Major factors of protein evolution revealed by eigenvalue decomposition analysis

Xin Liu[1]
[1]Institute of Mechanics
Chinese Academy of Sciences
Beijing, China

Li-Mei Zhang[2], **Jun Yin**[1], **Ya-Pu Zhao**[1]
[2]School of Science
Beijing Jiaotong University
Beijing, China

Abstract *Here we attempt to characterize protein evolution by its dominant factors. These factors are revealed by top eigenvectors in the spectrums of eigenvalue decomposition analysis. To reduce the bias induced by closely related sequences in the database, we introduce a parameter, sequence identity by which proteins with sequence identity below certain level are involved in analysis. It is found that, with drop of sequence identity level, residue feature mainly conserved in mutation changes from hydrophobicity to volume. The transition point is at sequence identity* $\approx 45\%$. *As residue hydrophobicity no longer governs residue substitution, it is a doubt whether importance of hydrophobic interaction declines in conserving the family representative properties among remote homologues. So, we also investigate the contribution of hydrophobic interaction in near and remote homologues. In aligned homologues, hydrophobic interaction systems inbuilt in these proteins are aligned too; and can be deemed to be similar and substitutable with each other. With a comparison of aligned hydrophobic interaction systems, we can catch the representative character of hydrophobic interaction for the corresponding protein family. Here top weighted feature in the substitution of hydrophobic interaction systems is revealed as a function of sequence identity. It is found that a shift happens to the type of physical quantity which governs the substitution of hydrophobic interaction. The number of hydrophobic residue is the dominantly unchangeable part in aligned hydrophobic interactions as sequence identity* $> 30\%$. *Below this point, state of internal hydrophobic force which characterizes the residue-residue pairwise interaction is primarily conserved. With view of this shift, intrinsic requirement of protein evolution is sought in the discussion section.*

Keywords: Protein evolution; hydrophobic interaction; Eigenvalue decomposition analysis; sequence analysis and alignment; remote homologues

1 Introduction

Though efforts have ever been made to reveal the nature of protein evolution[1, 2, 3, 4, 5], major factors conserved in the evolvement is still obscure. A clarification of such factors will assist us in comprehending the development of protein universe, designing artificial proteins, and etc. Here, by a revealment of such factors, we attempt to present some outlines of protein evolution to readers.

Residue mutation is vital for protein evolution. Information obtained by the analysis of aligned homologous sequences can give us the significant characters of protein evolution. In numerous sequence alignment based works, the knowledge system BLOSUM(blocks substitution matrix) given by Henikoff and Henikoff[6] is typical, and provides us a basis for further investigation. In their approach, analysis was based on local sequence alignments in a high quality database-BLOCKS[7] where the most highly conserved regions of related proteins in PROSITE[8] catalog were collected. Aligned residues were deemed to be substitutable one another. To characterize substitutability/similarity of residues, statistics of residue substitution were converted into a log-odds ratio between a combined model and an independent one. In order to reduce the bias induced by closely related sequences in the database, the level of sequence identity was introduced as a parameter in the clustering of homologous proteins. Sequences that were identical for at least certain level were grouped together, and weighted as a single sequence in data counting. In this way, non-redundant samples, i.e. proteins which were alike below certain identity level were analyzed. Consequently, one matrix characterizes the

propensity of residue substitution in sequences below a certain sequence identity level. Further analysis of this scoring scheme can outline the governing residue features in the substitution of amino acid.

In this work, we firstly identity the governing feature in residue substitution by analyzing Henikoff's BLOSUM matrices with a general method, eigenvalue decomposition. It is revealed that there is an intrinsic transition point at sequence identity $\approx 45\%$ where the dominant feature related to residue substitution shifts. Hydrophobicity[9] and residue volume[10] are two significant residue features in protein evolution. Hydrophobicity is the major conserved residue feature as sequence identity is above this transition point. Whereas, residue volume is a governing feature below this point.

Hydrophobic interaction is confirmed as the dominant driving force for protein folding[11, 12]. Since hydrophobicity is not the governing residue feature in amino acid substitution of remote homologous proteins, it is a doubt whether importance of hydrophobic interaction declines in conserving the family representative properties among remote homologues. So, we are also interest in the contribution of hydrophobic interaction in sequences of different homologous levels.

