JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 12, 2010

Pp. 1679–1687
DOI: 10.1089/cmb.2008.0035

Substitution Matrices of Residue Triplets Derived from Protein Blocks

XIN LIU and YA-PU ZHAO

ABSTRACT

In protein sequence alignment, residue similarity is usually evaluated by substitution matrix, which scores all possible exchanges of one amino acid with another. Several matrices are widely used in sequence alignment, including PAM matrices derived from homologous sequence and BLOSUM matrices derived from aligned segments of BLOCKS. However, most matrices have not addressed the high-order residue-residue interactions that are vital to the bioproperties of protein. With consideration for the inherent correlation in residue triplet, we present a new scoring scheme for sequence alignment. Protein sequence is treated as overlapping and successive 3-residue segments. Two edge residues of a triplet are clustered into hydrophobic or polar categories, respectively. Protein sequence is then rewritten into triplet sequence with $2\times20\times2=80$ alphabets. Using a traditional approach, we construct a new scoring scheme named TLESUM$_{hp}$ (TripLEt SUbstitution Matrices with hydropobic and polar information) for pairwise substitution of triplets, which characterizes the similarity of residue triplets. The applications of this matrix led to marked improvements in multiple sequence alignment and in searching structurally alike residue segments. The reason for the occurrence of the "twilight zone," i.e., structure explosion of low identity sequences, is also discussed.

Key words: hydrophobicity, sequence alignment, twilight zone.

1. INTRODUCTION

SIMILARITY OF AMINO ACID IS THE BASIS OF PROTEIN SEQUENCE ALIGNMENT, protein design, and protein structure/function prediction. The mutation data matrices of Dayhoff and Eck (1968) and the substitution matrices of Henikoff and Henikoff (1992) are not only standard choices for the evaluation of residue similarity, but also the basis of amino acid classification (Murphy et al., 2000; Liu et al., 2002; Li et al., 2003). Efforts have been made to optimize these matrices by executing the iterative approach (Gonnet et al., 1992), including evolutionary information (Koshi and Goldstein, 1995). Usually, in generating these scoring schemes, homologous sequences were aligned column by column. Residues in each of the column were deemed to mutate to or substitute for each other during the evolution. The counts of residue mutations or substitutions observed in the database of aligned sequences were analyzed by statistical approaches, by which the scores of residue similarity were derived. Though widely used in protein studies, few of these scoring schemes have gone beyond the consideration of mono-residue substitution. Namely, it was assumed that residues in different positions of a polypeptide mutate independently of each other.

State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing, China.

However, it is known that there are complicated residue-residue interactions in protein. These high-order interactions are essential to the protein molecule. So the hypothesis of mono-residue substitution is not an optimized choice. A more useful approach is investigating the local high-order interactions of protein sequence by k consecutive letters (k-word). As residue-residue correlations are biologically meaningful, the k-word scheme has been used successfully in component analysis of nucleic acid and protein sequences (Karlin and Ghandou, 1985) and in the study of molecular phylogeny (Qi et al., 2004; Hao and Qi, 2004), for example. However, a 20^k dimension vector is required to characterize the word type in such an approach. Owing to the huge dimension size, a complete investigation is nearly impossible. A coarse-grained approach is the one and only choice. Using coarse-grained k-word, research has been performed, including protein secondary structure prediction (Zheng, 2004) and remote homologues detection (Ogul and Mumcuoglu, 2007).

As the physicochemical properties are alike for some naturally occurring amino acids (Mathews and Van Holde, 1995), the size of the residue alphabet could be reduced by grouping similar residues together. The feasibility of amino acid classification was supported by both experimental and theoretical approaches. Several experiments have shown that some proteins designed with fewer than 20 types of residues can maintain native structures of natural proteins (Regan and Degrado, 1988; Kamtekear et al., 1993; Davidson et al., 1995; Riddle et al., 1997). There were also several theoretical approaches focusing on residue classification, based on the substitution propensity of residues (Murphy et al., 2000; Liu et al., 2002; Li et al., 2003; Fan and Wang, 2003), interaction potential (Wang and Wang, 1999; Liu et al., 2002), and other properties (Gorban et al., 2010). Among the various classification schemes, a well-known and meaningful scheme is clustering residue alphabets into hydrophobic/polar groups. The importance of hydrophobic interaction owes much to the following two facts: (i) it is the driving force for protein folding (Dill, 1990; Li et al., 1997); and (ii) it is an important factor for protein-protein interaction (Jones and Thornton, 1996; Young et al., 1994). According to research on amino acid classification, it has been proven that the hydrophobic/polar feature is the dominant factor in clustering residue alphabets into two categories.

