Int. J. Mol. Sci. **2014**, *15*, 15963-15980; doi:10.3390/ijms150915963

OPEN ACCESS

International Journal of

Molecular Sciences

ISSN 1422-0067

www.mdpi.com/journal/ijms

Article

Neighbor Preferences of Amino Acids and Context-Dependent Effects of Amino Acid Substitutions in Human, Mouse, and Dog

Mingchuan Fu [1,2,†], Zhuoran Huang [1,2,†], Yuanhui Mao [1,2] and Shiheng Tao [1,2,*]

[1] College of Life Sciences and State Key Laboratory of Crop Stress Biology in Arid Areas, Northwest A&F University, Yangling 712100, China; E-Mails: fmcsky@nwsuaf.edu.cn (M.F.); sdrzh@nwsuaf.edu.cn (Z.H.); maoyuanhui@nwsuaf.edu.cn (Y.M.)

[2] Bioinformatics Center, Northwest A&F University, Yangling 712100, China

† These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: shihengt@nwsuaf.edu.cn; Tel.: +86-29-8709-1060; Fax: +86-29-8709-2262.

Received: 28 June 2014; in revised form: 27 August 2014 / Accepted: 2 September 2014 / Published: 10 September 2014

Abstract: Amino acids show apparent propensities toward their neighbors. In addition to preferences of amino acids for their neighborhood context, amino acid substitutions are also considered to be context-dependent. However, context-dependence patterns of amino acid substitutions still remain poorly understood. Using relative entropy, we investigated the neighbor preferences of 20 amino acids and the context-dependent effects of amino acid substitutions with protein sequences in human, mouse, and dog. For 20 amino acids, the highest relative entropy was mostly observed at the nearest adjacent site of either *N*- or *C*-terminus except C and G. C showed the highest relative entropy at the third flanking site and periodic pattern was detected at G flanking sites. Furthermore, neighbor preference patterns of amino acids varied greatly in different secondary structures. We then comprehensively investigated the context-dependent effects of amino acid substitutions. Our results showed that nearly half of 380 substitution types were evidently context dependent, and the context-dependent patterns relied on protein secondary structures. Among 20 amino acids, P elicited the greatest effect on amino acid substitutions. The underlying mechanisms of context-dependent effects of amino acid substitutions were possibly mutation bias at a DNA level and natural selection. Our findings may improve secondary structure prediction algorithms and protein design; moreover, this study provided useful information to develop empirical models of protein evolution that consider dependence between residues.

Keywords: protein; neighborhood; substitution; context-dependence; secondary structure

1. Introduction

Amino acid sequences are necessary to allow proteins to fold into their native conformations [1]. As such, protein sequence patterns should be characterized to understand protein structure, function, and stability. Previous studies revealed that amino acid compositions vary in secondary structures [2–5]. For example, M, A, K, E, and L are helix-preferred amino acids, whereas P and G likely disrupt helices. Likewise, V, I, T, F, W, and Y show high propensities for β-strand, whereas P, G, A, and E are poor β-strand-forming residues. In addition to amino acid preferences for different secondary structures, preferences for particular residue pairs in protein sequences have also been discovered. These preferred residue pairings are found in α-helices [6–11], parallel/antiparallel β-sheets [12–14], loops [15], and protein inter-domain linkers [16]. Such residue pairs are related to secondary structure formation and protein stabilization. In this work, the neighbor preferences of 20 amino acids were investigated and the neighbor preference patterns among different secondary structures were compared. This research may provide new insights into neighbor preferences of amino acids; furthermore, this study may improve secondary structure prediction algorithms and protein design.

Inspired by the research on neighbor preferences of amino acids, we further want to investigate whether or not amino acid substitutions also prefer neighborhood context and the specific context-dependence pattern of each substitution type. Thus, the second issue addressed in this study is to assess the context-dependent effects of amino acid substitutions. Nucleotide mutations are context dependent, and the most important mutation bias is the CpG effect [17–20]. Based on these empirical studies, several context-dependent evolutionary models for mammals have been established [21–24]. However, similar studies on amino acid substitutions are still very few. Understanding amino acid substitution patterns and constructing explicit protein evolution models are critical to phylogenetic analyses. Common protein evolution models, whether theoretical [25,26] or empirical [27–30], are usually constructed on the basis of the assumption that all protein sites evolve at the same rate and independent of other sites (*i.e.*, amino acid substitutions occurring at one site are independent from amino acids at other sites). Several models have been proposed to relax the assumption of equal evolutionary rates at all sites [31–36]; by contrast, the assumption of site-independence has been maintained. It is now widely agreed that the site-independence assumption is simplistic and biologically unrealistic. To relax such assumptions, researchers have developed some more elaborate models allowing for dependence between residues in recent years. Such models can be classified as knowledge-based (informational) models [37–39] and physics-based models [40,41]. However, these mechanistic models inaccurately describe protein thermodynamics and appear outperformed by some sophisticated site-independent models [42]. An alternative approach that may overcome the shortcomings of these mechanistic models is to use phenomenological models, which attempt to fit the results of sequence change without understanding the underlying biological process [43]. Wang *et al.* [44] first assessed the context-dependent effects of amino acid substitutions by using a dataset of 45,173 orthologous proteins. They found that amino acid substitutions are neighbor dependent, and the patterns of neighbor-dependence are similar between *N*- and *C*-termini. However, the specific context-dependence pattern of each amino acid substitution type has not been assessed. Moreover, context-dependence patterns in different protein secondary structures have not been considered in their research. In this work, the context-dependent effect of each amino acid substitution type and the context-dependence patterns in different secondary structures were comprehensively investigated. Our findings provided very useful information for further development of protein evolution empirical models that consider site dependence.

Relative entropy [45], also called Kullback-Leibler divergence or information gain, provides a measurement of the distance between two probability distributions *P* and *Q*. In general, *P* represents the observed probability distribution of a dataset and *Q* represents the expected or theoretical probability distribution. Applying relative entropy, we aimed to investigate the neighbor preference patterns of 20 amino acids and the context-dependent effects of amino acid substitutions.

2. Results

2.1. Neighbor Preferences of Amino Acids

We initially assessed the neighbor preference pattern of each amino acid type. Our results showed that all of the 20 amino acids were remarkably neighbor-preferred (Figure S1). The highest relative entropy was mainly observed at the nearest adjacent site of either *N*- or *C*-terminus, indicating that neighbor preferences were the strongest for the two immediate flanking sites; this value subsequently decreased when the distance to the 0 site increased. Relative entropy decreased rapidly in the nearest 5 to 7 flanking sites; afterward, relative entropy decreased very slowly and became not significant (Figure 1A). C and G are two exceptions. For C (Figure 2A), the highest relative entropies appeared at the third flanking sites of both *N*- and *C*-termini. Without these sites, relative entropies were relatively low. For G (Figure 3A), the relative entropies of *N*- and *C*-terminal flanking sites showed an evident periodic change, that is, relative entropies were remarkably high in all of the 3*n* (*n* = ±1, ±2, ±3…) flanking sites; other sites showed low values. The peaks at the 3*n* flanking sites decreased gradually when the distance to the 0 site increased. To determine the specific amino acids responsible for the high relative entropy at one flanking site, we further investigated the 20 relative entropies calculated using $P(a)\log_2\frac{P(a)}{Q(a)}$ ($a \in \{A, E, ..., Y\}$) at each flanking site of the 20 amino acids (Figures 1B, 2B, and 3B). We found that for more than half of the amino acids (A, L, V, P, G, S, T, Q, C, H, K, R, D, and E), the corresponding residues tended to show high propensity at the neighboring sites (e.g., in Figure 1B, amino acid A was a outlier which was above the upper whisker of the boxplot in each flanking site, which demonstrated that amino acid A was a type of preferred amino acid in the neighboring sites of amino acid A).

