Sci U S A*. 2006;103(38):14056-14061.
9. Leelananda SP, Kloczkowski A, Jernigan RL. Fold-specific sequence scoring improves protein sequence matching. *BMC Bioinform*. 2016;17(1):328.
10. Kent W, Sugnet C, Furey T, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006.

11. Jiao X, Yang L, An M, Chen W. A modified amino acid network model contains similar and dissimilar weight. *Comput Math Meth Med*. 2013; 2013:197892.
12. Nath Jha A, Vishveshwara S, Banavar JR. Amino acid interaction preferences in proteins. *Protein Sci*. 2010;19(3):603-616.
13. Dayhoff M, Schwartz R, Orcutt B. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*. Vol 5. Silver Spring, MD: National Biomedical Research Foundation; 1978:345-352.
14. Stamatakis A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313.
15. Bulka B, desJardins M, Freeland SJ. An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices. *BMC Bioinform*. 2006;7(1):329.
16. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89(22):10915-10919.
17. Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. 2007. G: profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*, 35(suppl_2):W193–W200.
18. McKiernan CJ, Friedlander M. The retinal rod Na+/Ca2+,K +Exchanger contains a noncleaved signal sequence required for translocation of the n terminus. *J Biol Chem*. 1999;274(53):38177-38182.
19. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;185(4154):862-864.
20. Sneath P. Relations between chemical structure and biological activity in peptides. *J Theor Biol*. 1966;12(2):157-195.
21. Ng PC, Henikoff JG, Henikoff S. Phat: a transmembrane-specific substitution matrix. *Bioinformatics*. 2000;16(9):760-766.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Primetis E, Chavlis S, Pavlidis P. Evolutionary models of amino acid substitutions based on the tertiary structure of their neighborhoods. *Proteins*. 2021;1-12. https://doi.org/10.1002/prot.26178