In this article, we introduce a residue clustering scheme for the representation of the residue triplet, which decreases the dimension of phase space drastically. We then derive the substitution matrices of the residue triplet from the counts of 3-residue pairwise substitutions based on the aligned residue segments in the BLOCKS9 database (Henikoff and Henikoff, 1991). The capability of such scoring scheme is examined in multiple sequence alignment and secondary structure identification.

2. METHODS

2.1. Database

In 1992, Henikoff and Henikoff (1992) obtained their amino acid substitution matrices (BLOSUM, blocks substitution matrix) from the BLOCKS database. This high-quality database is based on local sequence alignment, and derived from the homologous proteins in PROSITE catalog (Bairoch, 1991) by PROTOMAT algorithm (Henikoff and Henikoff, 1991). The BLOCKS database contains the most highly conserved regions (involving biologically significant sites, patterns, and profiles) of related proteins. In order to set up a general scoring scheme, we used the BLOCKS9 database (published December 1995) in matrices construction.

In the BLOCKS database, a group of ungapped multiple aligned segments is called a ''block,'' with each row a different protein segment and each column an aligned residue position. A block represents a conserved region of a protein family. In total, 3179 blocks are involved in BLOCKS9 database.

2.2. Residue classification scheme

In 1996, based on a database of native protein structure, Miyazawa and Jernigan (1996) derived their knowledge-based potential from the frequencies of structural contacts between different amino acids. This knowledge system provides the basis of many works in folding mechanism (Li et al., 1996, 1997) and residue classification (Wang and Wang, 1999; Liu et al., 2002).

In our approach, a 2-letter scheme based on the Miyazawa-Jernigan matrix is adopted in amino acid classification (hydrophobic h = {M, F, I, L, V, A, W}, polar p = {C, Y, Q, H, P, G, T, S, N, R, K, D, E}). According to a former work (Liu et al., 2002), this classification scheme has a strong correlation with residue hydrophobicity (Branden and Tooze, 1991). Actually, it has been shown by several researchers that there is

no large difference among various clustering schemes when residue alphabets are grouped into two categories. So, different selections of clustering schemes should have limited impact on our results. Therefore, the properties derived are expected to be robust. This is the first merit of a two-categories clustering scheme. The other consideration is sample counts. In our approach, approximately $200\times\xi^4$ elements of a matrix were estimated, where ξ is the category count of a clustering scheme. Sample counts decrease drastically with the increase of ξ, especially at the low homologous level. A two-categories clustering scheme makes the sample counts abundant enough for statistical approach.

2.3. Construction of triplet substitution matrix

We attempted to construct the scoring scheme for the similarities among k-words. As parameter k is large, the resulting scheme would be more specific. But the corresponding sample counts per matrix element are lower, which results in difficult statistical analysis. A good tradeoff is the adoption of a residue triplet. In our approach, each protein sequence was treated as successive triplets of amino acids. In each of the 3-residues segment, two neighbors of the central residue were mapped into h/p letters. With the $2\times20\times2=80$ letters alphabet set, protein sequences were rewritten into triplet sequence.

Generally, samples in a block are biased; in other words, many segments are closely related to each other. To reduce the bias, similar members were clustered within blocks, and each cluster was weighted as a single sequence in data counting. A parameter $\Theta\%$ was specified as the threshold of sequence identity. Residue segments that are identical for at least that percentage were grouped together within a block. Consequently, one matrix characterizes the similarity of triplet among sequences below a certain homologous level.

It was considered that triplets in a column can substitute with each other in protein evolution. We counted all possible pairs of triplet substitutions in each column of every block. All these counts were summed. The result of this counting is a frequency table listing the counts for each of the $80+79+..1=3240$ different triplet pairs that occur in the BLOCKS9 database. The table was then used to calculate a matrix representing the log-odds ratio between these observed frequencies and those expected by chance.