Figure 1. Neighbor preference pattern of the amino acid A. (**A**) Relative entropies of the neighboring sites of A of *N*- and *C*-termini. The vertical coordinate represents the value of relative entropy, and the horizontal coordinate represents the relative distance to the 0 site. The flanking sites of *N*-terminus are indicated by negative numbers and the flanking sites of *C*-terminus are indicated by positive numbers. The gray lines are the thresholds representing 0.001 significance level. The standard deviation of the relative entropy in each site is indicated by black error bars; (**B**) Boxplot of the 20 relative entropies calculated by $P(a)\log_2 \dfrac{P(a)}{Q(a)}$ ($a \in \{A, E, ..., Y\}$) at each flanking site of the amino acid A. The bold line in each box represents the median of the 20 values. The top and bottom lines of each box indicate the upper and lower quartiles, respectively. The upper and lower whiskers represent the largest data point which was less than the sum of the upper quartile plus 1.5 times the interquartile range (IQR), and the lowest data point which was greater than the lower quartile minus 1.5 IQR, respectively. In order to determine which amino acids were apparently preferred (or not preferred) in the neighboring sites, the outliers were represented by the corresponding one letter codes of amino acids.

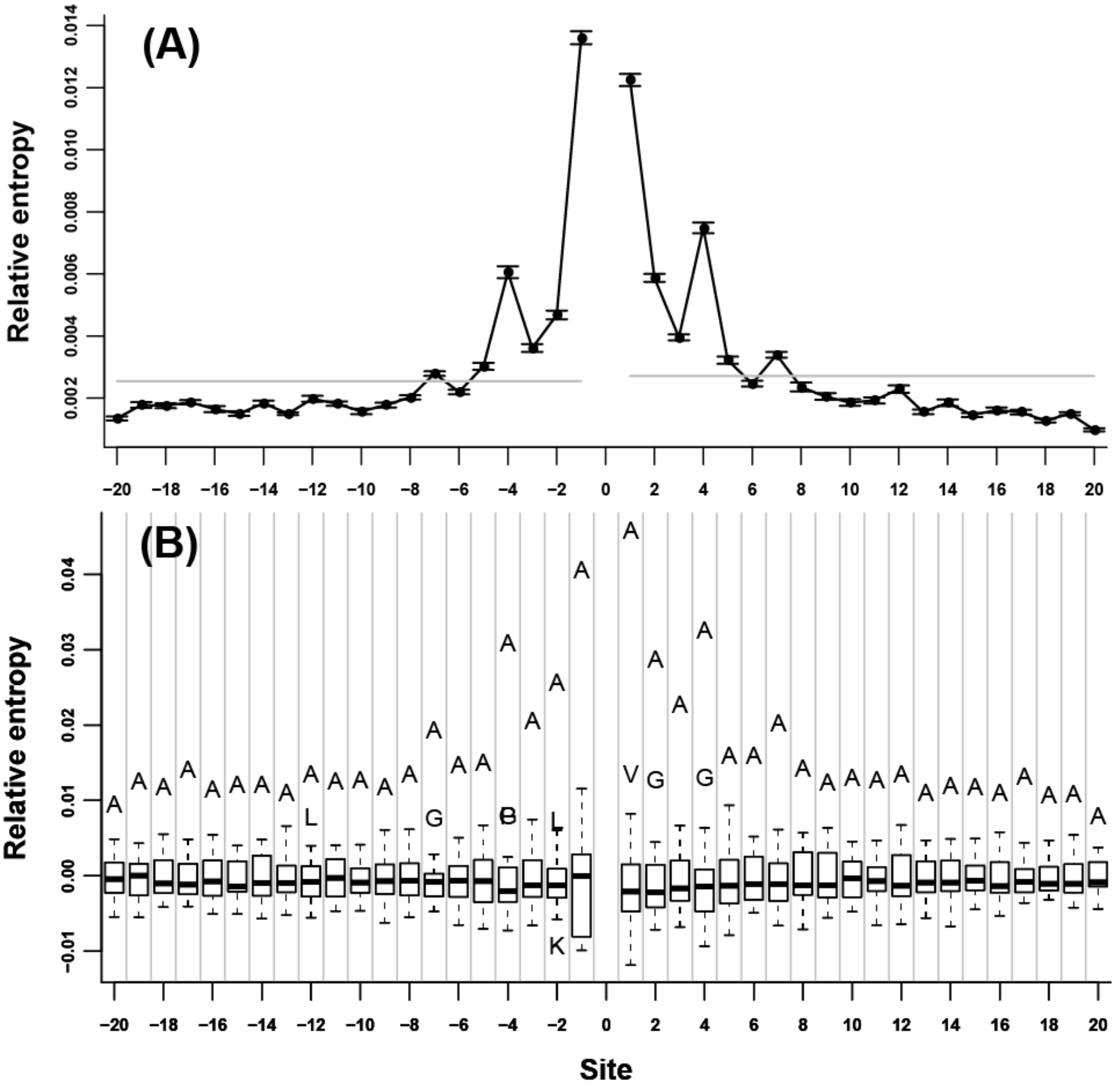


Figure 2. Neighbor preference pattern of the amino acid C. (**A**) Relative entropies of the neighboring sites of C of *N*- and *C*-termini; (**B**) Boxplot of the 20 relative entropies at each flanking site of the amino acid C.

