Journal of Theoretical Biology 261 (2009) 593–597

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# A scheme for multiple sequence alignment optimization—an improvement based on family representative mechanics features

Xin Liu, Ya-Pu Zhao *

*The State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, No. 15 Beisihuanxi Road, Beijing 100190, China*

ARTICLE INFO

*Article history:*
Received 22 April 2009
Received in revised form
26 August 2009
Accepted 26 August 2009
Available online 3 September 2009

*Keywords:*
Scoring scheme
Hydrophobic interaction
Multiple sequence alignment
Homologous proteins

ABSTRACT

As a basic tool of modern biology, sequence alignment can provide us useful information in fold, function, and active site of protein. For many cases, the increased quality of sequence alignment means a better performance. The motivation of present work is to increase ability of the existing scoring scheme/algorithm by considering residue–residue correlations better. Based on a coarse-grained approach, the hydrophobic force between each pair of residues is written out from protein sequence. It results in the construction of an intramolecular hydrophobic force network that describes the whole residue–residue interactions of each protein molecule, and characterizes protein's biological properties in the hydrophobic aspect. A former work has suggested that such network can characterize the top weighted feature regarding hydrophobicity. Moreover, for each homologous protein of a family, the corresponding network shares some common and representative family characters that eventually govern the conservation of biological properties during protein evolution. In present work, we score such family representative characters of a protein by the deviation of its intramolecular hydrophobic force network from that of background. Such score can assist the existing scoring schemes/algorithms, and boost up the ability of multiple sequences alignment, e.g. achieving a prominent increase ($\sim 50\%$) in searching the structurally alike residue segments at a low identity level. As the theoretical basis is different, the present scheme can assist most existing algorithms, and improve their efficiency remarkably.



Many computer-based tools in modern biology involve protein sequence alignment as an essential process. These alignments usually provide not only important insights into the bio-properties of genes and proteins, but also a cheap and swift technical route. With the increase in the ability of sequence alignment, many experiments in wet lab could be assisted, even accomplished in dry lab. Therefore the power of alignment schemes/tools could impact the present, even the future of biology.

There are complicated residue–residue interactions in a protein molecule. Most traditional scoring schemes were based on the hypothesis of single point mutation (Dayhoff and Eck, 1968; Henikoff and Henikoff, 1992; Karlin and Altschul, 1990; Altschul, 1991). Some advance methods took such correlations into account by hidden Markov model (Karplus et al., 1998; Finn et al., 2008), k-string (Ogul and Mumcuoğlu, 2007), or other considerations. However, owing to the unsolved secret in protein evolution, the suitable way in treating the residue–residue correlations is still an open question. With a proper consideration of the high order correlations, there should be large room in optimizing the ability of the existing alignment schemes.

Hydrophobic interaction is important for protein folding (Dill, 1990; Li et al., 1997) and function (Jones and Thomton, 1996; Young et al., 1994). In an aqueous solution, hydrophobic residue is loaded a force by water. Surrounding an amino acid, water molecules attract each other, and have the effect of squeezing the hydrophobic residue. On the contrary, no such force is loaded on a polar residue. In consequence, there is a force along the line between each pair of residues. These force vectors form a complicated network that characterizes the whole hydrophobic interaction of a molecule. Our former investigation has suggested that, in an aspect of hydrophobicity, contribution of such network is top weighted in conserving the family representative biological properties of a protein (Liu et al., 2008). Coinciding with this, mechanics feature was suggested to be vital to protein evolution (Tokuriki and Tawfik, 2009). Therefor the biological properties of a protein are related, even mainly determined by such network.

This discovery provides the theoretical basis of a new scoring scheme in evaluating the propensity of a protein to be a member of certain family, and in optimizing the quality of multiple

* Corresponding author. Tel.: +86 10 82543932.
*E-mail addresses:* liuxin@lnm.imech.ac.cn (X. Liu).
yzhao@imech.ac.cn (Y.-P. Zhao).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.08.028

sequence alignment. The scoring scheme is established in three steps:

(i) *Define force representation from protein sequence.* It was shown by several works that, as residue alphabets are grouped into two catagories, no large difference exists among various clustering schemes (Wang and Wang, 1999; Liu et al., 2002). The amino acid classifications correlate strongly with residue hydrophobicity (Carl and John, 1991). Our former work has shown that, as residue sequence is degenerated into a 2-letter chain, the force along the line of a residue pair could be clustered into three states (see Eq. (1) in Method section). In network formed by such coarse-grained force vectors, force strength and direction are determined by residue type, but not its three dimensional coordinates. Namely we can define a force network merely from protein sequence.