We denote the total number of triplet pair i, j $(1\le j\le i\le 80)$ by f_{ij}. Then the observed probability of the occurrence of pair i, j is

$$q_{ij}=f_{ij}/\sum_{i=1}^{80}\sum_{j=1}^{i}f_{ij} \tag{1}$$

The probability for triplet i to occur is then

$$p_i=q_{ii}+\sum_{j\neq i}q_{ij}/2 \tag{2}$$

The expected probability e_{ij} of occurrence of pair i, j is then p_ip_j for $i=j$ and $p_ip_j+p_jp_i=2p_ip_j$ for $i\neq j$. An odds ratio matrix is calculated where each entry is q_{ij}/e_{ij}. The logarithm of odds ratio is defined as lod ratio, which characterizes the deviation of the sample counts observed in BLOCK9 from those of the background. We calculated the lod ratio in half bit units as $s_{ij}=2\log_2(q_{ij}/e_{ij})$, which is rounded to the nearest integer value to produce TripLEt SUbstitution Matrices with hydropobic and polar information (TLESUM$_{hp}$ Θ).

For more details on matrix deriving, see Henikoff and Henikoff (1992).

3. RESULTS

In the TLESUM$_{hp}$ scoring scheme, the residue triplet is the unit of substitution. To the best of our knowledge, there is no similar scoring scheme; therefore, a direct comparison with existing methods is not possible. So we carried out a comparative evaluation of TLESUM$_{hp}$ by its performance in sequence alignment. The well-known BLOSUM62 matrix was used as a reference.

3.1. Multiple sequence alignment

Here we show the power of TLESUM$_{hp}$ in multiple sequence alignment by an example of single block motif search. We aligned a set of helix-turn-helix (HTH) proteins provided in 1993 by Lawrence (Lawrence

et al., 1993). There are thirty sequences in this data set. Each of them contains a HTH motif for DNA-binding involved in gene regulation. It has been found that there is a 18-residue common pattern in every sequence. Due to the varieties in residue composition and position, however, it is a challenge to identify the common pattern with the sequence alignment approach.

In 2005, Zheng searched this common motif with the BLOSUM62 scoring scheme (Zheng, 2005). Using a center-star approach, two segments with high similarity score were defined as close neighbors. A motif was regarded as a group of close neighbors in which one member shares the greatest similarity with the rest of the members. Taking each sliding windows of width 18 from every sequence as a seed, the string most similar to the seed is searched in each sequence other than the one that the seed is in. Namely, for seed $S_{i,l+1}S_{i,l+2}\ldots S_{i,l+18}$ in protein i, the highest scored segment in protein j ($i \neq j$; near neighbor of the seed) that has a similarity score $\phi_{ij,l} = \sum_{k=1}^{18} C(S_{i,l+k}, S_{j,m+k})$ not less than 10 bits is searched, where $C(x, y)$ is the xy entry in scoring scheme, and l, m are segment indices in sequence i and j, respectively. The score of a seed/center-star is defined as the score sum of its near neighbors, $Z_{il} = \sum_{j=1, j\neq i}^{30} \Phi(\phi_{ij,l})$, where $\Phi(x)$ is the step function with $\Phi(x) = x$ for $x \geq 10$ and $\Phi(x) = 0$ otherwise.

According to Zheng's work, using BLOSUM62, 23 near neighbors (including one incorrect case) was found in TOP1 star tree, i.e., the tree formed by the near neighbors identified by the highest scored seed and the seed itself. By rewriting the residue sequence with the triplet alphabet set and employing the TLESUM$_{hp}$62 scoring scheme, 29 near neighbors (only one protein is omitted) are found by the TOP1 star tree of this center-star approach. And all these near neighbors are HTH motifs (i.e., true signals).

3.2. Secondary structure identification

Homologous proteins usually share the same protein fold. Therefore, the ability of a scoring scheme could be evaluated by its performance in identifying structurally identical protein sequences. It is expected that, in a set of segments collected by sequence alignment, a more sensitive scoring scheme will find a higher proportion of members that share the same conformation. Namely, at a similar level of false signal noise, a good scheme can identify more true signals.

