The significance of molecular mechanics property to protein evolution

Li-Mei Zhang

School of Science,
Beijing Jiaotong University,
Beijing 100044, China

Xin Liu*

The State Key laboratory of Nonlinear Mechanics,
Institute of Mechanics, Chinese Academy of Sciences,
No. 15 Beisihuanxi Road, Beijing 100190, China
E-mail: liuxin@lnm.imech.ac.cn
*Corresponding author

Abstract: In the study of protein homology, multiple physical systems that correspond to different proteins must be jointly considered, such that the similarity of homologs can be explained in an aspect of the physics. Compared with the investigations that focus on the character of certain protein, the research object changes from one physical system to a set of systems. For this enlarged object, the importance of different physical quantities is still unclear. In an aspect of hydrophobic interaction, we rank the significance of different physical quantities here, according to their contribution to the conservation of family representative biological properties in protein evolution. Molecular mechanics property is suggested to be the governing factor responsible for protein homology.

Keywords: hydrophobic interaction; protein evolution; mechanics property.

Reference to this paper should be made as follows: Zhang, L-M. and Liu, X. (2013) 'The significance of molecular mechanics property to protein evolution', *Int. J. Data Mining and Bioinformatics*, Vol. 8, No. 1, pp.83–91.

Biographical notes: MS. Li-Mei Zhang is an Assistant Professor of the School of Science, Beijing Jiaotong University. She is expert in biophysics and bioinformatics, and interested in molecular modeling and analysis of protein. Her former works related to amino acid classification and analysis, governing residue features of protein evolution.

Dr. Xin Liu is an Associate Professor of the institute of mechanics, Chinese Academic of Sciences. He is expert in biomedical informatics, molecular dynamics simulation, modeling and experiment of protein molecule. His former works related to protein structure modeling and prediction, revealment of the underlying principle of protein evolution, physical mechanism of protein misfolding, and diseases.

The authors contributed equally to this work.

1 Introduction

Proteins are different from each other in amino acid sequence. For each protein, the molecules of protein and solution form a complicated physical system. Free energy of the physical system per protein is believed to play a vital role in protein folding. As the residue sequences are different among homologs, the residue interactions that contribute free energies should also be different in their corresponding physical systems. Consequently, the free energies are not similar among homologous proteins, especially for those of remote homologs. Therefore, in the study of protein homology, the importance of free energy is not as important as that in protein folding. There is a requirement to reevaluate the significance of different physical quantities in their contribution to protein homology.

Hydrophobic interactions have been suggested as the driving force of protein folding (Li et al., 1997; Dill, 1990), and play an important role in protein function (Jones and Thomton, 1996; Young et al., 1994). With revealment of features of hydrophobic interactions, scientists have achieved many progresses in characterising protein's properties, such as folding mechanism (Li et al., 1996, 1997), marginal stability (Taverna and Goldstein, 2002; Bloom et al., 2004), kinetics of function (Gupta and Irbäck, 2004), and etc. Owing to its importance to protein molecule, hydrophobic interaction is an ideal aspect in evaluating the significance and contribution of different physical quantities to protein homology.

As the biological properties are conserved among homologous proteins, their physical systems can be deemed similar functionally. Once homologs are aligned site by site, hydrophobic interaction systems of these proteins can be deemed to be aligned too. We can compare the hydrophobic interactions of a protein with those of the other molecules, evaluate the significance of different physical quantities in their contributions to the similarities among these systems, and rank their importance for the conservation of bio-properties with the mathematical method such as eigenvalue decomposition analysis.