Though there is no difference in physical mechanism, every protein has its distinctive character in hydrophobic interaction. In a protein family, the hydrophobic interaction systems of different homologues can be deemed to be similar in some way. Once homologues are aligned site by site, hydrophobic interaction systems inbuilt in these proteins are aligned too. The respective character of each system can be compared one another. Difference exists in such comparison, especially for remote homologues. Factors conserved in such comparison should be essential for the family representative character of hydrophobic interaction. Using a coarse-grained model, we construct a new scoring scheme named TLESUM$_{hp}$(TripLEt SUbstitution Matrices with hydrophobic and polar information). According to the forementioned reason, sequence identity is introduced as a parameter. Eigenvalue decomposition analysis of these matrices achieves a new understanding of the top weighted feature in the similarity of aligned hydrophobic interaction systems. It is found that number of hydrophobic residue and internal force given by residue hydrophobicity are two physical quantities which govern the substitution of localized hydrophobic interaction at different identity stages. When residue-residue pairwise interaction is investigated in a protein molecule, a force contributed from hydrophobicity exists between each residue pair. As remote homologues are compared, it is significant to maintain the similarity between force states of a protein and those of other molecules. Whereas in near homologues, the most important feature in conserving the similarity of aligned hydrophobic interaction systems is keeping an equal number of hydrophobic residue approximately. Significance and implication of this observed shift are discussed. The major conclusions are:

1. For remote homologues, hydrophobic interaction is still vital in conserving family representative properties, but changes its aspect of emphasis, from conservation of the number of hydrophobic residue to the preservation of family specific hydrophobic force network.

2. Intramolecular hydrophobic force network has a crucial contribution to the representative hydrophobic interaction of a protein family, meets the intrinsic requirement of protein evolution, and help the conservation of family representative biological properties significantly.

2 Materials and methods

2.1 Eigenvalue decomposition analysis

In an eigenvalue decomposition approach, a given $N \times N$ real symmetric matrix M can be reconstructed as

$$M_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha V_{\alpha,i} V_{\alpha,j} \qquad (1)$$

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. In this way, information involved in a matrix would be decomposed and ranked according to significance. Item given by the top eigenvector, $\lambda_1 V_{1,i} V_{1,j}$ has the largest contribution to element M_{ij}.

A matrix is analyzed after subtracting its mean from each element of the matrix. Such a subtraction procedure is necessary to remove a trivial source of a large eigenvalue[12]. Any matrix with a nonzero mean m_0 can have one dominant eigenvalue proportional to Nm_0 if the dimension N of the matrix is large. Removing this trivial regularity enables us to clearly identify other intrinsic regularities which could be obscured in the spectrum of the unsubtracted matrix.

In eigenvalue decomposition approach, item

$$\lambda_1 V_{1,i} V_{1,j} = \frac{\lambda_1}{2}[V_{1,i}^2 + V_{1,j}^2 - (V_{1,i} - V_{1,j})^2] \quad (2)$$

has the largest contribution in reconstructing matrix element M_{ij}. If the first eigenvalue λ_1 is positive, the fewer the difference between $V_{1,i}$ and $V_{1,j}$, the more positive value is contributed to the element M_{ij}. In a scoring scheme, large value of a matrix element means the large propensity in substitution. Consequently, mutations with few difference in the component of $\mathbf{V}_1$ are conserved or may be positively favored.

2.2 Construction of triplet substitution matrices

We treat protein sequence as successive triplet words of amino acid. A Miyazawa-Jernigan matrix[13] based 2-letter scheme is employed in the classification of amino acid(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 the former work, this classification scheme[14] has a strong correlation with residue hydrophobicity). Two neighbors of the central residue are mapped into h/p letters. Consequently, the original $20 \times 20 \times 20$ types of triplet are clustered into $2 \times 20 \times 2$ alphabets. Residue sequences are rewritten with the 80 letters alphabet set.