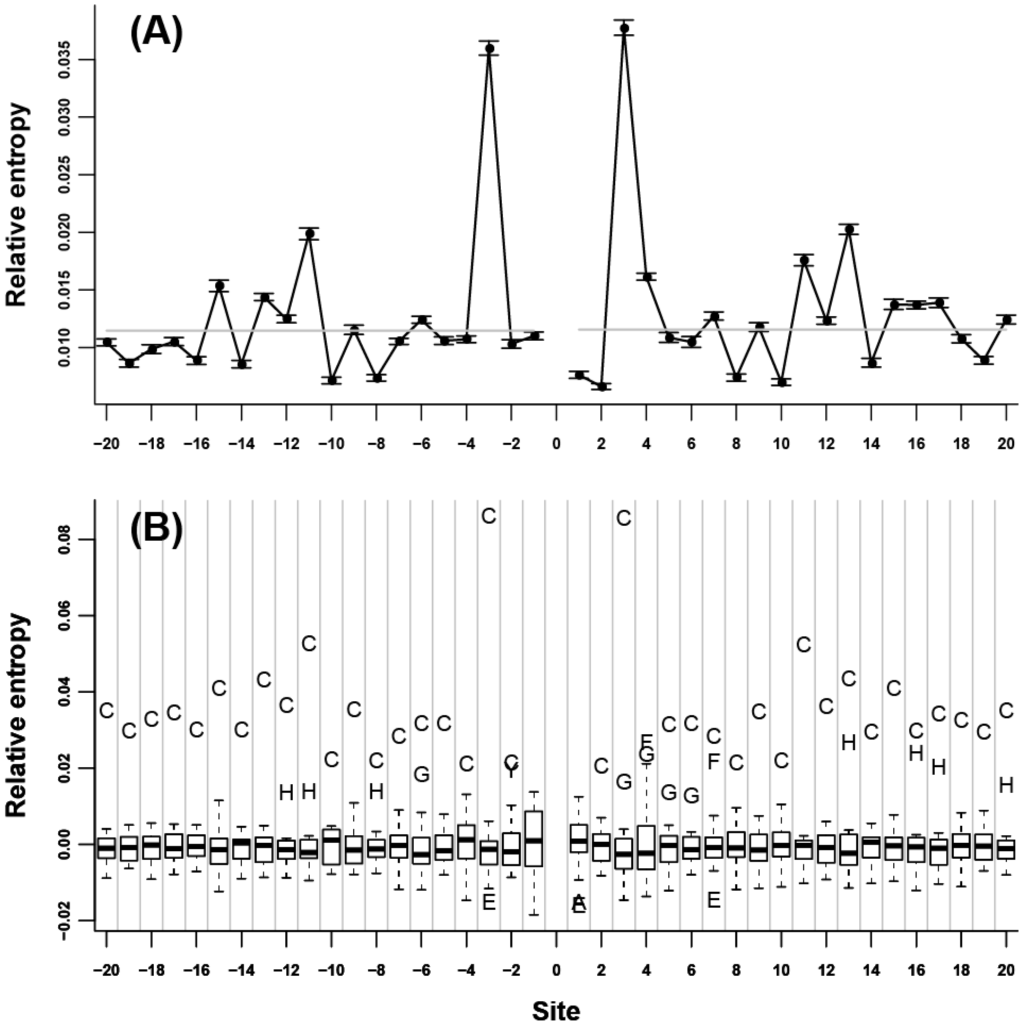


Figure 3. Neighbor preference pattern of the amino acid G. (**A**) Relative entropies of the neighboring sites of G of *N*- and *C*-termini; (**B**) Boxplot of the 20 relative entropies at each flanking site of the amino acid G.

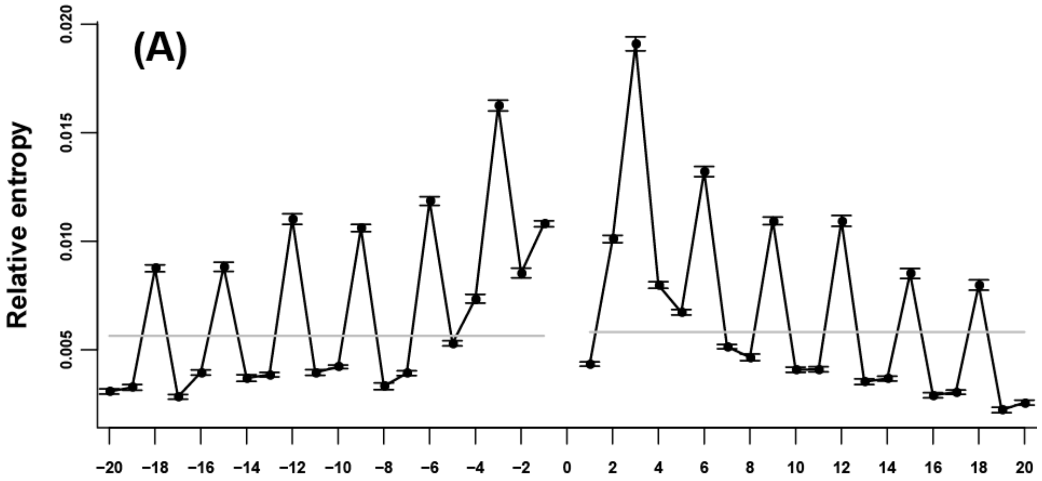
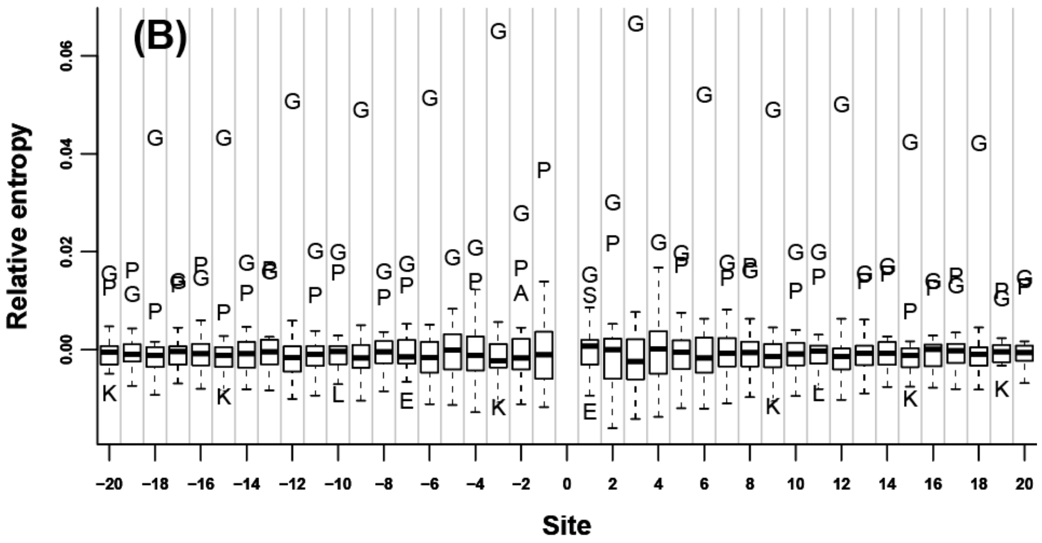