(ii) *Treat multi-aligned sequences as multi-aligned force networks.* With our force representation, one and only force network is derived from a protein molecule. We can then compare the networks one another, i.e. among different proteins. In multi-aligned sequences, residues are aligned column by column. Due to the definition of coarse-grained force, a residue pair contribute a force vector, two joint residue columns correspond to a column of forces. Consequently, forces are also aligned column by column, resulting in multiple network alignment. In this way, sequence alignment is mapped to network alignment.

(iii) *Score significance of a protein sequence by the information of multi-aligned force networks.* For a set of aligned networks/sequences, force states in each force column are analyzed by statistical approach. Significance of a sequence, the propensity to be a member of corresponding protein family is evaluated by the deviation of its inbuilt force network from that constructed by a background model.

We tested the ability of present scheme by assisting traditional scoring schemes of multiple sequence alignment, in searching the structurally alike residue segments at a low identity level. About 50 percentage increase was achieved. The power of present scheme in designing new members of a protein family, together with the conservation of family representative bio-chemical properties, is also introduced.

## 1. Methods

For the force networks of a set of homologous proteins, we suppose there is a family representative network that is the consensus of the set members. According to such consensus, significance of a protein can be evaluated by the deviation of its intramolecular force network from that of background. But there are $C_\chi^2$ forces in a $\chi$-residue protein. Cost of a complete consideration is extremely high. As a feasible choice, we focus on the network of localized hydrophobic interaction. Illustration of the scheme is shown in Fig. 1.

In a given multiple sequence alignment, protein sequence is rewritten as successive overlapping 5-residue units. By convention, we assign the first quintuplet to the third residue, the second to the fourth and so on, until finally the last quintuplet is assigned to the last residue but two. Consequently we get the multi-aligned quintuplet sequences with the same gap positions as those of the original multiple sequence alignment, but four additional gaps on the first two and last two sites in each sequence.

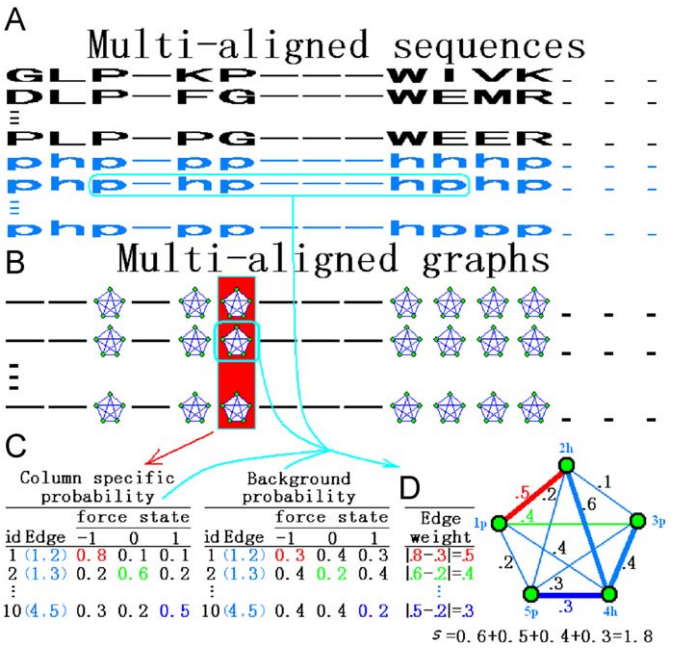


**Fig. 1.** Illustration of scheme HFnet (Hydrophobic Force network). (A) Group residue alphabets into $h$ (hydrophobic) or $p$ (polar) types. Rewrite protein sequence into 2-letter chain. (B) Draw a graph of residue-to-residue interaction in each quintuplet. Map multi-aligned sequences into multi-aligned graphs. Define state of hydrophobic force along each edge. (C) Calculate edge specific probability of the occurrence of force state for each column of aligned graphs, and that for background graph set. The background graph includes all graphs in sequences that are selected as background set. (D) Find the maximum spanning tree for each graph (the tree is drawn in bold line). Value beside each edge is the edge weight, which reflects the difference between column specific probability and that of background.