The aligned sequences in BLOCKS9 database[7] are used in our approach. In this database, a group of ungapped multiple aligned residue segments is called a block with each row a different protein segment and each column an aligned residue position. A single block represents a conserved region of a protein family. Totally, 3179 blocks are involved in BLOCKS9 database. According to the same reason of the construction of BLOSUM matrices, sequence identity is introduced as a parameter in segment clustering.

It is considered that triplet words in a column can substitute with each other in protein evolution. We count all possible pairs of triplet word substitutions in each column of every block. All these counts are summed. The result of this counting is a frequency table listing the number of times for each of the $80+79+..1 = 3240$ different triplet word pairs to occur in the BLOCKS9 database. The table is then used to calculate a 80×80 symmetric matrix representing the log-odds ratio between these observed frequencies and those expected by chance. We denote the total number of triplet word pair $i, j (1 \leq j \leq i \leq 80)$ by σ_{ij}. Then the observed probability of the occurrence of pair i, j is $q_{ij} = \sigma_{ij} / \sum_{i=1}^{80} \sum_{j=1}^{i} \sigma_{ij}$. The probability for triplet word i to occur is then $p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2$. The expected probability e_{ij} of occurrence of pair i, j is then $p_i p_j$ for $i = j$ and $p_i p_j + p_j p_i = 2 p_i p_j$ for $i \neq j$. A lod ratio is then calculated 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$_{hp}$).

3 Results

3.1 Views obtained by the analysis of BLOSUM matrices

In BLOSUM matrices, 90% contributions are made from the first 9-14 eigenvalues. Here we focus on the first eigenvalues which contribute $\sim 1/5$ to the total eigenvalue. Due to the positive values, components of these eigenvectors are dominantly conserved or may be positively favored in each BLOSUM matrix.

Table 1. Correlation coefficients r between the first eigenvectors of BLOSUM matrices and several residue features. Chou-Fasman's strand propensity correlates to hydrophobicity with r=0.59, and to residue volume with r=0.52. Molecular weight correlates to hydrophobicity with r=-0.25, and to residue volume with r=0.92.

Matrices	Residue features			
	hydrophobicity	residue volumn	strand propensity	molecular weight
BLOSUM95	0.77	0.54	0.84	0.75
BLOSUM90	0.77	0.54	0.84	0.73
BLOSUM85	0.77	0.54	0.84	0.74
BLOSUM80	0.77	0.54	0.84	0.76
BLOSUM75	0.77	0.54	0.84	0.77
BLOSUM70	0.77	0.53	0.84	0.77
BLOSUM65	0.77	0.54	0.83	0.77
BLOSUM60	0.77	0.54	0.84	0.78
BLOSUM55	0.76	0.57	0.84	0.79
BLOSUM50	0.75	0.58	0.84	0.78
BLOSUM45	0.71	0.64	0.83	0.69
BLOSUM40	0.59	0.74	0.80	0.58
BLOSUM35	0.47	0.75	0.73	0.26
BLOSUM30	0.31	0.73	0.57	0.22

To illustrate the meaning of these eigenvectors, we study the linear regression between eigenvector V_α and vector R_κ which is a 20-dimension vector introduced as a representation of residue feature κ, where κ refers to hydrophobicity, residue volume, secondary structure propensity[15, 16], and etc. Correlation coefficient r is calculated as

$$r = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}, \qquad l_{xy} = \sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y}) \quad (3)$$

where m is the vector dimension, $\overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$. The obtained correlation coefficients are shown in table 1.

It is found that there is a transition point at sequence identity $\approx 45\%$ where a switch occurs for the

dominant related residue feature. Hydrophobicity has a strong relationship with the principal eigenvector as sequence identity is above this transition point. However, below this point, residue volume is tightly related[17].

We want to point out that many residue features are inherently correlated. There are other features(correlating to both hydrophobicity and residue volume) which are more related to these eigenvectors. To make an indubitable analysis, we select two orthogonal vectors as the presentive features(The correlation coefficient between hydrophobicity and residue volume ≈ 0).