In accordance with the aforemationed deduction, based on an analysis of the similarity of local hydrophobic interaction, we find that the energetic property is not the factor that contributes most importantly to the conservation of bio-properties among remote homologous proteins. There is an intrinsic transition point at the sequence identity $\approx 30\%$ for the type of physical quantity significant for protein evolution. The importance of the similarity of hydrophobic/polar residue composition, which is related to the energetic item, could make sense only for near homologous proteins. For remote homologs, the force vector is the most important physical quantity responsible for the conservation of biological properties. As the protein design of remote homologs of a protein family needs an efficient process of identifying the eligible nonredundant candidate from a huge (20^N) sequence space, an algorithms of the direct description of the underlying principle of protein evolution is required to achieve such goal. Since an algorithm based on the network of intramolecular force would accomplish such a challenge, the conservation of mechanics property is suggested to be a basic requirement of the evolution of protein.

2 Materials and methods

In an aspect of hydrophobic interaction, we focus on the importance of different physical quantities in their contributions to the conservation of the

representative biological properties of a protein family. Due to the significant role of hydrophobic interaction, the physical quantity identified through such coarse-grained approach should also be important according to a scheme using more elaborate considerations. In this work, the hydrophobic interaction systems of different homologous proteins are deemed to be similar. The physical quantity that governs such similarity is identified by analysing the local hydrophobic interactions in the aligned homologous sequences. To reduce the bias induced by close related homologous proteins, we introduce the level of sequence identity as a parameter. Proteins with sequence identity below certain level are involved in an analysis. Top weighted factor that contributes to the similarity of hydrophobic interaction system is revealed as a function of sequence identity.

We focus on the similarities of hydrophobic interaction systems in aligned homologous proteins. The original data are analysed by statistical approach. To make a general conclusion, we choose the aligned sequences in BLOCKS9 (Henikoff and Henikoff, 1991) as our data set. In this high quality database, sequences of biologically significant sites, patterns and profiles of numerous protein families are involved. Totally, 3179 blocks, i.e., groups of ungapped multiple aligned homologous polypeptides are used in this analysis. Our results are derived from the similarity of short residue segments. We believe they capture the essential features at local level.

Information conserved in multiple homologous sequence alignment is a significant evolutionary source, and is the foundation of many most-important tools in bioinformatics, such as BLAST for sequence alignment (Altschul et al., 1997), MODELLER (Fiser and Sali, 2003) and I-TASSER (Zhang, 2008) for protein structure modeling and prediction, and etc. Although these approaches are primarily based on local scores, they achieve huge successes. It indicates that evolutionary information in local is vital for evolvement analysis, and sequence alignment can preserve dominating information of evolution. Therefore, it is reasonable to learn local level evolutionary rule from aligned homologous polypeptides. And such rule is also important enough.

Liu et al. (2003) has shown that the direct adjacent neighbors of a central residue have the most important pairwise interactions with the center. In order to clarify the most important evolutionary information by a succinct analysis, we select triplets as the basic units. Protein sequence $a_0a_1a_2a_3a_4a_5a_6\ldots$ is treated as successive overlapping triplet words of amino acid $(a_0a_1a_2)(a_1a_2a_3)(a_2a_3a_4)(a_3a_4a_5)(a_4a_5a_6)\ldots$ In aligned homologous proteins, triplets are aligned column by column. Samples in one column can be deemed to be substitutable for each another. Such approach of evaluating similarity with substitutablility is derived from analysis of homologous protein sequence alignment, and has been a standard method to analyse the similarities among residues (Henikoff and Henikoff, 1992) (monomers), k-words (Liu and Zhao, 2010a) (triplets), local conformations (Liu et al., 2008) (quartets), and so on. The higher the substitutability between two samples, the more similar they are. Here, each triplet is a subset of the intramolecular interactions of corresponding protein, and can be deemed as a joint unit composed of the central residue and the local hydrophobic interaction(LHI) that is provided by the two side amino acids. In aligned homologous proteins, such LHIs are aligned too, and can be deemed substitutable for each other. Substitutability/similarity of these LHIs can be observed in the similarities of triplets as they interchange with each other in aligned homologous proteins.

