Analyses of the general rule on residue pair frequencies in local amino acid sequences of soluble, ordered proteins

Matsuyuki Shiota^{1,2*} and Kengo Kinoshita^{1,2,3}

¹Department of Applied Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi 980–8579, Japan

²Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Miyagi 980–8573, Japan

³Department of *In Silico*, Institute for Development, Aging and Cancer, Tohoku University, Sendai, Miyagi 980–8575, Japan

Received 12 November 2012; Revised 26 January 2013; Accepted 14 March 2013

DOI: 10.1002/pro.2255

Published online 23 March 2013 proteinscience.org

Abstract: The amino acid sequences of soluble, ordered proteins with stable structures have evolved due to biological and physical requirements, thus distinguishing them from random sequences. Previous analyses have focused on extracting the features that frequently appear in protein substructures, such as α -helix and β -sheet, but the universal features of protein sequences have not been addressed. To clarify the differences between native protein sequences and random sequences, we analyzed 7368 soluble, ordered protein sequences, by inspecting the observed and expected occurrences of 400 amino acid pairs in local proximity, up to 10 residues along the sequence in comparison with their expected occurrence in random sequence. We found the trend that the hydrophobic residue pairs and the polar residue pairs are significantly decreased, whereas the pairs between a hydrophobic residue and a polar residue are increased. This trend was universally observed regardless of the secondary structure content but was not observed in protein sequences that include intrinsically disordered regions, indicating that it can be a general rule of protein foldability. The possible benefits of this rule are discussed from the viewpoints of protein aggregation and disorder, which are both caused by low-complexity regions of hydrophobic or polar residues.

Keywords: sequence analysis; protein structure; secondary structure; protein disorder

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Grant-in-Aid for Scientific Research on Innovative Areas; Grant number: 22136005; Grant sponsor: Grant-in-Aid for Scientific Research (C); Grant number: 24570176; Grant sponsor: Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology, Japan; Grant sponsor: Computational time was provided by the Super Computer System, Human Genome Center, The Institute of Medical Science, The University of Tokyo. This study was supported in part by MEXT Tohoku Medical Megabank Project.

*Correspondence to: Matsuyuki Shiota, Graduate School of Information Sciences, Tohoku University, 6-3-09, Aza-aoba, Aramaki, Aoba-ku, Sendai, Miyagi 980-8579, Japan. E-mail: mshiota@ecei.tohoku.ac.jp

Introduction

In aqueous conditions, most proteins spontaneously fold into their native states, which are determined by their amino acid sequences and particular physiological conditions.¹ The driving force in the folding of such soluble, ordered proteins is the hydrophobic effect, by which nonpolar side chains cluster inside the structure, whereas polar ones are located outside. To facilitate protein folding under aqueous conditions, an intrinsic balance between the hydrophobic and hydrophilic residues in the amino acid sequence is required, in order to accomplish the spatial segregation of these two types of residues.

The simplest and most studied feature that characterizes protein sequences, in contrast to

random sequences, is the amino acid composition or the relative frequencies of the 20 amino acids. The amino acid compositions in proteins are clearly different from the equal probabilities for the 20 amino acids, and the amino acid compositions of individual proteins significantly differ from each other. Some of the variations can be explained by biological factors, such as nutritional conditions² and environments,³ and by physical factors, such as secondary structures, protein size,⁴ and surface-to-volume ratio (SVR).⁵ Due to these differences, the amino acid composition is an important feature to consider when predicting the structure and function of an uncharacterized protein.^{6–8} In general, the amino acid compositions have evolved to satisfy the various physicochemical requirements that are necessary for the individual proteins to perform their functions.