3.2 Views obtained by the analysis of TLESUM$_{hp}$ matrices

We focus on the comparison of hydrophobic interaction systems at local level. Protein sequence is treated as successive triplet words of amino acid. Each triplet owns a subset of the hydrophobic interaction of corresponding protein. As the basic element of residue-residue interaction is residue pair, focusing on such fundamental element, we investigate features of hydrophobic interaction between a pair of residues. Two side residues of triplet provide a hydrophobic environment for the central one. In aligned homologous sequence, the inbuilt hydrophobic environments are aligned column by column. Hydrophobic environments in aligned samples are deemed to be substitutable with each other. Similarity/substitutability of these hydrophobic environments can be observed in similarities of triplet words as they interchange with each others in aligned homologous proteins.

Scoring scheme TLESUM$_{hp}$ is introduced to evaluate triplet's similarities. In this scheme, two neighbors of the central residue are classified into hydrophobic(h) or polar(p) groups respectively, and provide four kinds of coarse-grained hydrophobic environments(h_h, h_p, p_h, p_p) for the center, where '_' stands for the center residue. Counts of samples for a triplet pair will be sparse if no residue clustering is employed. But excessive clustering will make the scheme less sensitive. As the central residue of triplet is not degenerated, distinctiveness of localized hydrophobic interaction is kept moderately.

To compare the contribution of hydrophobic environment among sequences which are alike in identity meaning below certain level, 14 matrices with different levels of sequence clustering are analyzed by eigenvalue decomposition approach. In a 80 dimensional eigenvector $\mathbf{V}_\alpha$, each type of triplet corresponds to a component of this vector. For central residue Ω, there are four relevant components, hΩh, hΩp, pΩh, and pΩp. Since the central residues are identical, difference of values in these components are due to type difference of hydrophobic environment. So, in case of central residue Ω, contribution of hydrophobic environment can be described as the following unit vector $\mathbf{C}_{\alpha\Omega}$

$$C_{\alpha\Omega,k} = \frac{V_{\alpha,\Omega k} - \overline{V_{\alpha,\Omega}}}{|V_{\alpha,\Omega} - \overline{V_{\alpha,\Omega}}|} \tag{4}$$

where k stands for the four kinds of hydrophobic environments, $\overline{V_{\alpha,\Omega}} = \sum_{k=1}^{4} V_{\alpha,\Omega k}/4$. Then, we calculate vector $\overline{\mathbf{C}_\alpha} = \sum_{\Omega=1}^{20} \mathbf{C}_{\alpha\Omega}$, and rescale it into a four dimension unit vector $\mathbf{Q}_\alpha$. $\mathbf{Q}_\alpha$ describes the general contribution of hydrophobic environment in eigenvector $\mathbf{V}_\alpha$.

Table 2. Vector $\mathbf{Q_1}$ derived from the first eigenvectors of TLESUM$_{hp}$ matrices. It describes the general contribution of hydrophobic environment in eigenvector $\mathbf{V_1}$.

Matrices	$\mathbf{Q_1}$ of the 1st eigenvector (h_h , h_p , p_h , p_p)
TLESUM$_{hp}$95	(0.70 , 0.03 , -0.01 , -0.72)
TLESUM$_{hp}$90	(0.69 , 0.05 , -0.02 , -0.72)
TLESUM$_{hp}$85	(0.69 , 0.05 , -0.02 , -0.72)
TLESUM$_{hp}$80	(0.69 , 0.04 , -0.01 , -0.72)
TLESUM$_{hp}$75	(0.69 , 0.04 , -0.01 , -0.72)
TLESUM$_{hp}$70	(0.69 , 0.04 , -0.02 , -0.72)
TLESUM$_{hp}$65	(0.70 , 0.03 , -0.01 , -0.72)
TLESUM$_{hp}$60	(0.70 , 0.02 , -0.00 , -0.71)
TLESUM$_{hp}$55	(0.70 , 0.02 , -0.01 , -0.71)
TLESUM$_{hp}$50	(0.71 , 0.03 , -0.03 , -0.71)
TLESUM$_{hp}$45	(0.71 , 0.04 , -0.04 , -0.71)
TLESUM$_{hp}$40	(0.71 , 0.06 , -0.07 , -0.70)
TLESUM$_{hp}$35	(-0.68 , -0.18 , 0.18 , 0.68)
TLESUM$_{hp}$30	(-0.13 , -0.69 , 0.70 , 0.13)