2.1 Score triplet's similarity by TLESUM$_{hp}$ matrix

Scoring scheme TLESUM$_{hp}$ is introduced to evaluate similarity of triplets (Liu and Zhao, 2010a). Each element of TLESUM$_{hp}$ matrix scores the substitutable propensity of two triplets which interchange with each other in homologous proteins. In matrix element ij, a large score means the corresponding triplets, i and j, are similar and easy to be substituted with each another. In this scheme, due to the limit of data size, two neighbors of the central residue are classified into hydrophobic(h) or polar(p) groups respectively (Liu et al., 2002), and provide four kinds of coarse-grained LHIs(h_h, h_p, p_h, p_p) for the center, where '_' stands for central residue. Consequently, the original $20 \times 20 \times 20 = 8000$ types of triplets are clustered into $2 \times 20 \times 2 = 80$ alphabets. As a result, size of TLESUM$_{hp}$ is 80×80. It is important to classify triplets into fewer types. Otherwise, there will be sparse samples for a triplet pair. But the detail difference among various residue types are omitted as these members are classified into the same group, that is, specificity among different types of triplets decreases after a clustering approach. Therefore, excessive clustering results in a drastic decrease of specificity. As the central residue of triplet is not degenerated, a tradeoff is introduced in this scheme. Distinctiveness of localised hydrophobic interaction is kept moderately.

Samples in database BLOCKS9 are biased, i.e., many segments are closely related. In constructing TLESUM$_{hp}$ scoring scheme, such bias was reduced by clustering similar members within blocks. This is done by specifying a parameter, sequence identity, by which residue segments that are identical for at least that percentage are grouped together. Each cluster is weighted as a single sequence in data counting. Consequently, one matrix characterises triplet's similarity in homologues that have sequence identity below a certain level. To investigate the similarity of LHI among homologs of different identity level, 14 matrices with various sequence clustering levels are analysed (TLESUM$_{hp}30$, TLESUM$_{hp}35, \ldots$, TLESUM$_{hp}95$; see supporting information: About TLESUM$_{hp}$ matrices). Namely sequence identity is introduced as a parameter in our analysis (range from 30% to 95%).

2.2 Deduce the major factors responsible for the similarities of LHIs from TLESUM$_{hp}$ matrix

The major factors responsible for the similarities of LHIs are deduced from the similarities of residue triplets. With a general method, eigenvalue decomposition, information involved in a matrix can be ranked according to its significance. After subtracting the mean of corresponding matrix from each element(see supporting information:About eigenvalue decomposition), eigenvalue decomposition analysis is applied to each of the 80×80 real symmetric TLESUM$_{hp}$ matrices. In this 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}.

$\mathbf{V}_1$ of TLESUM$_{hp}$ matrix is significant for the similarity of triplet. In this 80 dimensional vector, each type of triplet corresponds to a component of this eigenvector. For each kind of central residue Ω, there are four relevant components corresponding to hΩh, hΩp, pΩh, and pΩp respectively. As the central residues are identical, difference of values of these components comes from the difference of the type of LHI(h_h, h_p, p_h, p_p). So, in case of central residue Ω, contribution of LHI can be described as the following unit vector $\mathbf{C}_{1\Omega}$

$$C_{1\Omega,k} = \frac{V_{1,\Omega k} - \overline{V_{1,\Omega}}}{|V_{1,\Omega} - \overline{V_{1,\Omega}}|} \tag{2}$$

where k stands for the four kinds of LHIs, $\overline{V_{1,\Omega}} = \sum_{k=1}^{4} V_{1,\Omega k}/4$. Then, we calculate the mean vector $\overline{\mathbf{C}_1} = \sum_{\Omega=1}^{20} \mathbf{C}_{1\Omega}/20$, and rescale it into a four dimensional unit vector $\mathbf{Q_1}$. $\mathbf{Q_1}$ describes the general contribution of LHI to triplet similarities according to the top weighted eigenvector $\mathbf{V}_1$, that is, the major factor responsible for the similarity of LHIs.