Given the amino acid composition of each protein, the second simplest sequence feature is the amino acid pair frequencies, which are defined as the occurrence of two types of amino acids with a particular residue separation along the sequence. Protein structures include regular secondary structures, such as α -helices and β -strands, in which many of the frequently appearing patterns of two amino acids at particular separations have been reported. For example, α -helices have characteristic amino acid compositions at each residue position,^{9–12} and β -strands also have similar position-specific patterns, although their effects on the sequence are weaker, as compared with those of helices.^{13–15} Furthermore, α -helices and β -strands are also known to include periodic arrangements of hydrophobic and hydrophilic residues,^{16–18} and some of the loop regions have their own unique patterns.¹⁹ Although these previous reports included important information about natural protein sequences, they are insufficient for a statistical and comprehensive understanding of the amino acid pair frequencies, in terms of the following three aspects: (1) These reports focused on the amino acid pair propensities of particular substructures of proteins, but they did not reveal the common features in the amino acid sequences of soluble, ordered proteins, (2) previous analyses revealed the importance of several frequently occurring amino acid pairs in forming stable structure in solvents, but few studies have discussed the less frequently observed pairs, and (3) the effects of variations in the amino acid compositions of individual proteins on the amino acid pair frequencies were not discussed explicitly. The bottleneck in addressing these points was the lack of data for ordered protein sequences, because the number of proteins with 3D structures is much smaller than the number of protein sequences within the sequence databases.

Recently, the number of protein structures deposited in the Protein Data Bank²⁰ has been

increasing rapidly, representing the accomplishments of many structural genomics initiatives.^{21–23} This large body of structural data enables us to comprehensively discuss the abundance and the absence of amino acid pairs, as compared with their expected frequencies. Our analysis will help to clarify the characteristic features of protein sequences that distinguish them from random arrangements of the 20 types of amino acids.

In this report, we first focused on a nonredundant data set of protein sequences that fold into stable 3D structures. We then examined the trends in the frequencies of amino acid pairs, and discussed the physicochemical interactions that generate the trends. We also showed that the observed trend was limited to the ordered protein sequences, but was not found in the sequences that are not guaranteed to fold into a stable structure. Among the ordered protein sequences, we further discussed the effects of their secondary structure content.

Results and Discussion

Data sets of protein sequences

We downloaded 10,569 nonredundant amino acid sequences of protein domains from SCOP v1.75, in which the maximum sequence identity between any two sequences is below 40%.²⁴ From these sequences, we selected domains with structures solved by X-ray crystallography at a resolution better than 2.5 Å, so as to focus on the amino acid sequences of ordered proteins. Membrane proteins, which were identified either by having the MeSH term “Membrane Protein” or by the SOSUI program,²⁵ were excluded in order to focus on the sequence–structure relationship of soluble proteins. Our final dataset consisted of 7368 protein domains. From them, the amino acid sequences were obtained by reading the ATOM records, to exclude the regions without a stable structure. In addition, any short terminal sequences resembling His-tags were eliminated from the sequences. We referred to this data set as the “Ordered” set.

Local co-occurrence scores of amino acid pairs

We first analyzed the frequency of an amino acid pair between a and b , observed in the “Ordered” data set, along the sequence with a particular separation k . The abundance and the absence of residue pairs were estimated by the co-occurrence score (CoS),

$$\text{CoS}(a, b, k) = \ln \frac{F(a, b, k)}{E(a, b, k)}, \quad (1)$$

where $F(a, b, k)$ and $E(a, b, k)$ are the observed and expected frequencies, respectively, of the amino acid pair a and b separated by k residues, with

consideration of the order of the amino acid pair. In other words, $F(a, b, k)$ is different from $F(b, a, k)$. The expected frequency in the data set was calculated as the sum of the expected frequencies in individual proteins, by assuming that the expected frequency of a residue pair in a protein will be proportional to the product of the relative occurrences of each residue in the protein (see Materials and Methods section for details.)

Although $\text{CoS}(a, b, k)$ can provide comprehensive information about the amino acid pair preferences, the number of combinations is large (i.e., 400 amino acid pairs times the number of possible separations), and thus it is difficult to obtain general trends about amino acid sequences from these scores. To decrease the number of combinations and to focus on the local sequence preference, we averaged the CoS values over 1–10 residue separations, and defined the local CoS (LCoS) as

$$\text{LCoS}(a, b) = \frac{1}{10} \sum_{k=1}^{10} \text{CoS}(a, b, k). \quad (2)$$

Positive values of LCoS indicate that the two specified residues are abundant in the local proximity up to 10 residue separations. We will first discuss the abundance and the absence of the 400 residue pairs by using LCoS, and then we will observe the preferences at each separation, by using CoS for several notable pairs or groups of pairs.