Vectors $\mathbf{Q_1}$ derived from the first eigenvectors of different TLESUM$_{hp}$ matrices are shown in table 2. As the first eigenvalue is positive for each TLESUM$_{hp}$ matrix, mutations with few difference in the component of $\mathbf{Q_1}$ are conserved in hydrophobic environment substitution. Obviously, there are only two kinds of $\mathbf{Q_1}$ vectors in table 2. A transition of $\mathbf{Q_1}$ happens at sequence identity $\approx 30\%$. Substitution between h_p and p_h are primarily conserved as sequence identity > 30. As sequence identity ≤ 30, interchange h_h$\leftrightarrow$p_p is favored.

Vector $\mathbf{Q_1}$ characterizes the dominant feature in similarity of localized hydrophobic interaction. As sequence identity > 30, after simple translation and rescaling, $\mathbf{Q_1}^{>30}$ is nearly equal to vector $\mathbf{G} = (2, 1, 1, 0)$ which describes the number of hydrophobic residue in environment (h_h, h_p, p_h, p_p). Correlation coefficients $|r|$ between $\mathbf{Q_1}^{>30}$ and $\mathbf{G}$ are no less than 0.97. Namely, in the substitution of hydrophobic environment, mutations with few difference in the component of $\mathbf{G}$ is favored, i.e. number of hydrophobic residue is dominantly conserved for

near homologues. Contrast to such simpleness, implication of the other type $\mathbf{Q}_1^{\leq 30}$ is not obvious. To clarify meaning of this vector, we introduce internal hydrophobic force $\mathbf{f}$, a coarse-grained physical quantity contributed by an environment.

In an aqueous solution, water molecules attract each other, and have the effect of squeezing the hydrophobic residue. Such force does not exist for polar residue. Then there is a non zero resultant force for a residue pair which own single hydrophobic residue. For example, in environment h_p, hydrophobic force squeezes the hydrophobic side of the triplet, but not the polar side. Once we consider a force along the residue-residue virtual line, there will be a non zero resultant force F(h→p) pointing to the C-terminal residue. With a neglect of detail residue type, we can define hydrophobic force along the virtual line of a residue pair as

$$f_{ij} = \begin{cases} 1 & if \quad a_i = h, a_j = p \\ 0 & if \quad a_i = a_j \\ -1 & if \quad a_i = p, a_j = h \end{cases} \tag{5}$$

where $i, j (i < j)$ are site indices, a_i, a_j are the classified alphabets of residue i and j. For triplet words, $i = 0, j = 2$. Hydrophobic force loaded on h_p is roughly evaluated by value 1. As pointing to an opposite direction, force of p_h is defined as -1. Hydrophobic environments with identical type of members(h_h or p_p) are considered to receive 0 resultant force along the virtual line. The obtained vector $\mathbf{f} = (0, 1, -1, 0)$ is nearly equal to $-\mathbf{Q}_1^{\leq 30}$ after a simple rescaling. Correlation coefficient $|r(\mathbf{Q}_1^{\leq 30}, \mathbf{f})| \geq 0.98$. So mutations with few difference in internal hydrophobic force $\mathbf{f}$ is dominantly conserved in environment substitution as sequence identity ≤ 30.

4 Discussion

Here we reveal the dominant factors in residue substitution and localized hydrophobic interaction interchange. Our analysis is based on statistics of thousands sets of un-gapped multi-aligned fragments or blocks. Consequently, these results adapt to most protein catalogues, in other words, to substitutions of most sites and localized hydrophobic interactions. As a result of these general phenomenons, the double transitions may lead to concerted switch of residues' substitutability on multiple sites of homologous sequences; and hinder the efforts to deduce property from analysis of single point mutation.