In eigenvalue decomposition approach, item

$$\lambda_1 V_{1,i} V_{1,j} = \frac{\lambda_1}{2}[V_{1,i}^2 + V_{1,j}^2 - (V_{1,i} - V_{1,j})^2] \tag{3}$$

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 TLESUM$_{hp}$, large value of a matrix element means the large substitutable propensity between members of a triplet pair. Consequently, as LHIs substituting with each other, mutations with few difference in the component of $\mathbf{Q_1}$ are conserved or may be positively favored. Physical quantity corresponding to $\mathbf{Q_1}$ is top significant for the conservation of the representative biological properties of a protein family.

3 Results

We have revealed the top weighted factors responsible for the similarities of LHIs in 59325 segments. According to different levels of sequence identity, vectors $\mathbf{Q_1}$ derived from the first eigenvectors of corresponding TLESUM$_{hp}$ matrices are shown in Table 1. 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 LHI substitution. It is quite obvious that only two kinds of $\mathbf{Q_1}$ vectors exist in Table 1, i.e., vector (-0.13, -0.69, 0.70, 0.13) from TLESUM$_{hp}$30 and the representative vector (0.70, 0.03, -0.01, -0.72) from TLESUM$_{hp}$95. As vector (-0.68, -0.18, 0.18, 0.68) from TLESUM$_{hp}$35 can been largely deemed as the negative vector from TLESUM$_{hp}\theta$ ($\theta > 35$), a transition of $\mathbf{Q_1}$ happens at sequence identity $\approx$30%. Mutations between h_p and p_h are conserved as sequence identity $>$30. As sequence identity $\leq$30, interchange h_h $\leftrightarrow$ p_p is favored.

Vector $\mathbf{Q_1}$ is vital to the similarity of LHI. As sequence identity $>$30%, after simple translation and rescaling, $\mathbf{Q_1^{>30}}$ is nearly equal to vector $\mathbf{G} = (2, 1, 1, 0)$

Table 1 Vector $\mathbf{Q_1}$ derived from the first eigenvectors of TLESUM$_{hp}$ matrices. It describes the top significant factors for the conservation of the representative biological properties of a protein family

Matrices	$\mathbf{Q_1}$ *of the 1st eigenvector* (h_h, h_p, p_h, p_p)	$\lvert CC(\mathbf{Q_1 G})\rvert$	$\lvert CC(\mathbf{Q_1 f})\rvert$
TLESUM$_{hp}$30	(−0.13, −0.69, 0.70, 0.13)	0.18	0.98
TLESUM$_{hp}$35	(−0.68, −0.18, 0.18, 0.68)	0.97	0.26
TLESUM$_{hp}$40	(0.71, 0.06, −0.07, −0.70)	1.00	0.09
TLESUM$_{hp}$45	(0.71, 0.04, −0.04, −0.71)	1.00	0.06
TLESUM$_{hp}$50	(0.71, 0.03, −0.03, −0.71)	1.00	0.04
TLESUM$_{hp}$55	(0.70, 0.02, −0.01, −0.71)	1.00	0.02
TLESUM$_{hp}$60	(0.70, 0.02, −0.00, −0.71)	1.00	0.01
TLESUM$_{hp}$65	(0.70, 0.03, −0.01, −0.72)	1.00	0.03
TLESUM$_{hp}$70	(0.69, 0.04, −0.02, −0.72)	1.00	0.04
TLESUM$_{hp}$75	(0.69, 0.04, −0.01, −0.72)	1.00	0.04
TLESUM$_{hp}$80	(0.69, 0.04, −0.01, −0.72)	1.00	0.04
TLESUM$_{hp}$85	(0.69, 0.05, −0.02, −0.72)	1.00	0.05
TLESUM$_{hp}$90	(0.69, 0.05, −0.02, −0.72)	1.00	0.05
TLESUM$_{hp}$95	(0.70, 0.03, −0.01, −0.72)	1.00	0.03