Figure 1 illustrates the LCoSs for 400 amino acid pairs. In this figure, we roughly divided the pairs into three groups: hydrophobic–hydrophobic (HH, left), polar–polar (PP, middle), and hydrophobic–polar (HP, right), under the assumption that the following nine residues (A, C, M, I, L, V, F, Y, and W) were hydrophobic, and the others were polar. For a more detailed discussion, the 20 amino acids were color-coded according to their physicochemical properties: the nine hydrophobic residues are colored green, the positively charged residues (H, K, and R) are blue, the negatively charged ones (D and E) are red, G and P are cyan, and the noncharged polar residues (S, T, N, and Q) are magenta.

The figure revealed the general trend that the frequencies of the HH pairs and the PP pairs are decreased from their expected values (i.e., negative LCoSs), whereas most of the local HP pairs are found more frequently than expected (positive LCoSs). The average and standard deviations of LCoS values of the HH, PP, and HP pairs were -0.033 ± 0.030 , -0.017 ± 0.028 , and $+0.019 \pm 0.029$, respectively, and the three groups showed significantly different distributions from each other ($P < 0.01$ by Wilcoxon test for all combinations of distributions). If we simulated a data set of random protein sequences by permutating randomly the

residues in each protein and calculated the LCoS value for each amino acid pair from the randomized sequence data set, the average and standard deviations of LCoSs of the HH, PP and HP pairs were much smaller, -0.0004 ± 0.0006 , -0.0002 ± 0.0003 , and -0.0002 ± 0.0002 , respectively, indicating that most of the observed LCoS values are significant in the native protein sequences. These results indicated that the mixing of hydrophobic and polar residues was favored, whereas the clustering of residues from the same group was avoided, in the amino acid sequences of the “Ordered” proteins during the course of evolution.

The suppression of the local occurrences of HH pairs may imply a strategy for preventing protein aggregation. Since protein aggregation is related to various pathological conditions,²⁶ the sequence mutations that enhance the features to promote aggregation will be a disadvantage for survival, and thus be suppressed in evolution.^{27,28} For example, amino acid sequences reportedly have safeguards against aggregation, by placing residues that prevent aggregation, such as arginine, lysine, and proline, in the sites flanking long stretches of hydrophobic residues.^{29,30}

The contribution of each hydrophobic residue to the decrease in local HH pairs was estimated by comparing the distribution of LCoSs of those HH pairs which includes at least one residue of the corresponding residue type. Since we defined nine residues as hydrophobic, each residue was included in 17 HH pairs, consisting of $8 \times 2 = 16$ heterogeneous pairs and one identical pair, and the median LCoS of these 17 pairs was compared. As results, isoleucine (-0.057), phenylalanine (-0.052), tryptophan (-0.052), and leucine (-0.045) strongly contributed to the trend of decrease in HH pairs, indicating that strong hydrophobicity is responsible for the decrease in HH pairs in local proximity.

In a similar way, the decreased occurrence of local PP pairs represents an important feature of proteins with stable structures, when we focus on the consistency with the sequence features of intrinsically disordered proteins.³¹ The disordered regions are characterized by a low content of hydrophobic residues (or a high content of polar and charged residues),^{32,33} and low complexity due to highly repeated residues.^{34,35} These features are consistent with our observation of the suppressed local clustering of polar and charged residues in structured proteins.

The reason why the local proximity of HP pairs is increased in the amino acid sequences may be to compensate for the suppressed HH and PP pairs, but protein structures may have local features that promote the mixing of H and P amino acids. For example, regular secondary structures, such as helices and strands, have hydrophobic and polar

Figure 1. Local co-occurrence scores (LCoS) of the 400 amino acid pairs. The pairs are evenly divided into 40 columns on the x-axis and arranged such that the HH, PP and HP pairs are sorted from left to right. The y-axis indicates the LCoS value of each pair. The amino acids are color coded, as described in the text. The identical residue pairs are highlighted with arrows, and the regions enclosed by dashed or dotted lines are discussed in the text.

residues that periodically appear in turn, in small globular proteins. Another example is the supersecondary structure, in which regular secondary structure elements (SSEs) with hydrophobic residues are followed by a loop region with polar residues.