Figure 3. *Cont.*

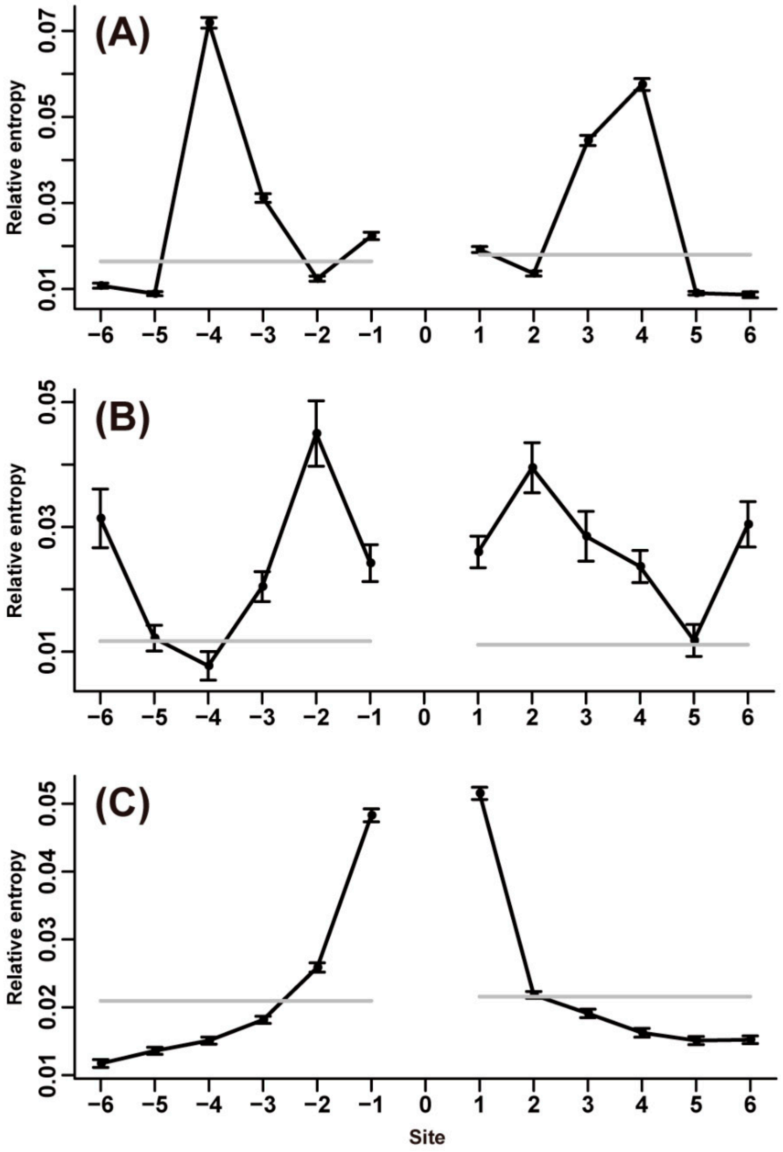


2.2. Neighbor Preferences of Amino Acids in Different Protein Secondary Structures

Considering the limitation of the average sequence length of each secondary structure (α-helix, β-strand and coil), we extracted six instead of twenty flanking sites for each type of amino acid. Neighbor preference patterns in the three secondary structures were different. In α-helix, the highest relative entropy was commonly seen at the third or fourth flanking site for 14 of the 20 amino acids (F, A, V, L, I, M, W, C, Y, Q, R, K, E, and D; Figure 4A; Figure S2). The majority of these amino acids favorably interacted with residues at spacing *i*, *i* + 3 or *i*, *i* + 4 of the same kind or similar polarity, that is, nonpolar/nonpolar or polar/polar pairings were predominant for the *i*, *i* + 3 and *i*, *i* + 4 combinations. The exceptions were C and Y, which showed high propensity for the residues of opposite polarity. L was frequently found at the third or fourth neighboring sites of nonpolar amino acids. Charged amino acids preferred to interact with residues at spacing *i*, *i* + 3 or *i*, *i* + 4 of opposing charge (e.g., R-E, K-E, and K-D pairings). Some exceptional amino acid pairings, such as L-V, L-I, L-L, F-L, A-A, Q-Q, and E-R, at spacing *i*, *i* + 4 observed in our study have also been documented in other studies [7,11]. For the six other amino acids (P, G, T, N, S, and H), the highest relative entropy was mainly observed at the second flanking site of *N*- or *C*-terminus (except P). In contrast to *i*, *i* + 3 and *i*, *i* + 4 pairings, the exceptional *i*, *i* + 2 amino acid combinations were all polar/nonpolar pairings except G. L was the most preferred residue at the ±2 sites of these amino acids.

Studies have investigated residue pairing preferences on adjacent β-strands [12–14]. In contrast to these studies, our study investigated the neighbor preferences of amino acids along the same strand. Although intimate interactions between residues in one strand, which is almost fully extended, are rare because of a large C_α-C_α distance [13], evident neighbor preferences were still found. For the amino acids in β-strand, the highest relative entropy commonly appeared at the immediate or second adjacent site (Figure 4B; Figure S3). The preferred *i*, *i* + 1 pairings were mainly polar/nonpolar combinations (e.g., V-S, I-Q, N-I, F-H, F-S, D-I, and V-E); by contrast, polar/polar or nonpolar/nonpolar combinations were predominant for *i*, *i* + 2 pairings (e.g., G-I, V-G, N-T, and N-S). V and I were frequently found at neighboring sites of other amino acids. Although G is a poor β-strand-forming residue, G showed moderately high propensity at neighboring sites of other amino acids, such as L, V, I, and F.

Figure 4. Neighbor preference patterns of V, T, and E in different secondary structures. (**A**) Neighbor preference pattern of the amino acid V in α-helix; (**B**) Neighbor preference pattern of the amino acid T in β-strand; and (**C**) Neighbor preference pattern of the amino acid E in coil. (Boxplots of these amino acids can be found in Figures S2–S4).

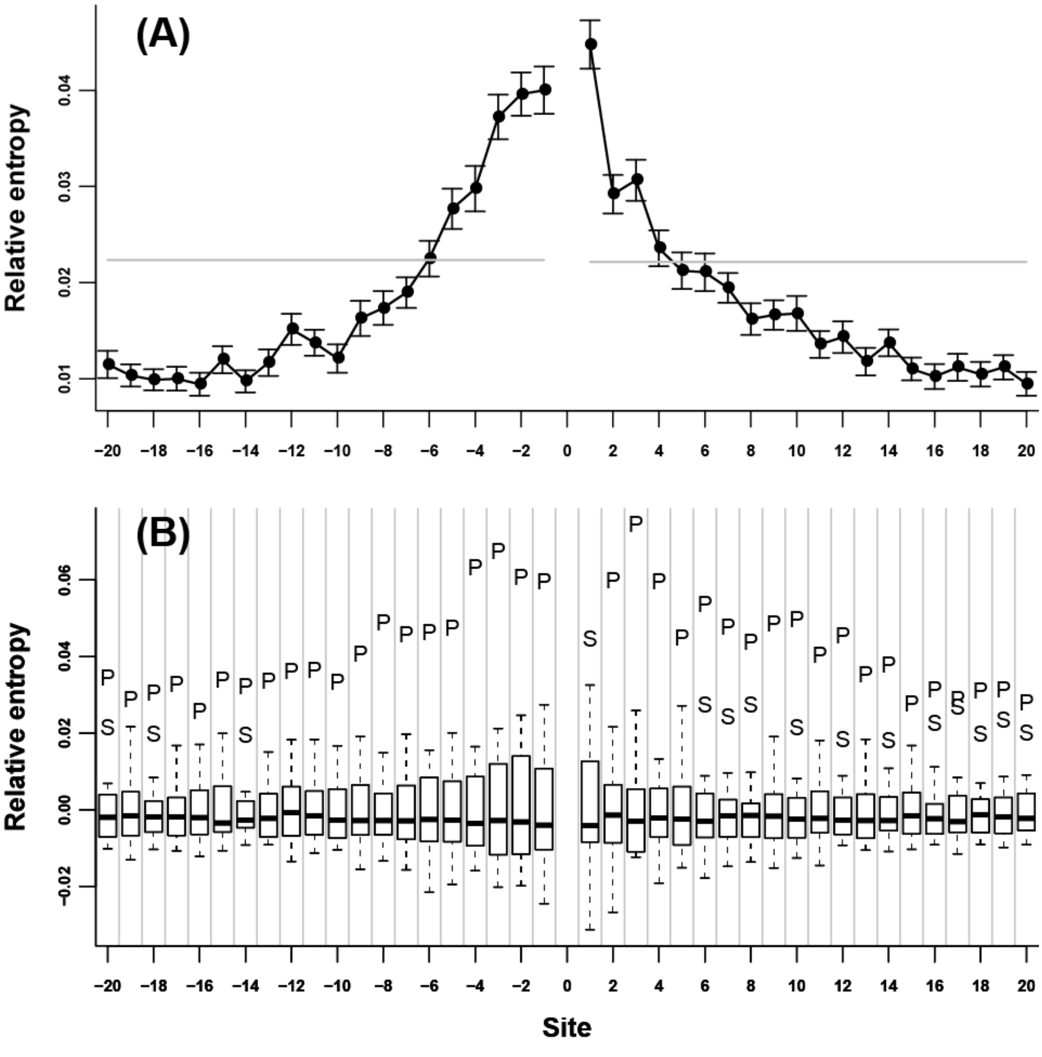


In coil, all of the amino acids showed evident neighbor preferences; the highest relative entropy was mainly observed at the nearest flanking site (Figure 4C; Figure S4). For C and G, the highest relative entropies were detected at the third flanking sites of both *N*- and *C*-termini. In addition, both residues showed apparently high relative entropies at ±6 sites. More than half of the amino acids in coil (A, P, G, S, Q, C, Y, H, K, R, D, and E) showed high propensity for the same type of residues. P occurred frequently at neighboring sites of almost all of the amino acids.