### 1.1. Force and force network representations

For each quintuplet, after drawing the $C_5^2 = 10$ residue-to-residue virtual lines, we get a graph of the local network of residue–residue pairwise interactions. Complete graph $G = (V, E)$ is introduced to characterize this network. Each residue of a quintuplet is represented by a vertex in the vertex set $V$. Edge $(i,j) \in E$ between residues $i$ and $j$ denotes a pairwise interaction.

Based on our former work, we define the state of hydrophobic force $f_{ij}$ along each edge $(i,j)$ as

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

where $h/p$ is for hydrophobic/polar residue, $1 \le i < j \le 5$, $a_i, a_j$ are the classified alphabets of residue $i$ and $j$. Scheme of residue classification is: hydrophobic $h = \{M, F, I, L, V, A, W\}$, polar $p = \{C, Y, Q, H, P, G, T, S, N, R, K, D, E\}$ (Liu et al., 2002). Owing to the differences of forces loaded by aqueous solution, the resultant force between a residue pair is different from one another. With the introduction of amino acid classification, these force states could be clustered into several categories. There are a lot of works focusing on the coarse-grained representation of protein sequences (Liu et al., 2002; Fan and Wang, 2003; Li et al., 2003; Bacardit et al., 2009). It has been proven that the hydrophobic/polar feature is the dominant factor in clustering residue alphabets into two categories. Therefore the force states defined with $h/p$ letters are expected to be robust. In our force representation, water molecules squeeze the hydrophobic residue of the residue pair $h\_p$, resulting in a resultant force measured by

1 along the virtual line (pointing at the C terminal) of the two residues. As pointing to the opposite direction, the force state of *p_h* is defined as −1. Since the solution contributes nearly equal but opposite forces on the two residues, respectively, the resultant force along the virtual line is approx. zero for *h_h*. The state of *p_p* is defined as 0 due to the absent of hydrophobicity. In 2008, we analyzed the database of homologous proteins using residue triplets, and found that the aforementioned coarse-grained force representation is general and top weighted in the hydrophobic interactions conserved in various protein families. Here the definition we learned is extended from triplet to quintuplet. Such extension is local and moderate, introduces acceptable complexity, should coincide with the theoretical basis learned from triplet analysis.

With such force representation, a force network is derived from each quintuplet. Each of them is a subset of the whole hydrophobic force network of corresponding protein. In this way, the representation of the whole force network is simplified to a sequence of successive graphs with definite force state along each edge. In consequence, when we treat multiple sequence alignment as aligned force networks, the aligned residue quintuplets correspond to aligned force networks/graphs of quintuplets.

### 1.2. Model of hydrophobic force network

In multiple sequence alignment, the force graphs of quintuplets are aligned column by column. For edge $(i,j)$ of column $k$, probability of the occurrence of force state $l$ is calculated as

$$P_{ij}^{k}(l)=\frac{\sum_{n=1}^{M^k}\delta(l,f_{ij}^{nk})}{M^k} \tag{2}$$

where $n$ is the sequence index, $M^k$ is the size of aligned graph in column $k$, $\delta(x,y)$ is the step function with $\delta(x,y)=1$ for $x=y$ and $\delta(x,y)=0$ otherwise, $f_{ij}^{nk}$ is the force state given by graph $G^{nk}$. As gap is skipped, $M^k$ may be less than the number of aligned sequences. For a set of sequences, a background graph can be constructed with the information of all graphs contained in these sequences. For a background graph, probability of the occurrence of force state $l$ can be calculated as

$$Q_{ij}(l)=\frac{\sum_{k=1}^{L}\sum_{n=1}^{B^k}\delta(l,f_{ij}^{nk})}{\sum_{k=1}^{L}B^k} \tag{3}$$

where $L$ is the column number of the multi-aligned quintuplet sequences, $B^k$ is the size of aligned graphs in column $k$ of the background set (There is only one background graph in each iteration. All graphs in background set are used in evaluating $Q$. Except the first iteration, the background set is only a subset of the total sequences.). In graph $G^{nk}$, the difference between hydrophobic force $f_{ij}^{nk}$ and that of the background model can be characterized as

$$D_{ij}^{nk}=|P_{ij}^{k}(f_{ij}^{nk})-Q_{ij}(f_{ij}^{nk})| \tag{4}$$

which is considered as the weight of edge $(i,j)$. We suppose that graphs in biologically significant proteins should be remarkably different from that of the background. Thus, for graph $G^{nk}$, we find the maximum spanning tree (Kruskal, 1956) that has a weight sum termed as $s(G^{nk})$, and indicates the most significant non-redundant interactions contributed by the 5-residue collectivity. Then, the significance of sequence $n$ is scored by mean of $s(G^{nk})$ as

$$S^{n}=\frac{\sum_{k=1,k\neq g_n}^{L}s(G^{nk})}{\sum_{k=1,k\neq g_n}^{L}1} \tag{5}$$

where $g_n$ stands for gap position.