In near homologous proteins, large amount of aligned residues are identical. Similarity of biological properties owes much to the identical physical/chemical features contributed by the same residues. What we do is peeling such simple contribution off with identity threshold decreased level by level. In remote homologues, the contribution from identical residue is not high anymore. Since trivial source of homologue's similarity is reduced, analysis based on the data of remote homologues is more suitable for touching the truth of evolution. We can extend the definition of residue-residue internal hydrophobic force to a scope larger than triplet. Then a complicated intramolecular network of hydrophobic force will be obtained for each protein. For remote homologues, internal hydrophobic force $\mathbf{f}$ is top weighted. So property given by such intramolecular network should

1. contribute much to the family representative hydrophobic interaction, and
2. provide some clues to intrinsic requirement of protein evolution.

Based on this idea, we have developed a simple 2-letter model to characterize the family representative intramolecular network of hydrophobic force. We extend the definition of residue-residue hydrophobic force from triplet to quintuplet. In given multiple sequence alignment, hydrophobic force networks of quintuplets are aligned column by column. Based on the column specific statistical information of hydrophobic force, significance of a sequence is characterized by the deviation of its inbuilt network from that of background[18].

We apply this algorithm to case of Socolich's 42 natural WW sequences[1]. With a threshold $T_S = 1.657$, almost all natively folded proteins(24 out of 28) can be identified. Percent of natively folded proteins predicted correctly(sensitivity s_n) and percent of predictions correct for natively folded proteins(specificity s_p) achieve 85.7% and 82.8% respectively. This model is absolutely based on the network of hydrophobic force, and can estimate whether a protein folds to WW structure or not sufficiently. It means that the contribution of hydrophobic force network to family representative hydrophobic interaction is quite significant.

There is distinct difference between this algorithm and HP model[19]. Traditional HP model investigates whether a fold is designable or not. We are interested in 'Why does a protein select fold A as its native structure, but not other competitors B, C, and etc?' As this character is essential for protein evolution, we want to quest for the relationship between contribution of hydrophobic force network and intrinsic requirement of protein evolution. This is

done by design artificial members of WW domain almost solely based on the contribution of hydrophobic force network.

For WW domain, it has ever been claimed by Socolich that site-independent residue propensity in multiple sequence alignment is weak in determining family specific fold of WW protein. So the column specific residue type can provide poor information, but candidate amino acid. We further diminish residue's contribution by restraining the sequence identity at a low level. In this way, the evolutionary demands required by remote homologues are also simulated. Five artificial proteins are designed. All of them have low pairwise sequence identity(30%) with each others, and with each proteins in the learning set. The only requirement is that significance evaluated by the model of hydrophobic force network is more than threshold $T_S = 1.657$. (see supporting information:Design of remote homologous proteins in WW domain). Usually it is difficult to write such sequences out, and saying nothing of sharing a family specific fold.

Results of molecular dynamics simulation show that artificial proteins fold into similar structures to the wild type WW proteins(see supporting information:Molecular dynamics simulation of artifical proteins). We find, expect residues in loop1 and N/C terminals which are less important for keeping biological properties of WW domain, residues at the vital sites fold into nearly identical structure to the natively folded protein. Similarity of these structures is not only in secondary structure representation, but also in the twisting direction, and side chain coordinate of the evolutionary conserved residues at position 7 and 31(expect residue 31 of protein 3). Although not perfect, these results are sound enough, and indicate the significant role of hydrophobic force network in meeting the intrinsic requirement of protein evolution. As only two types of letters are used in the representation of protein sequence, success of this simple model reflects the creditability of our results on the other side.

The algorithm is powerful. In another test set of 1612 nonredundant sequences, utilization of this model achieves prominent increase in searching the structurally alike reside segments at a low identity level. Therefore, though residue hydrophobic is not governing feature in residue substitution of remote homologues, hydrophobic interaction is still vital in conserving family representative properties. But it changes its emphasis from conservation of the number of hydrophobic residue to the preservation of family specific hydrophobic force network.