2.3. Context-Dependent Effects of Amino Acid Substitutions

To investigate whether or not amino acid substitutions are dependent on neighborhood context, we then assessed the pattern of context-dependent effects of amino acid substitutions. Our results showed that nearly half of the 380 amino acid substitution types were remarkably context dependent, and the highest relative entropies were mainly observed at the two nearest flanking sites (Figure 5A; Figure S5). Among the 20 amino acids, P, E, S, A, and G were frequently found at neighboring sites of amino acid substitutions, particularly substitutions between nonpolar residues or between nonpolar and polar residues. For substitutions between two polar amino acids, P, E, S, K, and Q appeared to be the preferred neighbors. A common characteristic of the amino acid substitutions was that when they substituted to P, E, S, A, G or K, the neighboring sites tended to show high propensity for the amino acid type as the post-substituted one (Figure 5B). Considering the self-preference of these amino acids (P, E, S, A, G and K), we found that a residue with a specific neighborhood context in proteins was possibly substituted in another residue, which exhibited a high propensity of this neighborhood context. Among 20 amino acids, P was the most preferred residue at neighboring sites of many substitution types. For some substitution types, this high propensity of P could extend to nearly all of the flanking sites (Figure 5B). Although P was not the most heterogeneous amino acid at neighboring sites in some other cases, P showed relatively high frequency in association with other amino acids.

Figure 5. Context-dependence pattern of *S*→*P* substitution. (**A**) Relative entropies of the neighboring sites of the *S*→*P* substitution of *N*- and *C*-termini; (**B**) Boxplot of the 20 relative entropies at each flanking site of the *S*→*P* substitution.

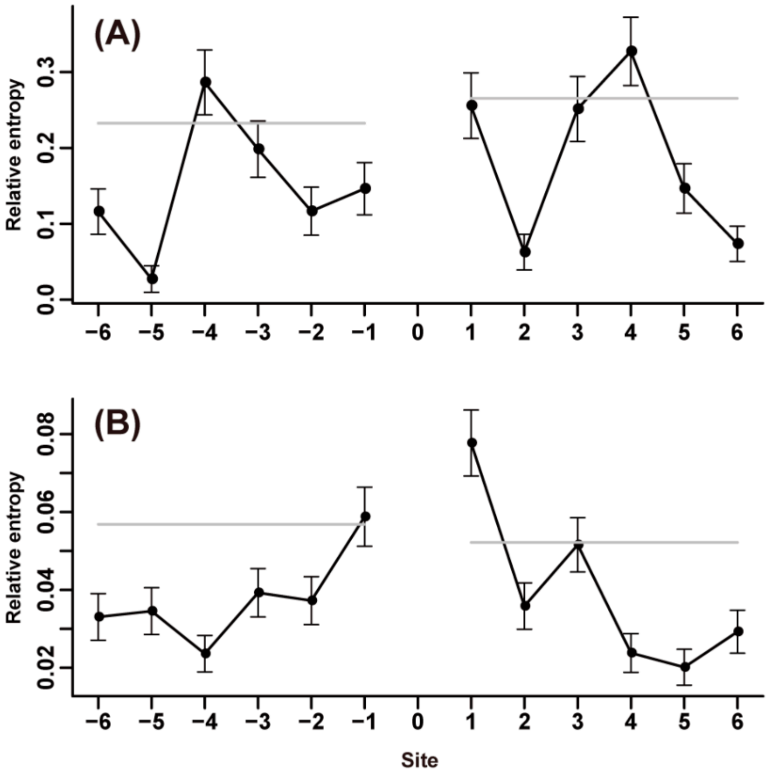


2.4. Context-Dependent Effects of Amino Acid Substitutions in Different Protein Secondary Structures

Some amino acid substitution types in each secondary structure were excluded from this research because of limited data size. In β-strand, only two substitution types were considered and neither of them showed an evident context-dependent effect.

In α-helix, 43 out of the 96 considered substitution types showed evident context-dependent effects (Figure S6). Similar to the neighbor preference patterns of amino acids in α-helix, the highest relative entropy of most amino acid substitution types was observed at the third or fourth flanking site (Figure 6A). Notably, the ±3 or ±4 sites of these amino acid substitutions always showed high propensity for one or more residues of E, Q, K, and R. Eight substitution types showed the highest relative entropy at the second flanking site, and L was the most preferred neighbor at the ±2 sites in most cases. Interestingly, all of these eight substitution types (except *L→V*) were the ones substituted to G, T, N, or S (e.g., *A→G*, *I→T*, *K→N*, and *A→S*) showing the highest neighbor preference for the ±2 sites in α-helix. In addition, some substitution types showed the greatest heterogeneity at the nearest flanking sites (particularly +1 site); most of these substitutions involved L, I, or V (e.g., *I→M*, *L→F*, and *T→V*).

Figure 6. Context-dependence pattern of *G→E* substitution in α-helix and coil. (**A**) Context-dependence pattern of the *G→E* substitution in α-helix; (**B**) Context-dependence pattern of the *G→E* substitution in coil. (Boxplots of this amino acid substitution type in α-helix and coil can be found in Figures S6 and S7).



In coil, 163 substitution types were considered. Among these substitutions, 40 showed evident context-dependent effects (Figure S7). In these substitution types, the highest relative entropy mostly appeared at the nearest flanking sites (Figure 6B). Among the 20 amino acids, P occurred most frequently at the neighboring sites of amino acid substitutions. Furthermore, A, S, E, and G showed high propensities at neighboring sites. Substitutions to A or E showed similar propensities of neighborhood context as the corresponding post-substituted residue.

3. Discussion

In this research, more than half of the 20 amino acids showed propensity toward the amino acid types as themselves at the neighboring sites. This self-preference is consistent with that in previous studies and is likely caused by replication slippage and nucleotide mutation bias [46–48]. Self-clustering of amino acids contributes to the emergence of novel proteins and protein-protein interaction networks [49,50]. I, M, F, W, N, and Y do not show evident self-preference pattern, suggesting that tandem repeats of these residues are not favored in proteins [51].

The exceptional neighbor preference patterns of C and G are structurally and functionally important. Miseta and Csutora [52] initially revealed that two Cs are frequently found with the separation of two other residues in various proteins of several species. They suggested that this result is attributed to the first and fourth residues in α-helices or β-turns that are closest to each other. Contrary to previous studies, this exceptional pattern of C was only observed in coils in our study. The periodic occurrence of G was observed in this study. Periodic patterns play a role in structural packing and atom interactions [51]. Previous studies showed that G at every third position is essential for the formation of a collagen triple helix [53]. Using Pfam [54], we further grouped the proteins according to the presence or absence of the collagen triple-helix domain. The result showed that a remarkable periodic occurrence of G occurred in the collagen proteins; by contrast, no such pattern was observed in the other proteins (Figure S8). Hence, the periodic occurrence of G observed in this study resulted from the influence of collagen proteins.

Cellular location greatly influences protein structure and function. The same type of amino acid may behave distinctively in different protein environments. To assess whether or not exceptional neighbor preference patterns of C and G are identical in different protein environments, we further investigated the neighbor preferences of these two amino acids in different cellular locations (results of C are shown in Figure S9, results of G are shown in Figure S10). In nuclear proteins, the relative entropies were strikingly high at the ±3 sites of C. In cytosol proteins, the ±3 sites of C also showed relatively high relative entropies. For G in proteins located in the cytosol, endoplasmic reticulum, extracellular environment, lysosome, and mitochondria, the pattern of periodic change was observed. No such pattern was observed in other protein environments. We also repeated this work by investigating the protein sequences with corresponding GO [55] entries (about one-third of the dataset) which were retrieved by DAVID [56,57]. The results showed that the neighbor preference patterns of C and G in different cellular locations were similar between these two datasets (Figure S11). These findings revealed that neighbor preference patterns for both amino acids greatly differed in various protein environments.