The whole aligned sequences are used as background data set firstly. Then we array protein sequences by their scores in a decrease order. Sequences at top of the rank are deemed to be more significant than those at the tail. Using sequences at the tail as new background set, we update $Q$ and run iteration until convergence of the algorithm. Although only sequence information is used, as complete graph is employed in our approach, the present scheme is a model of three-dimensional network.

## 2. Result

We present a scheme HFnet (Hydrophobic Force network) to evaluate the significance of each sequence in a given multiple sequence alignment. The multi-aligned sequences can be obtained either by the existing sequence alignment algorithms or by the methods developed/optimized by users. All the need of HFnet is multi-aligned sequences. Structural information is unnecessary. Whether these sequences are aligned with structural information or not relies on alignment algorithm, but not the scoring scheme HFnet.

In HFnet, only two letters are used in sequence representation. Such coarse-grained model results in large amount of positive signals, therefore cannot be used alone. The ability of HFnet is assisting the existing scoring scheme. And as HFnet can assist most schemes, it is not appropriate to compare HFnet with other schemes. What is important is the ability enhancement contributed by HFnet. Hereby searching structurally alike polypeptides of query segments, we show the power of HFnet in assisting traditional sequence alignment schemes/approaches.

A non-redundant set of 1612 non-membrane proteins from PDB_SELECT25 (Hobohm and Sander, 1994) (issued on September 25 of 2001) is selected as our test set where no pair of sequences share sequence identity more than 25%. By sliding a window along protein sequence, each 21-residue segment of this data set is chosen as a query segment. For each query segment, we search its remote homologs in our data set, i.e. the homologs in subset $\{\mathbf{SI}_{\leq 25}\}$ for which each member shares no more than 25% residues with the query one. To construct a multiple sequence alignment, a traditional position-specific matrix is learned from a set of near homologs of the query segments (share no sample with subset $\{\mathbf{SI}_{\leq 25}\}$, see the supporting information).

We score each segment in subset $\{\mathbf{SI}_{\leq 25}\}$ according to the position-specific matrix, and rank them in a decrease order. The top 50 segments are collected. A simple approach is employed to show the power of HFnet—whether rerank these 50 segments with HFnet algorithm or not. Then the position-specific matrix is updated with the top 30 segments, further iteration is processed in subset $\{\mathbf{SI}_{\leq 25}\}$. After iteration, as output of sequence alignment, the top 30 segments are checked. We say segment $v$ is a remote homolog of $\mu$ if the structural similarity $drms(\mu,v)<\theta$ Å (Park and Levitt, 1995), where $\theta$ corresponds to different threshold for structural similarity.

A jackknife test is performed. The protein that contains the query segment is removed from the data set. Then for each query segment, remote homologs are searched in the remaining data set according to the aforementioned process. In this test method, the sum of the output segments is equal for different cases such as the times of iteration, the choices of threshold $\theta$, the scoring scheme in calculating position-specific matrix, and whether HFnet is employed or not. We can evaluate the performance of sequence alignment by a ratio between the sum of remote-homologous segments obtained (True-Positive, TP) and that of the left (False-Positive, FP). Scoring schemes BLOSUM$\omega$ (Henikoff