To evaluate the performance of TLESUM$_{hp}$62, we calculated the all-against-all pairwise similarity scores for 10-width segments in a nonredundant set. In this data set, 1612 nonmembrane proteins from PDB_SELECT25 (Hobohm and Sander, 1994) are collected, and no pair of sequences share sequence identity of more than 25%. Given two segments X and Y that have a similarity score $\psi_{XY} = \sum_i C(X_i, Y_i)$ more than threshold $T_{TLESUM_{hp}}$, if their secondary structure representations $\omega(X_i)$ and $\omega(Y_i)$ are the same in any column i of the alignment, then there is one count of a "True Positive" (TP) sample; otherwise, there is a "False Positive" (FP) one. The secondary structures were taken from the DSSP database (Kabsch and Sander, 1983). According to this representation, there are eight types of protein secondary structures defined with the hydrogen bond: H, G, I, E, X, T, S, and B. As in most methods, we considered 3 states $\{h, e, c\}$ generated from the 8 by the coarse-graining $H, G, I \rightarrow h$ for helices, $E \rightarrow e$ for strands and $X, T, S, B \rightarrow c$ for coils.

A similar approach was performed with BLOSUM62 matrix. At certain threshold T_{BLOSUM}, we obtained the counts of TP and FP samples. Then by varying threshold $T_{TLESUM_{hp}}$, we tuned the total counts of FP to be equal for both schemes. Consequently, we can evaluate the performance of a scheme by the improvement in True Positive counts. For example, as threshold $T_{BLOSUM} = 23$ units, we collected 86601 TP and 1389714 FP events with BLOSUM62 matrix. In next step, TLESUM$_{hp}$62 detected 1403259 FP samples after an adjustment of the threshold $T_{TLESUM_{hp}}$. At a nearly equal FP level, TLESUM$_{hp}$62 detected 100644 TP samples, gained an increase of nearly 16.2% compared with that of BLOSUM62 scheme.

The results of secondary structure identification are shown in Figure 1. Compared with BLOSUM62, the improvement contributed by TLESUM$_{hp}$62 is remarkable. The TP counts increase 9.2-16.2%. As shown in the insertion of Figure 1, the FP/TP ratio decreases as TLESUM$_{hp}$62 is used. Furthermore, in the counts of FP, the cases with only tiny structural discrepancy increase 8.8–19.8%. For example, when the counts of FP is about 244000, the proportion of FP with single mismatch in secondary structure are 22027 and 18381 for TLESUM$_{hp}$62 and BLOSUM62, respectively. The present scheme achieved an increase of 19.8 percents. Therefore with an adoption of TLESUM$_{hp}$62 matrix, the population of pairwise segment alignment shifts towards the region with fewer mismatches.

In a detail analysis of the improvement, we found that TLESUM$_{hp}$62 is more sensitive to samples at low identity level. As $T_{BLOSUM} = 23$, in the TP segment pairs obtained by BLOSUM62, 44.4 percentages

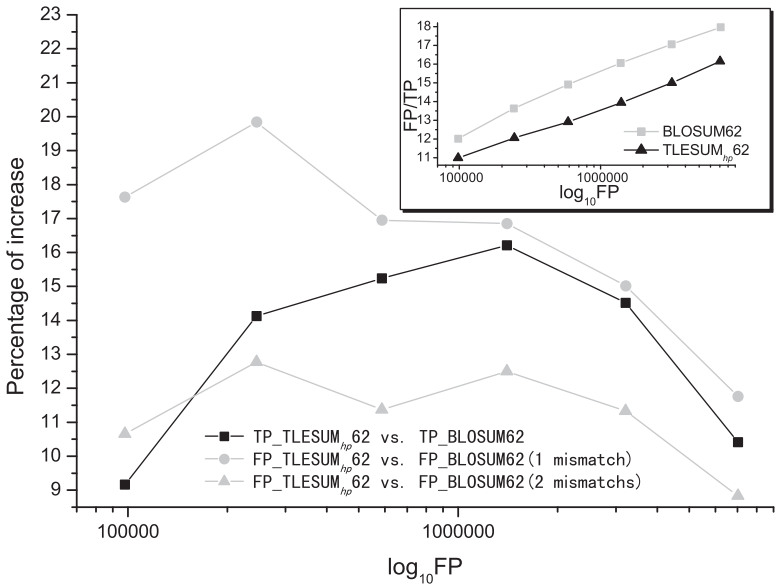

FIG. 1. Results of secondary structure identification.