The neighbor preference patterns of amino acids in different secondary structures are necessary to maintain the corresponding structural conformation. Distinct neighbor preference patterns of amino acids were found in α-helix, β-strand, and coil. In α-helix, the highest relative entropy occurred at the third or fourth flanking sites in the majority of the amino acids. This result is mainly attributed to the residue pairs at spacing i, $i + 3$ or i, $i + 4$ appearing on the same side of α-helix; this spatial proximity induces the two side chains of these pairs to favorably interact and stabilize helices by salt bridges, hydrogen bonds, or hydrophobic interactions of particular amino acid combinations [7,11]. In particular, H-bonding between residues at spacing *i*, *i* + 4 contributes greatly to α-helix stabilization. The high propensity for nonpolar/nonpolar or polar/polar combinations of *i*, *i* + 3 and *i*, *i* + 4 residue pairings is mainly attributed to the amphiphilic nature of α-helix, which contains one hydrophobic side and one hydrophilic side. The exceptional *i*, *i* + 4 polar/nonpolar pairings observed in this study indicate that the interactions between some residues of opposite polarity may be necessary to stabilize a α-helix structure [8]. The preferred *i*, *i* + 2 pairings were predominant polar/nonpolar combinations mainly because the residues at spacing *i*, *i* + 2 are on the opposite side of α-helix.

One characteristic of the neighbor preference pattern in β-strand is the high propensity for polar/nonpolar combinations of *i*, *i* + 1 pairings and polar/polar or nonpolar/nonpolar combinations of residues at spacing *i*, *i* + 2. This observation is reasonable because the alternating pattern of nonpolar and polar residues is a general characteristic of β-strands and is necessary to determine β-strand structure [58,59].

In coil, P is frequently found at neighboring sites of almost all of the amino acids. P is the sole imino acid among 20 amino acids. The amino nitrogen of P is bonded to two alkyl groups rather than one alkyl group; therefore, no amide hydrogen can be donated to form H-bonding. This unique characteristic allows P to break α-helix and β-strand conformations and lead to form irregular secondary structures [15]. Consequently, P was the preferred neighbor of residues in coils.

In this work, the context-dependent effects of amino acid substitutions were comprehensively investigated. The underlying mechanisms of the context-dependent effects remain unclear. Two possible reasons are nucleotide mutation bias and natural selection. Misawa and Kikuno [60] found that approximately 14% of synonymous and nonsynonymous substitutions in human genes are caused by CpG hypermutations [61]. Considering that a nonsynonymous substitution is involved in the CpG effect (e.g., a nonsynonymous substitution from Val to Ile, *i.e.*, *GTT→ATT*, *GTC→ATC* or *GTA→ATA*, provided that the third position of the 5'-adjacent codon is C), it may be retained rather than purified if such substitution does not apparently change protein structure and function. Thus, some amino acid substitutions possibly occur in specific neighborhood contexts because of mutation bias. Another reason is natural selection. One notable finding in this work is that amino acid substitutions likely showed similar propensities of neighborhood context to those of post-substituted residue, particularly in the substitutions to P, E, S, A, G, and K in all of the proteins, to G, T, N, and S in α-helices, and to A and E in coils. This result indicated that one amino acid with a specific neighborhood context in proteins was possibly substituted by the amino acid with the propensity of such neighborhood context. This characteristic of amino acid substitutions is reasonable because natural selection favors the maintenance of protein local structures and functions.

4. Experimental Section

4.1. Neighbor Preferences of Amino Acids

We investigated the neighbor preference pattern of each amino acid type by using the protein sequences of human (*Homo sapiens*), mouse (*Mus musculus*), and dog (*Canis lupus familiaris*) downloaded at Ensembl database [62,63]. A total of 20 flanking sites of *N*- and *C*-termini of each amino acid were extracted. The sites containing less than 20 flanking sites of either *N*- or *C*-terminus were excluded from the research. We extracted six instead of twenty flanking sites when neighbor preference patterns were investigated in different secondary structures because of the limitation of the average sequence length of protein secondary structures.

4.2. Context-Dependent Effects of Amino Acid Substitutions

The orthologous information of mammals was downloaded from OrthoDB database [64,65], which provides the hierarchical catalog of orthologs, including 252 eukaryotic species and 1115 bacteria. The OrthoDB database can provide the corresponding Ensembl Protein IDs of each species in each orthologous group. We extracted the orthologous groups of human, mouse, and dog (a total of 11,007 orthologous groups), and downloaded the corresponding protein sequences from the Ensembl database.

We aligned the orthologous sequences of human, mouse, and dog by using clustalw [66] with default parameters. Using the PHYLIP [67] format files produced by clustalw as input files, the codeml program in the PAML (Phylogenetic Analysis by Maximum Likelihood) package [68] was then used to reconstruct the ancestral sequences. The control file for the codeml program was codeml.ctl. The empirical model of jones.dat (the parameter aaRatefile in codeml.ctl) was used in this study. The parameters "seqfile" and "outfile" were changed for each orthologous (alignment) group by a simple Perl script. A phylogenetic tree being constructed by the taxonomy tools of NCBI [69] was used (Figure S12). The average accuracies of the two ancestral sequences were >96% and 98% (PAML calculates the posterior probability of each site in ancestral node by using maximum likelihood method [70], which can be used as the measurement of the accuracy of the site. The overall accuracies of the ancestral sequences for each orthologous group could be found in the rst output file, which was produced by the codeml program).

There were four branches along the phylogenetic tree in total (Figure S12). In each branch, there were two nodes, which represented the descendant node and the corresponding ancestral node. To infer the amino acid substitutions along one branch, pair-wise comparisons between the sequences of the two nodes were conducted (the rst files produced by the codeml program could give us the summary of amino acid substitutions along each branch). At last, above 1,000,000 substitutions were inferred along the tree. To assess the context-dependent effects of amino acid substitutions, we extracted 20 flanking sites of *N*- and *C*-termini from each substitution. We excluded substitution sites with <20 flanking sites. Six flanking sites were extracted to investigate context dependence patterns of amino acid substitutions in different secondary structures.

4.3. Calculation of Relative Entropy

The relative entropy [45] of an amino acid *a* at one particular flanking site of one type of amino acids or substitutions was calculated as follows:

$$D(P_a \| Q_a) = P(a)\log_2 \frac{P(a)}{Q(a)} \tag{1}$$

where $P(a)$ represents the observed frequency of amino acid *a* at a given flanking site and $Q(a)$ represents the background or expected frequency of amino acid *a* at the site.

We define the relative entropy of one given flanking site as the sum of the relative entropies of the 20 amino acids at the site, as expressed in the following equation:

$$D(P\,||\,Q) = \sum_{a\in\{A,E,\ldots,Y\}} P(a)\log_2 \frac{P(a)}{Q(a)} \tag{2}$$

This summation is always nonnegative and is equal to zero if and only if $P(a) = Q(a)$ for all of the 20 amino acids at one flanking site.