In near homologues, residue hydrophobicity is the dominant feature for residue substitution. Number of hydrophobic residue is seldom impacted in mutation, therefore conserved. Such conservation is like a coat beclouding the preservation of family specific hydrophobic force network, although the second one does work inherently. But in remote homologues, hydrophobicity is no longer a feature dominating residue substitution. The forementioned foundation of being a governing physical quantity disappears. Significance of hydrophobic force network emerges. So the transition might be more like a fadeout of significance of the former physical quantity, but not strengthening of the second.

We find that, in remote homologues, it is not hydrophobicity but volume acts as the dominant residue feature for the substitution of amino acid. Then, in remote homologous proteins, the major challenge of mutation is to pile up a given structure with residues of suitable size of side chain(Other features, e.g. hydrophobicity, charge, and etc, are the secondary considerations.). Contribution of hydrophobic force network meets this requirement and help the conservation of family representative biological properties significantly. For example, in residue triplet, hydrophobic force loads on non-polar side chains of the edge residues, induces different side chain rotations(with shapes ($_{v-v}$, /$_{-}$ /, \ $_{-}$\, $_{v-v}$) for (h_h, h_p, p_h, p_p) respectively, where $_{v}$ means zero resultant force is loaded on this residue; side chain rotation along the residue-residue virtual line is weak). This shapes a bed for the stacking of central residue's side chain. Consequently, similar hydrophobic force results in a similar way of side chain packing, and benefits the conservation of family specific fold.

In methodology, without comparison, we will not know all animals of quadrumana have five fingers in one hand. In the same way, no matter what we success in the investigating the physical mechanism of individual protein, without a comparison among different homologues, we are still a blind to some significant mechanism common for the whole set of homologous proteins. When we are interested in properties of a protein family, the primary object we focus is the set, but not its members. As an analogy, rule of government is different from that of citizen. Though there are many works on hydrophobic interaction, main focus of them is single protein, but not homologue set. Topics of former work and us are two different subject.

There is complicated intramolecular forces in protein, hydrophobic force is only a vital component of them. And since such component is already significant for conservation of family representative properties, it is reasonable that intramolecular forces and

molecule's mechanics property should be worth to be more concerned than ever.

To ensure a healthy development of modern biology, a patent is applied for corresponding method. We encourage pure scientific research. Contact authors when the method is to be used.

We are grateful to professor Lu-Hua Lai and Chao Tang for their helpful discussions. This work was supported in part by the Special Funds for Major National Basic Research Projects and the National Natural Science Foundation of China.

References

[1] Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H. and Ranganathan, R. "Evolutionary information for specifying a protein fold."; Nature, 437, 512-518, 2005.

[2] Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B. and Ranganathan, R. "Natural-like function in artificial WW domains."; Nature, 437, 579-583, 2005.

[3] Kinjo,A. R. and Nishikawa, K. "Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservation in proteins."; Bioinformatics, 20, 2504-2508, 2004.

[4] Pál, C., Papp, B., Lercher, M.J. "An integrated view of protein evolution."; Nat. Rev. Genet., 7, 337-348, 2006

[5] Luz H., and Vingron M. "Family specific rates of protein evolution."; Bioinformatics, 22, 1166-1171, 2006.

[6] Henikoff,S. and Henikoff,J.G. "Amino acid substitution matrices from protein blocks."; Proc. Natl. Acad. Sci., 89, 10915-10919, 1992.

[7] Henikoff,S. and Henikoff,J.G. "Automated assembly of protein blocks for database searching."; Nucleic Acids Res., 19, 6565-6572, 1991.

[8] Bairoch, A. "PROSITE: a dictionary of sites and patterns in proteins."; Nucleic Acids Res., 19, 2241-2245, 1991.

[9] Carl,B. and John,T. "Introduction to Protein Structure."; Garland Publishing,Inc., 1991.

[10] Zamyatin, A.A. "Protein Volume in Solution."; Prog. Biophys. Mol. Biol., 24, 107-123, 1972.

[11] Dill,K.A. "Dominant forces in protein folding."; Biochemistry, 29, 7133-7155, 1990.

[12] Li,H. Tang,C. and Wingreen,N.S. "Nature of driving force for protein folding: A result from analyzing the statistical potential."; Phys. Rev. Lett., 79, 765-768, 1997.