(38464/86601) of them have no more than 4 identical residues. Whereas, at a nearly equal FP level, 75 percentages (75474/100644) of TP samples collected by TLESUM$_{hp}$62 meet this low identity standard.

4. DISCUSSION

We have constructed a series of triplet substitution matrices from the BLOCKS9 database. Since three-order residue-residue interaction is considered, these matrices are expected to be more sensitive. Improvements in sequence alignment, protein design, and protein structure/function prediction will be attained by existing methods after the modification of using TLESUM$_{hp}$ matrices.

Insertion or deletion is important for sequence alignment. As the size of TLESUM$_{hp}$ is about 16 times of that of ordinary matrices, it is not convenient to evaluate performance before a modification of the existing tools. But once a residue sequence is rewritten into triplet sequence, a gap can certainly be introduced. It does not matter whether a gap represents a triplet or a residue.

We found that, with the variety of sequence identity level, the propensity of the transitions among two coarse-grained residues of triplet (TCGRT) alters accordingly. For example, the effect of p*h to h*p transition can be evaluated by mean $m(p^*h \rightarrow h^*p) = \frac{1}{20}\Sigma_{\Omega=1}^{20} TLESUM_{hp}\Theta(p\Omega h, h\Omega p)$, where Ω stands for central residue. The effect of h*h to p*p can be defined as $m(h * h \rightarrow p * p)$ in a similar way. According to TLESUM$_{hp}$95, the ratio $m(p * h \rightarrow h * p)/m(h * h \rightarrow p * p)$ is 1.85. But for the TLESUM$_{hp}$30 matrix, the ratio decreases to 1.11. Namely, the propensity of h*h to p*p transition increases at low identity level.

As the extension of BLOSUM matrices, TLESUM$_{hp}$ should correlate with BLOSUM. But in a direct score comparison, there are various differences between the two types of schemes. In the BLOSUM62 scheme, the substitution score between F and W is 3 units (half-bits). As shown by the examples in Table 1, a mismatch of the TCGRT induces a notable score shift in the TLESUM$_{hp}$ matrices, though the central residues are identical. In TLESUM$_{hp}$62, the substitution score between hFh and pWp is -3 units. It has the maximum deviation compared to the 3 units evaluated by BLOSUM62 (i.e., a decrease of 6 units). Moreover, as shown in Table 1, the score deviation induced by sequence identity threshold Θ is obvious. So the complexity of comparison is beyond the reach of a simple observation. There is a need for mathematical analysis.

To investigate the correlation between the two types of matrices, a theoretical approach—eigenvalue decomposition analysis—is applied to these matrices. According to this mathematical analysis, the components of a matrix can be ranked based on their significance. Different matrices can be compared by the most important components of them. In this way, the most significant similarity/dissimilarity among different matrices could be revealed. For a $N \times N$ real symmetric matrix M, the element of the matrix can be reconstructed as

Table 1. The Score of Triplet Substitution as F and W Are Central Residues

Triplet pair	*TLESUM$_{hp}$30*	*TLESUM$_{hp}$62*	*TLESUM$_{hp}$95*
hFh ↔ hWh	4	**4**	4
hFh ↔ hWp	1	2	1
hFh ↔ pWh	2	0	−1
hFh ↔ pWp	*−10*	−3	*−4*
hFp ↔ hWh	0	1	1
hFp ↔ hWp	1	**3**	3
hFp ↔ pWh	−1	−2	−4
hFp ↔ pWp	*−5*	−1	*−1*
pFh ↔ hWh	1	1	1
pFh ↔ hWp	0	−2	−3
pFh ↔ pWh	0	**3**	3
pFh ↔ pWp	1	0	−1
pFp ↔ hWh	*−1*	−2	*−5*
pFp ↔ hWp	−3	0	−3
pFp ↔ pWh	−4	0	−2
pFp ↔ pWp	1	**2**	2