To investigate the neighbor preferences of amino acids, we assigned the background distribution of amino acids as the observed frequencies of the 20 amino acids in the corresponding dataset (all of the protein sequences, sequences in one secondary structure, or in one cellular location). To analyze the context-dependent effects of amino acid substitutions and prevent the neighbor preferences of amino acids from providing biased results, we calculated the background distribution of amino acids as the distribution of the 20 amino acids at each flanking site of one particular amino acid type in the corresponding dataset. For example, the background distribution of amino acids at each flanking site in *A*→*D* substitution corresponded to the amino acid distribution at the corresponding flanking site of A.

To estimate thresholds, we randomly shuffled the 20 (or 6) sites of all *N*- or *C*-terminal flanking sequences of one type of amino acids or substitutions and then recalculated the relative entropy of each flanking site [71]. This process was repeated 500 times, and 10,000 (20 × 500) simulated relative entropies were obtained. We chose the tenth-highest value as threshold representing the 0.001 significance level. For investigations on different protein secondary structures, we selected the third-highest value among 3000 (6 × 500) simulated relative entropies as threshold. The thresholds of the two sides commonly showed differences because the data sizes of *N*- and *C*-terminal flanking sequences were usually distinct. Through bootstrap samplings of the dataset for 100 times, the standard deviations of relative entropies were calculated.

4.4. Prediction of Protein Secondary Structures

We predicted the secondary structures of protein sequences by using the Jpred server [72]. Jpred is a protein secondary structure prediction server incorporating the Jnet algorithm [73]. The Jnet method was developed by seven-fold cross-validated training on the dataset derived from SCOP [74] database at the superfamily level. The server can update in sync with the major updates of SCOP and UniProt [75], which makes it maintain high-accuracy predictions. Now, the average accuracy of this server is >81%. This server predicts three categories of secondary structures (α-helix, β-strand, and coil) of a protein sequence. In this study, we used Advanced Jpred [76] for batch submission of the protein sequences. The input files were in FASTA format (a format begins with a single description line, then followed by sequence lines), with each sequence being given a unique name.

4.5. Prediction of Protein Cellular Locations

We predicted the cellular locations of our dataset by using WoLF PSORT [77,78]. The server can convert amino acid sequences into numerical localization features based on known sorting signal motifs and several other sequence features. A wrapper method is used to select the most relevant features. The dataset of WoLF PSORT is comprised of fungi, plant and animal. The information of cellular location used in the server is obtained from UniProt [75] and GO [55] databases. At present, the average prediction accuracy of the server is >80%. This server can be used to predict approximately 11 cellular locations (e.g., cytosol, extracellular, nuclear, and plasma membrane) in animal sequences.

5. Conclusions

In this work, amino acids were evidently neighbor preferred and the amino acid substitutions were apparently context dependent. These findings could be exploited in the improvement of secondary structure prediction algorithms and further development of protein evolution models. Further studies should be conducted to investigate these neighbor preference patterns in more species and proteins with different functions. Further studies could also be performed to construct a context-dependent model of protein evolution incorporating the results of this work.

Supplementary Materials

Supplementary materials can be found at http://www.mdpi.com/1422-0067/15/9/15963/s1.

Acknowledgments

We would like to thank all the members in the Bioinformatics Centre in Northwest A&F University for the daily useful discussions. This work was supported by the PhD Programs Foundation of the Ministry of Education of China (20100204110026) and project from National Science Foundation of China (31125007). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

MCF and ZRH participated in the design of the research, performed the data analyses and drafted the manuscript. YHM participated in its design and coordination. SHT conceived of the study and helped to draft the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Dwyer, D.S. Nearest-neighbor effects and structural preferences in dipeptides are a function of the electronic properties of amino acid side-chains. *Proteins* **2006**, *63*, 939–948.
2. Chou, P.Y.; Fasman, G.D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **1974**, *13*, 211–222.
3. Chou, P.Y. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45–148.
4. Chou, P.Y.; Fasman, G.D. Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **1978**, *47*, 251–276.

5. Chakrabartty, A.; Kortemme, T.; Baldwin, R.L. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **1994**, *3*, 843–852.
6. Padmanabhan, S.; Baldwin, R.L. Helix-stabilizing interaction between tyrosine and leucine or valine when the spacing is *i*, *i* + 4. *J. Mol. Biol.* **1994**, *241*, 706–713.
7. Padmanabhan, S.; Baldwin, R.L. Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. *Protein Sci.* **1994**, *3*, 1992–1997.
8. Andrew, C.D.; Penel, S.; Jones, G.R.; Doig, A.J. Stabilizing nonpolar/polar side-chain interactions in the α-helix. *Proteins Struct. Funct. Bioinform.* **2001**, *45*, 449–455.
9. Wang, J.; Feng, J.A. Exploring the sequence patterns in the α-helices of proteins. *Protein Eng.* **2003**, *16*, 799–807.
10. Fonseca, N.A.; Camacho, R.; Magalhães, A.L. Amino acid pairing at the *N*- and *C*-termini of helical segments in proteins. *Proteins Struct. Funct. Bioinform.* **2008**, *70*, 188–196.
11. De Sousa, M.M.; Munteanu, C.R.; Pazos, A.; Fonseca, N.A.; Camacho, R.; Magalhães, A. Amino acid pair-and triplet-wise groupings in the interior of α-helical segments in proteins. *J. Theor. Biol.* **2011**, *271*, 136–144.
12. Wouters, M.A.; Curmi, P.M. An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: The differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins* **1995**, *22*, 119–131.
13. Hutchinson, E.G.; Sessions, R.B.; Thornton, J.M.; Woolfson, D.N. Determinants of strand register in antiparallel β-sheets of proteins. *Protein Sci.* **1998**, *7*, 2287–2300.
14. Fooks, H.; Martin, A.; Woolfson, D.; Sessions, R.; Hutchinson, E. Amino acid pairing preferences in parallel β-sheets in proteins. *J. Mol. Biol.* **2006**, *356*, 32–44.
15. Crasto, C.J.; Feng, J. Sequence codes for extended conformation: A neighbor-dependent sequence analysis of loops in proteins. *Proteins* **2001**, *42*, 399–413.
16. George, R.A.; Heringa, J. An analysis of protein domain linkers: Their classification and role in protein folding. *Protein Eng.* **2002**, *15*, 871–879.
17. Blake, R.D.; Hess, S.T.; Nicholson-Tuell, J. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.* **1992**, *34*, 189–200.
18. Zhao, Z.; Boerwinkle, E. Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome. *Genome Res.* **2002**, *12*, 1679–1686.
19. Nevarez, P.A.; DeBoever, C.M.; Freeland, B.J.; Quitt, M.A.; Bush, E.C. Context dependent substitution biases vary within the human genome. *BMC Bioinform.* **2010**, *11*. doi:10.1186/1471-2105-11-462.
20. Panchin, A.; Mitrofanov, S.; Alexeevski, A.; Spirin, S.; Panchin, Y. New words in human mutagenesis. *BMC Bioinform.* **2011**, *12*, doi:10.1186/1471-2105-12-268.
21. Baele, G.; van de Peer, Y.; Vansteelandt, S. Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate non-coding sequences. *BMC Evol. Biol.* **2010**, *10*, doi:10.1186/1471-2148-10-244.