Structural backgrounds of frequent HH and PP pairs

Despite the general trend of negative LCoS values for HH and PP pairs, some had positive LCoSs. As shown in Figure 1, most of the HH pairs with positive LCoSs contained alanine or cysteine. These residues have relatively low hydrophobicity, as compared with the other hydrophobic residues, which may explain their different trends from the other typical HH pairs. On the other hand, the PP pairs with positive LCoSs included various types of residues. The sequence–structure relationship of these exceptional PP pairs is interesting, because they may play important roles in the stability of protein structures.

First, two histidines (Column 8, LCoS 0.066) is the most favored pair among the PP pairs. They are reportedly included in many local motifs, such as zinc fingers, and their important roles in functional sites explain their exceptional abundance in protein structures.

Among the pairs between the remaining four charged residues (D, E, R, and K), the eight pairs of opposing charges were found significantly more frequently than those of the same charge (Fig. 1,

Columns 8–11 and 14, dashed rectangles for different charge pairs and dotted ones for the same charge pairs, $P < 0.01$ by Wilcoxon test). This observation is physically reasonable, because counter-charged pairs are electrically favorable.³⁶ However, it is quite interesting that the opposing charged pairs had directionality in protein structures. In other words, the negative charge (D and E) frequently occurred at the N-terminus, whereas the positive one (R and K) was at the C-terminus (Fig. 1, Columns 9–11, dashed rectangles. For example, see the difference between EK and KE). A close inspection of these residue pairs, using CoSs from 1 to 10 sequence separations, revealed that these pairs frequently appeared at 1, 3, 4, and 7 separations [Fig. 2(A)], which suggested the periodicity of α -helices. In α -helices, the charged residues reportedly have a distinctly asymmetric distribution, due to the interaction between the charged side chains and the helix dipole moments³⁷; glutamic acid and aspartic acid are favored near the N-termini to interact with the main-chain nitrogens and lysines, and arginines are favored near the C-termini to interact with the main-chain oxygens.^{9,10} Thus, α -helices may impose an ordered arrangement of charged residues on protein sequences.

We also found other types of pairs with a directional preference; that is, PE (Column 13, dashed circle) had a positive LCoS of 0.023, whereas EP (Column 12, dashed circle) had a large negative LCoS of -0.102 . The analysis of the CoS values at

Figure 2. The co-occurrence propensities for (A) EK, KR, and their counterparts, (B) PE and EP, and (C) GS, GT, and their counterparts. The amino acid pairs in the opposing directions (e.g., EK and KE) are depicted by a solid line (the more frequent one) and a dashed line (the less frequent one) with the same color.

each separation revealed that PE was more favored than EP at any distance from 1 to 10, with the largest difference occurring at a separation of 1 [Fig. 2(B)]. This difference may be explained in terms of the terminal motifs of an α -helix. Although proline cannot form regular backbone hydrogen bonds in helices, it is strongly favored as either the first residue of the helix N-terminus or the residue just after the C-terminus, whereas on the other hand, glutamic acid is favored inside an α -helix near the N-terminus, as described.¹⁰ Thus, proline was frequently followed by glutamic acid, but the inverse pair was observed less, as a result of the combination of the two known sequence preferences.

GT and GS are another example of frequent PP pairs (Fig. 1, Column 17, dashed circle). They are much more favored than their opposing counterparts, SG (Column 18) and TG (Column 20). The analysis of these pairs using CoSs revealed that for most of the separations from 2 to 10, GT and GS occurred more frequently than TG and SG [Fig. 2(C)]. When secondary structures were assigned to these pairs, the majority of these residue pairs were involved in loop regions. Thus, the GT and GS pairs may play an important role in the stability of loop regions, as compared to their opposing counterparts, although their atomic details have not been clarified.

Suppression of identical pairs in ordered proteins

In the data set of soluble, ordered proteins, 12 of the 20 identical pairs had negative LCoSs, with a median LCoS value of -0.021 . The suppression of the identical pairs, however, may not necessarily be observed in protein sequences that are not guaranteed to fold into stable structures.

