Available online at www.sciencedirect.com
ScienceDirect
Physics Letters A 372 (2008) 2282–2285
PHYSICS LETTERS A
www.elsevier.com/locate/pla

Significant residue features revealed by eigenvalue decomposition analysis of BLOSUM matrices ☆

Li-Mei Zhang [a], Xin Liu [b,*]

[a] *School of Science, Beijing Jiaotong University, Beijing 100044, China*
[b] *Institute of Mechanics, Chinese Academy of Sciences, Beijing 100080, China*

Received 17 August 2007; received in revised form 9 October 2007; accepted 10 October 2007

Available online 12 November 2007

Communicated by C.R. Doering

Abstract

Here we attempt to characterize protein evolution by residue features which dominate residue substitution in homologous proteins. Evolutionary information contained in residue substitution matrix is abstracted with the method of eigenvalue decomposition. Top eigenvectors in the eigenvalue spectrums are analyzed as function of the level of similarity, i.e. sequence identity (SI) between homologous proteins. It is found that hydrophobicity and volume are two significant residue features conserved in protein evolution. There is a transition point at SI $\approx 45\%$. Residue hydrophobicity is a feature governing residue substitution as SI $\geqslant 45\%$. Whereas below this SI level, residue volume is a dominant feature.

PACS: 87.90.+y

Keywords: Protein evolution; Residue feature; Eigenvalue decomposition; BLOSUM

1. Introduction

Our ability to characterize the biological properties of a protein is almost exclusively obtained from properties conserved through evolutionary time. Although many efforts have been made to reveal the principle governing protein evolution, it is still a field filled with many secrets [1–3]. A draft characterizing the dominant factors in protein evolution will be great helpful to us.

Essential characters of protein evolution can be learned from analysis of aligned protein sequences. A typical knowledge system is BLOSUM (BLOcks SUbstitution Matrix) matrices derived by Henikoff et al. [4]. In their scoring schemes, residue similarity was evaluated based on analysis of local sequence alignments in a high quality database-BLOCKS [5] where the most highly conserved regions (involving biologically significant sites, patterns and profiles) of related proteins in PROSITE [6] catalog were collected. Statistics of residue substitution were converted into a log-odds ratio between a combined model and an independent one. To describe fluctuation in the substitutability of residue pairs, they introduced the level of sequence similarity/identity ($x\%$, $x = 30, 35$, etc.) as a parameter in the clustering of homologous sequences. This series of scoring matrices (BLOSUM30, BLOSUM35, etc.) provides the basis for uncovering the nature of protein evolution.

In this work, we apply a general method of matrix analysis, eigenvalue decomposition, to Henikoff's BLOSUM matrices. Top weighted components of the evolution information contained in these scoring schemes are catched. To uncover the origin of residue similarity fluctuation, 14 matrices are involved (BLOSUM30, BLOSUM35, ..., BLOSUM95). It is revealed that, at SI $\approx 45\%$, there is an intrinsic transition point for the dominant residue feature related to residue substitution. Hydrophobicity [7] and volume [8] are two significant residue features in protein evolution. Residue hydrophobicity is the dominant feature conserved in protein evolution as SI is above this transition point. However, below this point, residue volume

☆ Authors contribute equally to this work.
* Corresponding author.
E-mail address: liuxin@lnm.imech.ac.cn (X. Liu).

0375-9601/$ – see front matter
doi:10.1016/j.physleta.2007.10.091

acts as the dominant factor and controls residue substitution in remote homologous proteins.

2. Materials and methods

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

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

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. Item given by the top eigenvector, $\lambda_1 V_{1,i} V_{1,j}$ has the largest contribution to element M_{ij}.

We have applied eigenvalue decomposition analysis to Henikoff's BLOSUM matrices. Each of them corresponds to a specific SI level by which segments that are identical for at least that percentage of amino acids are grouped together and weighted as a single sequence in data counting. Consequently, each matrix characterizes residue similarity for sequences below certain SI level. In our approach, SI is introduced as a parameter which varies between 30% and 95%. At each SI level, the corresponding BLOSUM matrix is analyzed after subtracting its mean from each element of the matrix. To illustrate the meaning of these eigenvectors, we further study the linear regression between eigenvector $\mathbf{V}_\alpha$ and vector R_κ which is a 20-dimensional vector introduced as a representation of residue feature κ, where κ refers to hydrophobicity, volume, secondary structure propensity [9,10], etc. Correlation coefficient r is calculated as

$$r = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}, \quad l_{xy} = \sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}), \tag{2}$$

where $\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i$. The obtained correlation coefficient is analyzed as function of SI.

3. Results

In BLOSUM matrices, 90% contributions to the total eigenvalues are made from the top 9–14 eigenvalues. Here we focus on top two eigenvalues of each matrix which contribute $\sim 1/3$ to the total. Due to positiveness of these eigenvalues, the components of corresponding eigenvectors are conserved or may be positively favored in each BLOSUM matrix.