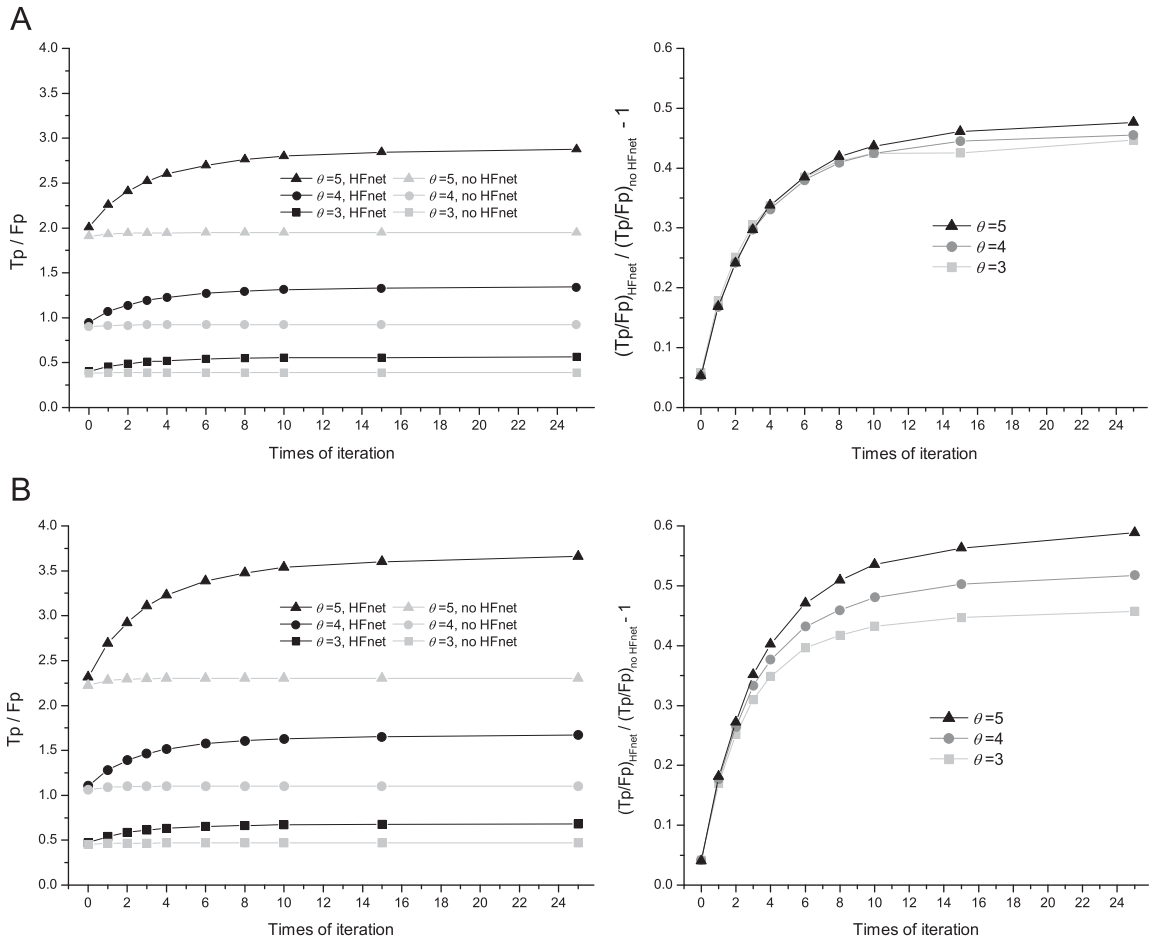


**Fig. 2.** Results in searching structurally alike residue segments at a low identity level. Ratios between sum of the True-Positive (TP) segments and that of the False-Positive (FP) segments are shown for different choices of threshold $\theta$, i.e. different levels of structural similarity. Increase of TP/FP ratio shows the ability of HFnet in assisting traditional single-point-mutation based scoring scheme. (A) Results with data of BLOSUM30 matrix in calculating position specific matrix. (B) Results with data of BLOSUM62 matrix.

and Henikoff, 1992), $\omega = 30$ and 62 are used in calculating the position-specific matrix, respectively. As shown in Fig. 2, HFnet offers a significant shift toward the region with fewer mismatches. When HFnet is absent, the testing approach is a typical single-point-mutation based sequence alignment method, with a rank of E-value (expected value; Altschul et al., 1997). Therefore, increase of TP/FP ratio shows the ability of HFnet in assisting traditional scoring scheme. If we define the enhancement as $(TP/FP)_{HFnet}/(TP/FP)_{noHFnet} - 1$, nearly 50% improvement is achieved.

## 3. Discussions

Here we contribute a scheme in multiple sequence alignment optimization. Similar to the well-known HP model (Li et al., 1996), only two residue alphabets are used in HPnet. But the main focus of HFnet is 'Why does a protein select fold *A* as its native structure, but not other competitors *B*, *C*, etc.?'. We attempt to present a representation of the common and representative family characters exist in the inbuilt force networks of a group of homologs, then improve the power of alignment tools thereby.