In each subunit, the score changes induced by miss matches of the two coarse-grained residues are notable. For example, when the two coarse-grained residues are matchable (hFh ↔ hWh, hFp ↔ hWp, pFh ↔ pWh, and pFp ↔ pWp; their scores are shown in bold), the substitution scores evaluated by TLESUM$_{hp}$62 are 4, 3, 3, and 2 units, respectively, in a similar level to the score C(F, W) = 3(units) evaluated by BLOSUM62 scheme. Otherwise, the scores deviate distinctly. As the scores of different TLESUM$_{hp}\Theta$ matrices are compared, the changes are also notable. The top three high deviation cases between TLESUM$_{hp}$30 and TLESUM$_{hp}$95 are hFh ↔ pWp, hFp ↔ pWp, and pFp ↔ hWh, and are shown in italic.

$$M_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha V_{\alpha,i} V_{\alpha,j}, \tag{3}$$

where M_{ij} is the element of the matrix in row i and column j, λ_α is the αth eigenvalue, and $V_{\alpha,\,i}$ is the ith component of the αth eigenvector, $\mathbf{V}_\alpha = (V_{\alpha,\,i})$. According to the absolute values, eigenvalues are sorted in a descending order. The item given by the top eigenvector, $\lambda_1 V_{1,\,i} V_{1,\,j}$ has the largest contribution to element M_{ij}. In order to uncover the most significant similarity/dissimilarity, we focus on the relationships between the first eigenvectors of the two types of matrices.

After subtracting the mean of the corresponding matrix from each element, eigenvalue decomposition analysis is applied to each matrix. For matrix TLESUM$_{hp}\Theta$ ($\Theta = 30, 35, \ldots, 95$), according to types of the TCGRT (h_h, h_p, p_h, p_p), the 80 components of an eigenvector are grouped into four subsets. Each subset can be deemed a 20-dimension vector and compared with the eigenvector obtained by the BLOSUMΘ matrix. Correlation coefficient r is calculated as $r = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}$, $l_{xy} = \sum_{i=1}^{20} (x_i - \overline{x})(y_i - \overline{y})$, where $\overline{z} = \frac{1}{20} \sum_{i=1}^{20} z_i$, i is the residue type, x is a vector given by a subset of the first eigenvector's components of the TLESUM$_{hp}\Theta$ matrix, and y is the first eigenvector of the BLOSUMΘ matrix. It was found that, for different choices of subset and sequence identity level Θ, $|r|$ ranges from 0.74 to 1. So the two types of matrices are indeed tightly related. According to the first eigenvectors, the difference between the two types of matrices owes much to the introduction of TCGRT. There are obvious concert shifts in the value of the first eigenvector's components induced by TCGRT. It indicates that the differences among the four types of TCGRT would be critical in reducing the 80-component eigenvectors of TLESUM to 20-component vectors, to compare against the BLOSUM eigenvectors.

In the first eigenvector of TLESUM$_{hp}\Theta$ ($\Theta > 30$), for any central residue Ω, components of hΩp and pΩh are nearly equal. Taking the two components as references, there are obvious value shifts in the components of hΩh and pΩp, up and down, respectively. For each TLESUM$_{hp}$ matrix, the first eigenvalue is positive. As item $\lambda_1 V_{1,i} V_{1,j} = \frac{\lambda_1}{2}[V_{1,i}^2 + V_{1,j}^2 - (V_{1,i} - V_{1,j})^2]$ has the largest contribution in reconstructing matrix element M_{ij}, the smaller the difference between $V_{1,i}$ and $V_{1,j}$, the greater is the positive value attributed to the element. As $V_{1,h\Omega p} \approx V_{1,p\Omega h}$, mutations between h_p and p_h are conserved or may be positively favored as $\Theta > 30$. In a similar approach, we find that interchange $h\_h \leftrightarrow p\_p$ is recommended as $\Theta \leq 30$. This transition of the type of favored interchange can explain the reason for the occurrence of the "twilight zone" for sequence alignment.