It is found that propensity of residue substitution is related to special features of amino acid [11]. Uncovering the top weighted features will benefit the comprehension of protein evolution, and the studies in contriving molecule. For the top two eigenvectors (EV) of each BLOSUM matrix, we calculate the correlation coefficients to several residue features respectively. As shown in Fig. 1, for the 1st EV, there is a transition point at SI $\approx 45\%$ where a switch occurs for the dominant residue feature. Residue hydrophobicity has a strong relationship with the 1st EV as SI is above this transition point. However, below this point, residue volume is tightly related.

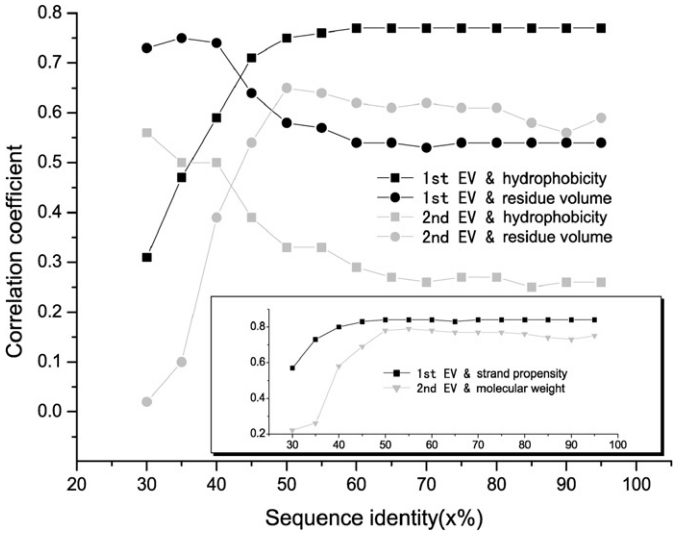

Fig. 1. Correlation coefficients between eigenvectors of BLOSUM matrices and residue features. The insert shows the coefficients of two other features related to the eigenvectors. For two features in the insert: Chou–Fasman's strand propensity correlates to hydrophobicity with $r = 0.59$, and to residue volume with $r = 0.52$. Molecular weight correlates to hydrophobicity with $r = -0.25$, and to residue volume with $r = 0.92$.

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 (see the insert of Fig. 1). To make an indubitable analysis and to indicate rotation of the 1st EV with the varying of SI, we select hydrophobicity and volume, the two orthogonal vectors in phase space as presentive features (the correlation coefficient between hydrophobicity and residue volume ≈ 0).

Origin of the forementioned transition is our next interest. Is it the result of a sharp changing of the 1st EV, or not? By plotting eigenvalue as function of SI (see Fig. 2(a)), we find that the first eigenvalues change successively. For detail analysis, we calculate correlation coefficients to measure the similarity between two 1st EV at successive SI levels, $x\%$ and $(x+5)\%$ ($x = 30, 35, \ldots, 90$). It shows that the correlation coefficients range from 0.8 to 1. No sharp transition of the 1st EV (claimed by Kinjo [3]) is found. As transition of the dominant feature would not be explained by the successive transformation of 1st EV, we further investigate similarities of these eigenvectors to some representative vectors of evolution. When SI ranges from 50% to 95%, only slightly changes occur to the top two eigenvectors (data not shown). So, we select the top two eigenvectors of BLOSUM80 as two orthogonal axes of protein evolution. Correlation coefficients between eigenvectors and the axes are shown in Fig. 2(b). We find that the top two eigenvectors contain obvious mix of the two axes as SI $< 45\%$. Since these two axes are highly related to hydrophobicity and volume respectively (shown in Fig. 1), this mix may induce a transition of the dominant residue feature.

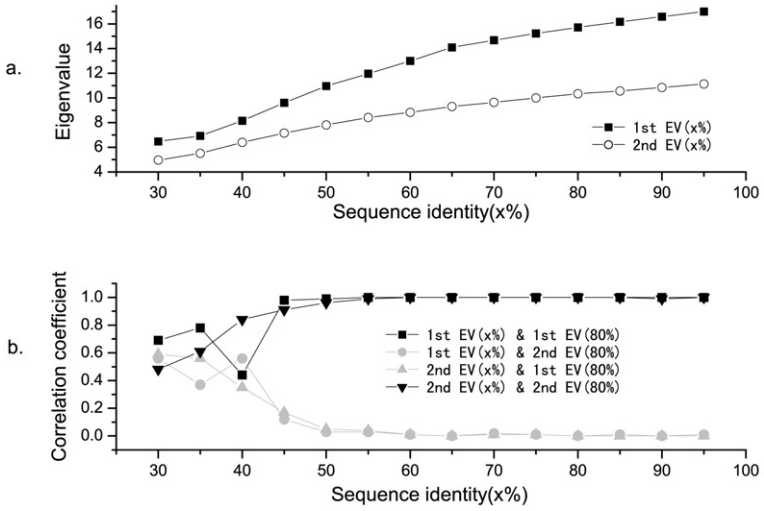

Fig. 2. Character of the variation of BLOSUM matrices' eigenvectors. (a) The top two eigenvalues as functions of sequence identity used for clustering homologous sequences. (b) Correlation coefficients between eigenvectors and those of BLOSUM80 which are selected as axes of protein evolution.