Analysis are applied on TLESUM$_{hp}$ matrices of different sequence identity levels. Absolute values of correlation coefficients (CC) for $\mathbf{Q_1} \leftrightarrow \mathbf{G}$ and $\mathbf{Q_1} \leftrightarrow \mathbf{f}$ are also shown.

which describes the number of hydrophobic residue in LHI (h_h, h_p, p_h, p_p). Namely, mutations with few difference in the component of $\mathbf{G}$ is favored, i.e., number of hydrophobic residue is dominantly conserved in LHI substitution as sequence identity >30. As hydrophobic residue transfers into any aqueous solution, water tends to form ordered cages around the non-polar molecule. This leads to a decrease in entropy, i.e., an increase in free-energy. By neglecting the detail residue type, free-energy of hydrophobic residue is roughly evaluated by value 0.35. Polar residue has value −0.35. Free-energy of a LHI is calculated by summing the contributions of the two edge residues independently and the resulting vector is (0.70, 0, 0, −0.70), nearly identical to vectors from TLESUM$_{hp}\theta$ ($\theta > 35$). Therefore, vector $\mathbf{Q_1^{>30}}$ and $\mathbf{G}$ are rough descriptions of the energetic item for different LHI types.

Although the meaning of $\mathbf{Q_1^{>30}}$ is obvious, implication of $\mathbf{Q_1^{\leq 30}}$ is not. To clarify meaning of this vector, we introduce internal hydrophobic force $\mathbf{f}$, a coarse-grained physical quantity contributed by an LHI. In an aqueous solution, water molecules attract one other, and have the effect of squeezing the hydrophobic residue. On the contrary, no such force is loaded on a polar residue. Then, for a residue pair owning single hydrophobic residue, there is a non zero resultant force along their virtual line. For example, in LHI h_p, hydrophobic force squeezes the hydrophobic side of the triplet. No such effect exists at 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, hydrophobic force along the virtual line of a residue pair is defined as $f_{ij} = 1$ if $a_i = h$ and $a_j = p$, $f_{ij} = -1$ if $a_i = p$ and $a_j = h$, and $f_{ij} = 0$ if $a_i = a_j$; 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(from C to N terminal), force of p_h is defined as -1. Since the solution contributes nearly equal but opposite forces on the two residues, the resultant force along the virtual line is approximately zero for the h_h case. So, LHIs with identical type of members(h_h or p_p) are considered to receive 0 resultant force along the virtual line. If we rescale the obtained vector $\mathbf{f} = (0, 1, -1, 0)$ into unit vector $\mathbf{f}' = (0, 0.71, -0.71, 0)$, $\mathbf{Q_1^{\leq 30}} \approx -\mathbf{f}'$. Data of linear regression show that this relationship is quite undoubted. So mutations with few difference in hydrophobic force $\mathbf{f}$ is dominantly conserved for the similarity of LHI as sequence identity ≤ 30.

4 Discussion

With a coarse-grained approach, we rank the importance of different physical quantities according to their contribution to the conservation of family specific bio-properties. The analysis is based on statistics of thousands sets of un-gapped multi-aligned homologous polypeptides. Consequently, this result adapts to most protein catalogues.

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. Consequently, the importance of energetic properties is usually first observed in near homologs. Whereas, this could not promise the energetic item a governing role in protein evolution. As protein evolve in a gradual manner generation by generation, a observation across short period of evolvement is weak in identifying the significant underlying mechanism that is conserved throughout long process of evolution. In remote homologues, the contribution from identical residue is not high anymore. As the 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. In this work, we identify the force vector a top weighted physical quantity for remote homologues. As the intramolecular residue-residue forces contribute the mechanics properties to a protein molecule, it indicates that mechanics properties contribute dominantly to the generation of remote homologs.