To address this point, we constructed four additional protein sequence data sets, for comparison to the “Ordered” sequences. (1) The “Disordered” data set was obtained from the DisProt database.³⁸ It consists of 653 protein sequences that were demonstrated to be intrinsically disordered; that is, not to

form stable structures. (2) The “UniRef” data set is a nonredundant subset from the UniProt database. It consists of 3,939,736 nonredundant protein sequences, among which each two share at most 50% sequence identity.³⁹ Datasets (3) “Human” and (4) “*Escherichia coli*” include 34,421 and 4190 protein sequences of each species, respectively, which were obtained from the RefSeq database.⁴⁰

Figure 3 illustrates the boxplots of the LCoSs for the 20 identical pairs, which were calculated from the five data sets. The results revealed that in the “Disordered” sequences, the identical pairs were increased, except for tryptophan, whereas cysteine, methionine, glycine, and proline, along with the charged residues, had large LCoSs, whereas the other hydrophobic residues, such as I, L, Y, F, and W, had smaller values. These high frequencies of local identical pairs result from the repetitive and low complexity regions included in disordered sequences.^{35,36} The “UniRef” data set also showed increased occurrences of identical pairs. A protein sequence in this data set can consist of a mixture of ordered and disordered regions. In such a sequence, the hydrophobic residues and polar residues will be concentrated in the ordered and disordered regions, respectively, and thereby the local co-occurrences of both types of identical pairs become increased. A similar trend was observed in the “Human” data set, in which about 40% of the sequences are predicted to contain long disordered regions.⁴¹ In contrast, the fractions of disordered regions in bacterial proteins are reportedly low, and the LCoSs of identical pairs in the “*E. coli*” data set were significantly lower than those of the other three sequence data sets. These results indicated that the suppression of the same residues is strongly related to the foldability of the amino acid sequences.

Effects of SSEs

SSEs, such as helices and strands, have periodic patterns of hydrophobic and polar residues. These patterns lead to the periodicity of the side chain

Figure 3. The LCoSs of the 20 identical pairs in the five data sets, depicted by boxplots. The thick solid lines indicate the median LCoSs. The top and bottom of each box indicate the first and third quartiles, respectively. The whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box.

orientations in these elements: residues with similar hydrophobicity appear at each three or four residues for helix segments and at each two residues for strand segments. To evaluate the effect of the SSEs on the residue co-occurrence in the amino acid sequence, the “Ordered” data set was further divided into all- α , all- β and α/β subsets, as follows. Secondary structures were determined by using the DSSP program.⁴² We calculated the α -fraction of a protein, as the ratio of the number of residues in helices (DSSP code “H,” “G,” or “I”) and the total number of residues in either helices or sheets (DSSP code “E” or “B”). The proteins with an α -fraction ≥ 0.8 were classified as all- α , whereas those with an α -fraction ≤ 0.2 were classified as all- β . Other proteins were classified as α/β . Overall, we obtained 1478 all- α , 977 all- β , and 4913 α/β proteins.

We divided the 400 residue pairs into three groups, according to the rough classifications shown in Figure 1, thus generating 81 HH (green), 121 PP (red), and 198 HP (black) pairs. Figure 4(A–D) shows the average CoSs of the three groups, as functions of residue separation. The results of the entire data set [Fig. 4(A)], which included all 7368 proteins regardless of their SSE contents, showed a strong oscillation of the average CoSs for each group, as a function of separation. This oscillation indicated a strong effect on the entire data set from the α -helix segments, in which residues with similar characteristics appear at separations of 3, 4, and 7. Nevertheless, HH pairs, as well as PP pairs, are essentially avoided in the local sequence, as shown by the LCoS values (dashed lines), which indicate the effects of similar pair suppression in the local proximity. As

expected, we observed similar patterns for the α/β proteins, because the data set is the mixture of α helices and β sheets [Fig. 4(D)].

The α -protein subset [Fig. 4(B)] and the β -protein subset [Fig. 4(C)] revealed quite different CoS patterns. The results for the α proteins indicated stronger oscillations of CoS values compatible with an α -helix than those for the entire data set, and the favorable interactions of HH and PP pairs in forming helices seemed to overcome their tendency to be suppressed at separations of 3, 4, and 7 [Fig. 4(B)]. On the other hand, since hydrophobic and polar residues are frequently aligned one after another in β -sheets, we expected to observe the co-occurrence of HH and PP pairs in each even number of residue separations for β proteins, but the results showed that an increased occurrence was only observed at a 2 residue separation [Fig. 4(C)]. There are two possible explanations for the limited effect of β -sheets. One is the smaller fraction of residues forming SSEs in β proteins than in α proteins: the average fraction of helices in the α -protein subset is 65%, whereas that of strands in the β -protein subset is 48% ($P < 0.001$, by Wilcoxon test). The other explanation is that a long stretch of alternating patterns of hydrophobic and polar residues is not favored in a protein sequence, because such a region would cause aggregation.^{43,44} In short, the difference between the amino acid sequences of α and β proteins lies at the separations from 2 to 4, where the effects of the different local periodicities are clearly observed.