There are distinct differences between our results and those of Kinjo's. We think that it mainly results from the procedure of the subtracting of matrix mean. This procedure is skipped by Kinjo. But, as discussed by Li [12], such a subtraction procedure is necessary to remove a trivial source of a large eigenvalue. 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. For example, it is claimed by Kinjo that the first eigenvector of BLOSUM80 has a high correlation coefficient (≈ -0.68) with relative mutabilities [13]. But, in our spectrum of BLOSUM80, the eigenvector most related to Kinjo's first eigenvector (correlation coefficient ≈ 0.99) contributes only -4% to the total eigenvalues.

4. Discussion

Our analysis is based on statistics of thousands sets of ungapped multi-aligned fragments or blocks. Consequently, as a general phenomenon, transition of the governing residue feature adapts to most protein catalogues, in other words, to substitutions on most residue site. This may lead to concerted switch of residues' substitutability on multiple site of homologous sequences; hinder the efforts to deduce protein property from analysis of single point mutation.

Hydrophobic interaction is confirmed as the dominant driving force for protein folding [12,14]. The transition which claims the significance role of residue volume (but not residue hydrophobicity) at low SI level is somewhat unexpected. Then, in an evolutionary aspect, does this mean a decline of the importance of hydrophobic interaction for remote homologues? To uncover in-depth nature of protein evolution, a further study of the contribution of hydrophobic interaction has been performed.

Using a coarse-grained model, we construct a new scoring scheme named $\text{TLESUM}_{\text{hp}}$ for 3-residue pairwise substitution. By further analysis of these matrices, we achieve a new understanding of protein evolution: Hydrophobic interaction is still significant, but changes its mechanism of action at low identity level. Cooperating with the forementioned transition, a shift happens to the type of physical quantity which dominantly characterizes the contribution of hydrophobicity. As residue volume acts as the dominant feature of residue substitution, in remote homologues, the most urgent task is to construct special structure with residues of suitable size of side chain. Consequently, importance of internal hydrophobic force (the typical physical quantity of protein evolution) increases: vector of hydrophobic force loads on residue's side chain, induces respective side chain rotations. So, similar network of internal hydrophobic force results in a similar way of side chain packing. This conclusion is robust, and has remarkable effect in keeping the biological properties of homologous proteins. Actually, we have developed a coarse-grained algorithm to characterize the internal hydrophobic force network of a protein family. Although only information of internal hydrophobic force is considered in this algorithm, it accomplishes an accuracy of more than 85% in singling out the native folded WW domain (true signal) from a 42 members protein set which is obtained by multiply sequence alignment [1]. Moreover, for this 34 letters WW domain, we design several artificial remote proteins with the information of internal hydrophobic force and column specific residue type. All these proteins have low pairwise sequence identities (30%, lower than the threshold [15] of twilight zone at sequence length $= 34$) with each others, and with each proteins in the learning set. Results of molecular dynamic simulation show that they fold to similar structures to the wild type proteins. Detailed description of this part will be published elsewhere.

References

[1] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, R. Ranganathan, Nature 437 (2005) 512.
[2] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffe, R. Ranganathan, Nature 437 (2005) 579.

[3] A.R. Kinjo, K. Nishikawa, Bioinformatics 20 (2004) 2504.
[4] S. Henikoff, J.G. Henikoff, Proc. Natl. Acad. Sci. 89 (1992) 10915.
[5] S. Henikoff, J.G. Henikoff, Nucleic Acids Res. 19 (1991) 6565.
[6] A. Bairoch, Nucleic Acids Res. 19 (1991) 2241.
[7] B. Carl, T. John, Introduction to Protein Structure, Garland Publishing, 1991.
[8] A.A. Zamyatin, Prog. Biophys. Mol. Biol. 24 (1972) 107.
[9] P.Y. Chou, G.D. Fasman, Biochemistry 13 (1974) 211.
[10] P.Y. Chou, G.D. Fasman, Biochemistry 13 (1974) 222.
[11] X. Liu, D. Liu, J. Qi, W.M. Zheng, Phys. Rev. E 66 (2002) 021906.
[12] H. Li, C. Tang, N.S. Wingreen, Phys. Rev. Lett. 79 (1997) 765.
[13] D.T. Jones, W.R. Taylor, J.M. Thornton, Comput. Appl. Biosci. 8 (1992) 275.
[14] K.A. Dill, Biochemistry 29 (1990) 7133.
[15] B. Rost, Protein Eng. 12 (1999) 85.