[13] Miyazawa,S. and Jernigan,R.L. "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, 1996.

[14] Liu,X., Liu,D. Qi,J. and Zheng,W. M. "Simplified Amino Acid Alphabets Based on Deviation of Conditional Probability from Random Background."; Phys. Rev. E, 66, 021906, 2002.

[15] Chou,P.Y. and Fasman,G.D. "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins."; Biochemistry, 13, 211-222, 1974.

[16] Chou,P.Y. and Fasman,G.D. "Prediction of protein conformation."; Biochemistry, 13, 222-245, 1974.

[17] Zhang, L.M. and Liu, X. "Significant Residue Features Revealed by Eigenvalue Decomposition Analysis of BLOSUM Matrices"; Phys. Lett. A, 372, 2282-2285, 2008.

[18] Contact Xin Liu for details.

[19] Li,H. Helling,R., Tang,C. and Wingreen,N. "Emergence of preferred structures in a simple model of protein folding."; Science, 273, 666-669, 1996.

Supporting Information

Design of remote homologous proteins in WW domain

Socolich's 42 natural WW sequences(Socolich et al, 2005) are analyzed with the model of internal hydrophobic force network. We use the information of multiple sequence alignment as they provided in their supplementary materials. For this 34 letters WW domain, we design artifical remote proteins with the information of internal hydrophobic force and column specific residue type. It was ever claimed by Socolich that site-independent residue propensity in multiple sequence alignment is weak in determining family specific fold of a protein. So the column specific residue type can provide poor information, but candidate amino acid. Further restraining the level of sequence identity makes such information much flimsier than ever. For each position in an artifical sequence, we assign a residue selected randomly from what have ever appeared at this position in the multiple sequence alignment. For an artifical sequence, if its score calculated from the forementioned model is more than threshold $T_S = 1.657$, and if its pairwise sequence identity with each protein of Socolich's 42 natural WW sequences and with each fore adopted artifical sequence is less than 30%, then we adopt it as a member of our designed proteins. Due to the limit of computing power, we design five sequences in this work. Detail sequences are listed here.

1) SVP-DG---WKEFKDEK-NVSFFWNIEAGTSSFAQPRLP
2) RVE-EP---WESRLSPM-DLIFFWDRFIRQSQWKDPSFI
3) PLQ-SN---YERGIDPA-ALTYYFHHQSKESSFAKPDFP
4) GVK-EV---YSIHSDLL-AVTFYFDAATNESTWAPPRME
5) SSA-EG---YQIYQTPH-AVVFFHNTESHTSQWTKPKGT

Molecular dynamics simulation of artifical proteins

All the MD simulations are performed by the GROMCAS package(van der Spoel, 2004), and the GROMOS-96 force field(van Gunsteren, 1996) is used for all simulations. Initial conformations are obtained by TASSER-Lite algorithm(Zhang and Skolnick, 2004). The proteins are enveloped in water boxes, and periodic boundary condition is used in all three directions. All simulations are performed with NVT ensemble and temperature coupling at 300K, where the coupling constant $\tau = 0.1$ps. Pressure coupling is used with a coupling constant $\tau = 0.1$ps, at a bar in all three directions. The cut-off of Lennard-Jones(LJ) interaction is 1.5nm, while the cut-off of electrostatic interaction is 2.0nm. Every system is equilibrated more than 10 ns, convergent after the equilibrium process, and the time step is 1fs.

References

[1]van der Spoel, D., van Buuren, A.R., Apol, E., Meulenhoff, P., Tieleman, D.P., Sijbers, A.L.T.M., Feenstra, K.A., van Drunen, R., Berendsen, H.J.C. *Gromacs user manual version 3.2*, (Nijenborgh 4,9747 AG Groningen, The Netherlands, 2004).

[2]van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P., Tironi, I.G. *Biomolecular Simulation: The GROMOS96 manual and user guide*, (Zurich, Switzerland:Hochschulverlag AG an der ETH Zurich, 1996).

[3]Zhang,Y. and Skolnick,J. (2004) *Proc. Natl. Acad. Sci. USA* 101:7594.