We then evaluated the effect of SSEs, by calculating the Spearman’s correlation coefficient between α and β proteins using the CoSs of the 400 pairs at each separation [Fig. 4(E)]. As a result, the separations were divided into three groups, according to the correlation values: positive correlations at 1, 5, and 6, negative ones at 2, 3, and 4, and no correlation from 7 to 10. The positive correlations coincide with the separations where the HH pairs and the PP pairs were suppressed in both the α and β proteins. The negative correlations confirmed the previous result that α -helices and β -strands had opposing effects on residue co-occurrence at the 2, 3, and 4 separations. If these CoSs were averaged over 1–10 separations to yield LCoSs, then the correlation between α and β proteins would be 0.55, which is higher than the correlations between the CoS values at any separation. As shown in the scatter plot of the LCoSs for α and β proteins [Fig. 4(F)], there was a tendency for the HH and PP pairs to decrease and the HP pairs to increase, regardless of the SSE types of the proteins. Therefore, we concluded that the avoidance of the local proximity of HH and PP pairs is a general feature in the amino acid sequences of soluble, ordered proteins, regardless of the SSE content of each protein.

Figure 4. Secondary structure-dependent co-occurrence scores (CoSs). The four graphs illustrate, from top to bottom, the results for (A) the entire data set, (B) the α -protein subset, (C) the β -protein subset, and (D) the α/β protein subset. The CoSs averaged over the residues in the HH, PP, and HP groups are depicted by green, red, and black solid lines, respectively. The average LCoSs of the three groups are shown by gray dotted lines. (E) The Spearman's correlation of CoSs between α and β proteins. (F) The scatter plot of LCoSs between α and β proteins.

Effect of residue separation of local proximity

In this study, the threshold of 10 residue separation for local proximity was chosen because we considered 10 residues are far enough to average out the effect of local structures such as secondary structures, turns and other functional motifs. The scarcity of signal past seven residues in Figure 4 suggests that shorter separations for local proximity can better discriminate HH, PP, and HP pairs. We examined the effect of threshold distance on LCoSs of residue pairs by varying the threshold from 3 to 13 separations (Supporting Information Figure 1). The result shows that the trend that HH and PP pairs are decreased and HP pairs are increased is observed the most significantly at threshold of six residue separation (HH: -0.047 ± 0.043 , PP: -0.027 ± 0.041 , and HP: 0.028 ± 0.042). The average LCoS of each group attenuates past six residue separations but the three groups still remain separated at 13 residue separation. Thus, the decrease in local HH and PP pairs is a general property of amino acid sequences of structured proteins, which can be observed at a wide range of threshold separations.

Effect of protein size and SVR

Our "Ordered" data set includes protein sequences of various sizes and various fractions of residues exposed to solvent in the surface and these structural features may influence local pairing propensities of residue pairs. To address this point, we

calculated LCoSs of amino acid pairs in protein subsets of different sizes (numbers of residues) or different SVRs⁵: in general, smaller proteins have larger SVRs (Supporting Information Tables I and II). As results, we did not observe any significant change in LCoSs of the HH, PP, or HP pairs if protein size is over 100 residues or if SVR is below $0.35 \text{ \AA}^{-1}$. In the subsets of the smallest size or the largest SVR proteins, significant decreases of LCoSs of HP pairs compared with those calculated from the whole data set were observed. These decreases, however, were not coupled with the decreases in the observed frequencies of local HP pairs, but they can be regarded as artifacts of small sample sizes of the two subsets (see Supporting Information Tables I and II and Supporting Information Discussion section). Thus, the local pairing patterns of amino acid pairs in HH, PP, and HP groups are not sensitive to protein structural features, such as protein size or SVR.