The "twilight zone" (Doolittle, 1986) is a crucial barrier to nearly all scientists who attempt to get information about protein structure and function by the sequence comparison method. Most protein pairs that have more than 30 out of 100 identical residues share similar protein structures (Sander and Schneider, 1991). Below this identity level, conventional sequence comparison methods often fail to align protein sequences. When two totally unrelated sequences composed of the 20 standard amino acids are aligned without any introduced gaps, random chance leads to about 6 percentages of identical residues ($\ll$30%). And according to the observation of Rost, the emergence of the "twilight zone" seems not to be based solely on statistics (Rost, 1999). Although a detailed description of structural deviation in the "twilight zone" has been described by Chung and Subbiah (1996), the reason for its occurrence is still not clear.

Here we introduce Li's HP model (Li et al., 1996) to describe the cause of the "twilight zone." In their model, a protein is treated as a self-avoiding chain of beads placed on a discrete lattice. Two types of beads are used to mimic polar (p) and hydrophobic (h) residues. The energy of a sequence folded into a certain structure is given by short-range contact interactions

$$H = \sum_{i<j} E_{\sigma_i \sigma_j} \Delta(r_i - r_j) \tag{4}$$

where $\Delta(r_i - r_j) = 1$ if beads r_i and r_j are adjoining lattice sites, but the two beads are not adjacent in position along the sequence, and $\Delta(r_i - r_j) = 0$ otherwise; σ_i is either h or p. Depending on the types of beads in contact, the interaction energies are evaluated as $E_{hh} = -2.3$, $E_{hp} = -1$, or $E_{pp} = 0$, corresponding to hh, hp, or pp contacts, respectively.

In Li's 2D model (Li et al., 1996), surrounding each bead of the core, there are two resultful neighbors that are not adjacent to the bead in position along the sequence. A bead pair have $2^2 \times 2^2$ kinds of surrounding neighbors in total. For neighbor type μ, we can calculate the energy $H_\mu^{A\_B}$ of the local structure, because bead pair A_B are surrounded by μ. When A_B changes to A'_B', there will be an energy modification $|\Delta H(\mu, A\_B \leftrightarrow A'\_B')| = |H_\mu^{A\_B} - H_\mu^{A'\_B'}|$, which reflects the impact to energy induced by interchange.

According to our analysis, interchange of h_p and p_h is recommended as sequence identity is >30%. As shown in Table 2, when they interchange with each other, only $|\Delta H| = 0 \sim 0.6$ rises in energy modification ($|\Delta H| = 0 \sim 1.2$ for 3D model). The tiny energy deviation does not impact homologous proteins significantly. Consequently, proteins tend to keep a similar structure. This is the reason that sequences at such an identity level are expected to have similar structure or function. But as sequence identity is less than 30%, interchange of h_h and p_p is favored. A large energy modification $|\Delta H| = 4 \sim 5.2$ is induced by such an interchange ($|\Delta H| = 8 \sim 10.4$ for 3D model). The low identity level further makes such mutations happen

TABLE 2. THE IMPACTS TO ENERGY H INDUCED BY $h\_p \leftrightarrow p\_h$ AND $h\_h \leftrightarrow p\_p$ INTERCHANGES

Neighbor categories	$\lvert\Delta H(h\_p \leftrightarrow p\_h)\rvert$	$\lvert\Delta H(h\_h \leftrightarrow p\_p)\rvert$
hhhh	0	5.2
hhhp	0.3	4.9
hhph	0.3	4.9
hhpp	0.6	4.6
hphh	0.3	4.9
hphp	0	4.6
hpph	0	4.6
hppp	0.3	4.3
phhh	0.3	4.9
phhp	0	4.6
phph	0	4.6
phpp	0.3	4.3
pphh	0.6	4.6
pphp	0.3	4.3
ppph	0.3	4.3
pppp	0	4

In Li's two-dimensional HP model, every residue in the core have two resultful neighbors that are not adjacent to the residue in position along sequence. Two coupled sites have 2^4 kinds of surrounding neighbors. The impact for every neighbor category is presented here.

frequently. Due to the large impact on energy, difficulty in preserving protein structure is drastically increased. This can explain the occurrence of the so-called ''twilight zone'' for sequence homology.

Details of the eigenvalue decomposition analysis of $TLESUM_{hp}$ matrices will be published elsewhere.

ACKNOWLEDGMENTS

We are grateful to professor Lu-Hua Lai for her helpful discussions. This work is supported in part by the National Natural Science Foundation of China (no. 10704077) and National Basic Research Program of China (973 Program, grant no. 2007CB310504).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bairoch, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* 19, 2241–2245.