22. Baele, G. Context-dependent evolutionary models for non-coding sequences: An overview of several decades of research and an analysis of laurasiatheria and primate evolution. *Evol. Biol.* **2012**, *39*, 61–82.
23. Baele, G.; van de Peer, Y.; Vansteelandt, S. Using non-reversible context-dependent evolutionary models to study substitution patterns in primate non-coding sequences. *J. Mol. Evol.* **2010**, *71*, 34–50.
24. Hwang, D.G.; Green, P. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13994–14001.
25. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862–864.
26. Miyata, T.; Miyazawa, S.; Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **1979**, *12*, 219–236.
27. Dayhoff, M.O. A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.* **1978**, *5*, 345–352.
28. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919.
29. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282.
30. Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699.
31. Overington, J.; Johnson, M.S.; Sali, A.; Blundell, T.L. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc. R. Soc. Lond. B* **1990**, *241*, 132–145.
32. Luthy, R.; McLachlan, A.D.; Eisenberg, D. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* **1991**, *10*, 229–239.
33. Overington, J.; Donnelly, D.; Johnson, M.S.; Šali, A.; Blundell, T.L. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci.* **1992**, *1*, 216–226.
34. Topham, C.M.; McLeod, A.; Eisenmenger, F.; Overington, J.P.; Johnson, M.S.; Blundell, T.L. Fragment ranking in modelling of protein structure: Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* **1993**, *229*, 194–220.
35. Thorne, J.L.; Goldman, N.; Jones, D.T. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **1996**, *13*, 666–673.
36. Goldman, N.; Thorne, J.L.; Jones, D.T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **1998**, *149*, 445–458.
37. Robinson, D.M.; Jones, D.T.; Kishino, H.; Goldman, N.; Thorne, J.L. Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* **2003**, *20*, 1692–1704.
38. Rodrigue, N.; Lartillot, N.; Bryant, D.; Philippe, H. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* **2005**, *347*, 207–217.
39. Kleinman, C.L.; Rodrigue, N.; Lartillot, N.; Philippe, H. Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.* **2010**, *27*, 1546–1560.

40. Rastogi, S.; Reuter, N.; Liberles, D.A. Evaluation of models for the evolution of protein sequences and functions under structural constraint. *Biophys. Chem.* **2006**, *124*, 134–144.
41. Grahnen, J.; Nandakumar, P.; Kubelka, J.; Liberles, D. Biophysical and structural considerations for protein sequence evolution. *BMC Evol. Biol.* **2011**, *11*, 361. doi:10.1186/1471-2148-11-361.
42. Rodrigue, N.; Philippe, H.; Lartillot, N. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol. Biol. Evol.* **2006**, *23*, 1762–1775.
43. Liberles, D.A.; Teichmann, S.A.; Bahar, I.; Bastolla, U.; Bloom, J.; Bornberg-Bauer, E.; Colwell, L.J.; de Koning, A.P.; Dokholyan, N.V.; Echave, J.; *et al.* The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci. Publ. Protein Soc.* **2012**, *21*, 769–785.
44. Wang, G.Z.; Chen, L.L.; Zhang, H.Y. Neighboring-site effects of amino acid mutation. *Biochem. Biophys. Res. Commun.* **2007**, *353*, 531–534.
45. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
46. Nishizawa, M.; Nishizawa, K. Local-scale repetitiveness in amino acid use in eukaryote protein sequences: A genomic factor in protein evolution. *Proteins Struct. Funct. Bioinform.* **1999**, *37*, 284–292.
47. Mularoni, L.; Veitia, R.A.; Alba, M.M. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **2007**, *89*, 316–325.
48. Haerty, W.; Golding, G.B. Genome-wide evidence for selection acting on single amino acid repeats. *Genome Res.* **2010**, *20*, 755–760.
49. Schaefer, M.H.; Wanker, E.E.; Andrade-Navarro, M.A. Evolution and function of cag/polyglutamine repeats in protein–protein interaction networks. *Nucleic Acids Res.* **2012**, *40*, 4273–4287.
50. Toll-Riera, M.; Radó-Trilla, N.; Martys, F.; Albà, M.M. Role of low-complexity sequences in the formation of novel protein coding sequences. *Mol. Biol. Evol.* **2012**, *29*, 883–886.
51. Katti, M.V.; Sami-Subbu, R.; Ranjekar, P.K.; Gupta, V.S. Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Sci.* **2000**, *9*, 1203–1209.
52. Miseta, A.; Csutora, P. Relationship between the occurrence of cysteine in proteins and the complexity of organisms. *Mol. Biol. Evol.* **2000**, *17*, 1232–1239.
53. Giddu, S.; Xu, F.; Nanda, V. Sequence recombination improves target specificity in a redesigned collagen peptide abc-type heterotrimer. *Proteins Struct. Funct. Bioinform.* **2013**, *81*, 386–393.
54. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **2014**, *42*, D222–D230.
55. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; *et al.* Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **2000**, *25*, 25–29.
56. Jiao, X.; Sherman, B.T.; Huang, D.W.; Stephens, R.; Baseler, M.W.; Lane, H.C.; Lempicki, R.A. David-ws: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **2012**, *28*, 1805–1806.
57. DAVID. Available online: http://david.abcc.ncifcrf.gov/home.jsp (accessed on 27 August 2014).
58. Eisenberg, D.; Weiss, R.M.; Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 140–144.

59. Mandel-Gutfreund, Y.; Gregoret, L.M. On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J. Mol. Biol.* **2002**, *323*, 453–461.
60. Misawa, K.; Kikuno, R.F. Evaluation of the effect of cpg hypermutability on human codon substitution. *Gene* **2009**, *431*, 18–22.
61. Duncan, B.K.; Miller, J.H. Mutagenic deamination of cytosine residues in DNA. *Nature* **1980**, *287*, 560–561.
62. Hubbard, T.; Barker, D.; Birney, E.; Cameron, G.; Chen, Y.; Clark, L.; Cox, T.; Cuff, J.; Curwen, V.; Down, T.; *et al.* The ensembl genome database project. *Nucleic Acids Res.* **2002**, *30*, 38–41.
63. Ensembl. Available online: http://www.ensembl.org/index.html (accessed on 27 August 2014).
64. Waterhouse, R.M.; Tegenfeldt, F.; Li, J.; Zdobnov, E.M.; Kriventseva, E.V. Orthodb: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **2013**, *41*, D358–D365.
65. OrthoDB. Available online: http://orthodb.org/orthodb7 (accessed on 27 August 2014).
66. Higgins, D.G.; Sharp, P.M. Clustal: A package for performing multiple sequence alignment on a microcomputer. *Gene* **1988**, *73*, 237–244.
67. Plotree, D.; Plotgram, D. PHYLIP-phylogeny inference package (version 3.2). *Cladistics* **1989**, *5*, 163–166.
68. Yang, Z. Paml: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **1997**, *13*, 555–556.
69. NCBI. Available online: http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi (accessed on 27 August 2014).
70. Yang, Z. *Computational molecular evolution*; Oxford University Press: New York, NY, USA, 2006; Volume 284.
71. Ma, L.; Zhang, T.; Huang, Z.; Jiang, X.; Tao, S. Patterns of nucleotides that flank substitutions in human orthologous genes. *BMC Genomics* **2010**, *11*, 416. doi:10.1186/1471-2164-11-416.
72. Cole, C.; Barber, J.D.; Barton, G.J. The jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **2008**, *36*, W197–W201.
73. Cuff, J.A.; Barton, G.J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins Struct. Funct. Bioinform.* **2000**, *40*, 502–511.
74. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.
75. Consortium, U. The universal protein resource (uniprot). *Nucleic Acids Res.* **2008**, *36*, D190–D195.
76. Advanced Jpred. Available online: http://www.compbio.dundee.ac.uk/www-jpred/advanced.html (accessed on 27 August 2014).
77. Horton, P.; Park, K.J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C.J.; Nakai, K. WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **2007**, *35*, W585–W587.
78. WoLF PSORT. Available online: http://www.genscript.com/psort/wolf_psort.html (accessed on 27 August 2014).

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).