Methods

Here, we describe the calculation of the CoS for an amino acid pair a and b , separated by k residues, from a data set of protein sequences.

First, for a given protein sequence p , we defined $f(a, b, k|p)$ as the occurrence of amino acid pair (a, b) with separation k . The total occurrence of the amino acid pair (a, b) at separation k in the whole data set is then given by $F(a, b, k) = \sum_p f(a, b, k|p)$, where the summation is taken over all of the proteins in the

data set. On the other hand, the expected occurrence of residue pair (a,b) at separation k in p with a random distribution is given by:

$$e(a,b,k|p) = \begin{cases} N(k|p) \times \frac{f(a|p) \times f(b|p)}{N(p) \times (N(p)-1)} & \text{for } a \neq b \\ N(k|p) \times \frac{f(a|p) \times (f(a|p)-1)}{N(p) \times (N(p)-1)} & \text{for } a=b \end{cases}, \quad (3)$$

where we define $f(a|p)$ as the occurrence of amino acid a , $N(p) = \sum_a f(a|p)$ as the total number of amino acids in p , and $N(k|p) = \sum_a \sum_b f(a,b,k|p)$ as the total number of amino acid pairs with separation k in p , where the two summations are taken over 20 amino acids and over 400 amino acid pairs, respectively. The total expected occurrence in the data set is then given by

$$E(a,b,k) = \sum_p e(a,b,k|p).$$

Finally, the CoS value of a residue pair is defined by the log ratio between the observed and expected occurrences, as in Eq. (1),

$$\text{CoS}(a,b,k) = \ln \frac{F(a,b,k)}{E(a,b,k)}. \quad (1')$$

Note that CoSs (and LCoSs) are defined relative to the occurrences expected from the amino acid compositions of individual proteins (i.e., $f(a|p)$ and $f(b|p)$). An alternative method to estimate the expected occurrences is to calculate from the amino acid composition of the entire data set,

$$E_{\text{DB}}(a,b,k) = N(k) \times p(a) \times p(b),$$

where $N(k) = \sum_p N(k|p)$ is the total number of residue pairs separated at k , and $p(a)$ and $p(b)$ are the relative frequencies of amino acids a and b in the entire data set. The CoS in this case is then defined by,

$$\text{CoS}_{\text{DB}}(a,b,k) = \ln \frac{F(a,b,k)}{E_{\text{DB}}(a,b,k)},$$

and the local CoS by,

$$\text{LCoS}_{\text{DB}}(a,b) = \sum_{k=1}^{10} \text{CoS}_{\text{DB}}(a,b,k). \quad (4)$$

The former estimation, $\text{CoS}(a,b,k)$, assumes that amino acids are randomly distributed *within proteins*, but does not assume any distribution of the amino acid frequencies in individual proteins. On the other hand, the latter one, $\text{CoS}_{\text{DB}}(a,b,k)$, assumes that amino acids are distributed randomly

both *within and between proteins* in the data set. In other words, the amino acid compositions of individual proteins are randomly determined from the amino acid composition in the entire data base; that is, $p(a)$. However, this assumption does not hold true, because many proteins have distinct amino acid compositions.²⁻⁵ Therefore, we used $\text{CoS}(a,b,k)$, in which the differences in the amino acid compositions are normalized.

Conclusions

We performed a comprehensive analysis of the abundances and absences of amino acid pairs in the local sequence proximity in soluble, ordered proteins. We observed a general trend that the pairs between two hydrophobic residues or those between two polar residues were decreased, and those between a hydrophobic residue and a polar one were increased, from their expected occurrences in the local proximity, up to a 10 residue separation. The suppression of the HH pairs and PP pairs was considered to be a consequence of avoiding the aggregation-prone regions and the disorder-promoting regions, respectively. Protein structures possess many local structural patterns, such as secondary structures and supersecondary structures, which promote the increased occurrence of HP pairs in local sequence proximity. We also confirmed that this trend was universally observed in proteins with different SSE contents, and that the low occurrence of HH and PP pairs was a characteristic feature observed in ordered regions but not in general protein sequences with a mixture of ordered and disordered regions. Therefore, we conclude that the suppression of HH and PP pairs is a universal feature in soluble, ordered proteins that fold into stable 3D structures.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230.
2. Hughes AL (2010) Evolutionary conservation of amino acid composition in paralogous insect vitellogenins. *Gene* 467:35–40.
3. Goodarzi H, Torabi N, Najafabadi HS, Archetti M (2008) Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene* 407:30–41.
4. Bastolla U, Demetrius L (2005) Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng Des Sel* 18: 405–415.
5. Shirota M, Ishida T, Kinoshita K (2008) Effects of surface-to-volume ratio of proteins on hydrophilic residues: decrease in occurrence and increase in buried fraction. *Protein Sci* 17:1596–1602.
6. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S (2002) Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* 319:1257–1265.

7. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S (2003) Prediction of human protein function according to gene ontology categories. *Bioinformatics* 19:635–642.
8. Roy S, Martinez D, Platero H, Lane T, Werner-Washburne M (2009) Exploiting amino acid composition for predicting protein–protein interactions. *PLoS One* 4:e7813.
9. Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* 240:1648–1652.
10. Kumar S, Bansal M (1998) Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins. *Proteins* 31:460–476.
11. Engel DE, DeGrado WF (2004) Amino acid propensities are position-dependent throughout the length of alpha-helices. *J Mol Biol* 337:1195–1205.
12. Fonseca NA, Camacho R, Magalhaes AL (2008) Amino acid pairing at the N- and C-termini of helical segments in proteins. *Proteins* 70:188–196.
13. Goliaei B, Minuchehr Z (2003) Exceptional pairs of amino acid neighbors in alpha-helices. *FEBS Lett* 537:121–127.
14. Fooks HM, Martin AC, Woolfson DN, Sessions RB, Hutchinson EG (2006) Amino acid pairing preferences in parallel beta-sheets in proteins. *J Mol Biol* 356:32–44.
15. Bhattacharjee N, Biswas P (2010) Position-specific propensities of amino acids in the beta-strand. *BMC Struct Biol* 10:29.
16. West MW, Hecht MH (1995) Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci* 4:2032–2039.
17. Xiong H, Buckwalter BL, Shieh HM, Hecht MH (1995) Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc Natl Acad Sci USA* 92:6349–6353.
18. Mandel-Gutfreund Y, Gregoret LM (2002) On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J Mol Biol* 323:453–461.
19. Crasto CJ, Feng J (2001) Sequence codes for extended conformation: a neighbor-dependent sequence analysis of loops in proteins. *Proteins* 42:399–413.
20. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35:D301–D303.
21. Terwilliger TC (2011) The success of structural genomics. *J Struct Funct Genomics* 12:43–44.
22. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C (2009) PSI-2: structural genomics to cover protein domain family space. *Structure* 17:869–881.
23. Weigelt J (2010) Structural genomics-impact on biomedicine and drug discovery. *Exp Cell Res* 316:1332–1338.
24. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res* 32:D189–D192.
25. Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14:378–379.
26. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 75:333–366.
27. Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep* 8:737–742.
28. Monsellier E, Ramazzotti M, de Laureto PP, Tartaglia GG, Taddei N, Fontana A, Vendruscolo M, Chiti F (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys J* 93:4382–4391.
29. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2009) Protein sequences encode safeguards against aggregation. *Hum Mutat* 30:431–437.
30. Rousseau F, Serrano L, Schymkowitz JW (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* 355:1037–1047.
31. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J Mol Biol* 293:321–331.
32. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001) Intrinsically disordered protein. *J Mol Graph Model* 19:26–59.
33. Uversky VN (2011) Intrinsically disordered proteins from A to Z. *Int J Biochem Cell Biol* 43:1090–1103.
34. Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 25:847–855.
35. Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 5:2985–2995.
36. Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41:415–427.
37. Hol WG (1985) The role of the alpha-helix dipole in protein function and structure. *Prog Biophys Mol Biol* 45:149–195.
38. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35:D786–D793.
39. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23:1282–1288.
40. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 40:D130–D135.
41. Pentony MM, Jones DT (2010) Modularity of intrinsic disorder in the human proteome. *Proteins* 78:212–221.
42. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
43. Broome BM, Hecht MH (2000) Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J Mol Biol* 296:961–968.
44. Ventura S (2005) Sequence determinants of protein aggregation: tools to increase protein solubility. *Microb Cell Fact* 4:11.