Branden, C., and Tooze, J. 1991. *Introduction to Protein Structure*. Garland Publishing, New York.

Chung, S.Y., and Subbiah, S. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure* 4, 1123–1127.

Davidson, A.R., Lumb, K.J., and Sauer, R.T. 1995. Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* 2, 856–863.

Dayhoff, M.O., and Eck, R.V. 1968. *Atlas of Protein Sequence and Structure. Vol. 3.* National Biomedical Research Foundation, Silver Springs, MD.

Dill, K.A. 1990. Dominant forces in protein folding. *Biochemistry* 29, 7133–7155.

Doolittle, R.F. 1986. *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acide Sequences*. University Science Books, Mill Valley, CA.

Fan, K., and Wang, W. 2003. What is the minimum number of letters required to fold a protein? *J. Mol. Biol.* 328, 921–926.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. 1992. Exhaustive matching of the entire protein sequence database. *Science* 256, 1433–1445.

Gorban, A.N., Kudryashev, M., and Popova, T. 2010. Informational way to protein alphabet: entropic classification of amino acids. Available at: http://arxiv.org/abs/q-bio/0501019. Accessed October 1, 2010.

Hao, B., and Qi, J. 2004. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *J. Bioinform. Comput. Biol.* 2, 1–19.

Henikoff, S., and Henikoff, J.G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19, 6565–6572.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 10915–10919.

Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Sci.* 3, 522–524.

Jones, S., and Thornton, J.M. 1996. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* 93, 13–20.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.

Kamtekear, S., Schiffer, J.M., Xiong, H, et al. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science* 265, 1680–1685.

Karlin, S., and Ghandou, G. 1985. Comparative statistics for DNA and protein sequences: single sequence analysis. *Proc. Natl. Acad. Sci. USA* 82, 5800–5804.

Koshi, J.M., and Goldstein, R.A. 1995. Context-dependent optimal substitution matrices. *Protein Eng.* 8, 641–645.

Lawrence, C.E., Altschul, S., Boguski, M., et al. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.

Li, H., Helling, R., Tang, C., et al. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* 273, 666–669.

Li, H., Tang, C., and Wingreen, N.S. 1997. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Lett.* 79, 765–768.

Li, T., Fan, K., Wang, J., et al. 2003. Reduction of protein sequence complexity by residue grouping. *Protein Eng.* 16, 323–330.

Liu, X., Liu, D., Qi, J., et al. 2002. Simplified amino acid alphabets based on deviation of conditional probability from random background. *Phys. Rev. E* 66, 021906.

Mathews, C.K., and Van Holde, K.E. 1995. *Biochemistry*. Benjamin Cumming, San Francisco.

Miyazawa, S., and Jernigan, R.L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* 256, 623–644.

Murphy, L.R., Wallqvist, A., and Levy, R.M. 2000. Simplied amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* 13, 149–152.

Ogul, H., and Mumcuoglu, E.U. 2007. A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets, *Biosystems* 87, 75–81.

Qi, J., Wang, B., and Hao, B. 2004. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.* 58, 1–11.

Regan, L., and Degrado, W.F. 1988. Characterization of a helical protein. designed from first principles. *Science* 241, 976–978.

Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., et al. 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* 4, 805–809.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12, 85–94.

Sander, C., and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignement. *Proteins* 9, 56–68.

Wang, J., and Wang, W. 1999. A computational approach to simplifying the protein folding alphabet. *Nat. Struct. Biol.* 6, 1033–1038.

Young, L., Jernigan, R.L., and Covell, D.G. 1994. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* 3, 717–729.

Zheng, W.M. 2004. Clustering of amino acids for protein secondary structure prediction. *J. Bioinform. Comput. Biol.* 2, 333–342.

Zheng, W.M. 2005. Relation between weight matrix and substitution matrix: motif search by similarity. *Bioinformatics* 21, 938–943.

Address correspondence to:
Dr. Ya-Pu Zhao
State Key Laboratory of Nonlinear Mechanics
Institute of Mechanics
Chinese Academy of Sciences
No. 15 Beisihuanxi Road
Beijing, 100190, China

E-mail: yzhao@imech.ac.cn