Actually, for a theory or scoring scheme, the soundest proving comes from wet experiment. We have designed five 34 residue artificial remote proteins of the WW domain using HFnet. All of them have low pairwise sequence identity ($< 30\%$) with each other and with each proteins in the learning set. Usually, it is hard to write them out, saying nothing of sharing family representative biological properties. Experiment of isothermal titration calorimetry showed that all of the five proteins exhibited detectable ligand-binding affinity. Four of them have the bio-activities in a similar level with wild-type proteins (to be published elsewhere). It means that HFnet scores not only the propensity of fold, but also that of biological properties. For an $N$ residue sequence, there are only about $(3-1)\times C_5^2 \times N = 20 \times N$ parameters to be estimated, i.e. in a similar complicity to protein sequence. For such a simple model, the aforementioned achievements indicates that the power of HFnet is creditable. As thus, HFnet has been used in the studies of detailed, donut-shaped topological feature of polypeptide relationship and significant sites responsible for initial pathogenic structural changes in conformational disease (Liu and Zhao, 2009).

As a rough model, there is room for further improvement of this scoring scheme. For example, a fold/family specific clustering scheme can be introduced in residue classification. Elaborate states can also be employed in characterizing hydrophobic force skillfully. Robusticity of this unpolished algorithm means a bright future of present scheme.

Finally, we want to say a few words on the usage of HFnet. Quality of multiple sequence alignment is crucial for this

scheme. The kernel of HFnet is distinguishing true signals from that of the background. Once most true signals drop to the background subset, the results are less satisfactory. Users should improve the alignment quality so that enough true signals are involved. The other important parameter is sequence length. As only two letters are used in HFnet, for such a simple scheme, signal of too short sequence may be missed. In several applications of HFnet, the sequence length ranges from 15, 17, 19 (Liu and Zhao, 2009) to 34 (WW domain). In all these cases, the scheme performs very well. Though the work focuses on the local interactions in proteins, there is not obvious length limit for long cases.

## Acknowledgments

A patent is applied for present scheme. 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 jointly supported by the National High-tech R&D Program of China (863 Program, Grant No. 2007AA021803), National Basic Research Program of China (973 Program, Grant No. 2007CB310500), and National Natural Science Foundation of China, No. 10704077.

## References

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219, 555–565.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bacardit, J., Stout, M., Hirst, J.D., Valencia, A., Smith, R.E., Krasnogor, N., 2009. Automated alphabet reduction for protein dataset. BMC Bioinform. 10, 6.

Carl, B., John, T., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York.

Dayhoff, M.O., Eck, R.V., 1968. Atlas of Protein Sequence and Structure, vol. 3. National Biomedical Research Foundation, Silver Springs, MD, pp. 33–45.

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

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

Finn, R.D., Tate, J., et al., 2008. The Pfam protein families database. Nucleic Acids Res. 36, D281–D288.

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

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

Jones, S., Thomton, J.M., 1996. Principles of protein–protein interactions. Proc. Natl. Acad. Sci. 93, 13–20.

Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. 87, 2264–2268.

Karplus, K., Barrett, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856.

Kruskal, J.B., 1956. On the shortest spanning tree of a graph and the traveling salesman problem. Proc. Amer. Math. Soc. 7, 48–50.

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

Li, H., Tang, C., 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., Wang, W., 2003. Reduction of protein sequence complexity by residue grouping. Protein Eng. 16, 323–330.

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

Liu, X., Zhang, L.M., Yin, J., Zhao, Y.P., 2008. Major factors of protein evolution revealed by eigenvalue decomposition analysis. In: Proceedings of the International Conference on Bioinformatics and Computational Biology BIOCOMP'08, Las Vegas, USA, pp. 91–97.

Liu, X., Zhao, Y.P., 2009. Donut-shaped fingerprint in homologous polypeptide relationships—a topological feature related to pathogenic structural conversion of conformational disease. J. Theor. Biol. 258, 294–301.

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

Park, B.H., Levitt, M., 1995. The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol. 249, 493–507.

Tokuriki, N., Tawfik, D.S., 2009. Protein dynamism and evolvability. Science 324, 203–207.

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

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