Our suggestion has been proven to be correct by a series subsequent works (Liu and Zhao, 2009a, 2009b, Liu and Zhao, 2010b, 2010c, 2010d), both in theoretical and in experimental one. When all residue-residue forces are considered, a complicated intramolecular force network can be obtained for each protein. Such network is a theoretical representation of protein's mechanics properties. Due to the aforementioned indication, property of such intramolecular network should contribute much to the conservation of family representative bio-properties. Based on this idea, Liu and Zhao have developed a simple 2-letter model in an aspect of hydrophobic force, by suggesting that there are some common and representative family characters in the intramolecular force networks of homologous proteins, which eventually govern the conservation of biological properties during protein evolution (Liu and Zhao, 2009a).

Liu and Zhao extend the definition of residue-residue hydrophobic force from triplet to quintuplet. For each quintuplet, after drawing the $C_5^2 = 10$

residue-to-residue virtual lines, they get a graph of the local network of residue-residue pairwise interaction. They define force state on each virtual line according to us. Then a local network of hydrophobic force is obtained for each quintuplet. Network of a quintuplet is a subset of the global intramolecular hydrophobic force network of corresponding protein. 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 quintuplet is characterised by the deviation of its inbuilt network from that of background. Then significance of a sequence, i.e., the propensity to be a member of corresponding protein family can be scored by mean of quintuplets' significance.

The algorithm can boost up capability of the existing tools in multiple sequence alignment(at least 50%) (Liu and Zhao, 2009a), and has successfully been used in uncovering the detailed donut-shaped topological feature of the polypeptide relationship (Liu and Zhao, 2009b), identifying the significant sites responsible for switching on the pathogenic structural changes in conformational disease (Liu and Zhao, 2010b, 2010c). Moreover, in order to prove the vital role of mechanics properties, they also designed some remote artificial members of WW domain family based on this fully computational approach exclusively (Liu and Zhao, 2010d). The bioactivities of the new members were confirmed with ligand-binding experiments. All the artificial members share similar function and folding with their natural counterparts, i.e., the conservation of mechanics properties is validated as a sufficient condition for designing proteins of WW domain. In the protein design of remote homologs, the trivial source for homolog's similarity is weighted less, the basic requirement of protein evolution contributes dominantly. In their study, molecular mechanics property is the only factor in remote homolog design, achieving a success. Consequently, it indicates that molecular mechanics property is the governing feature in generating new members of a protein family during evolution. Since protein dynamism was also suggested to be the foundation of protein evolvability (Tokuriki and Tawfik, 2009), it is clear that molecular mechanics properties are quite significant for protein evolution.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, pp.3389–3402.

Bloom, J.D., Wilke, C.O., Arnold, F.H. and Adami, C. (2004) 'Stability and the evolvability of function in a model protein', *Biophysical J.*, Vol. 86, pp.2758–2764.

Dill, K.A. (1990) 'Dominant forces in protein folding', *Biochemistry*, Vol. 29, pp.7133–7155.

Fiser, A. and Sali, A. (2003) 'Modeller: generation and refinement of homology-based protein structure models', *Methods Enzymol.*, Vol. 374, pp.461–491.

Gupta, N. and Irbäck, A. (2004) 'Coupled folding-binding versus docking: a lattice model study', *J. Chem. Phys.*, Vol. 120, pp.3983–3989.

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

Henikoff, S. and Henikoff, J.G. (1992) 'Amino acid substitution m atrices from protein blocks', *Proc. Natl. Acad. Sci.*, Vol. 89, pp.10915–10919.

Jones, S. and Thomton, J.M. (1996) 'Principles of protein-protein interactions', *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp.13–20.

Li, H., Helling, R., Tang, C. and Wingreen, N.(1996) ' Emergence of preferred structures in a simple model of protein folding', *Science*, Vol. 273, pp.666–669.

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

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.*, Vol. 79, pp.765–768.

Liu, X. and Zhao, Y.P. (2009a) 'A scheme for multiple sequences alignment optimization – an improvement based on family representative mechanics features', *J. Theor. Biol.*, Vol. 261, pp.593–597.

Liu, X. and Zhao, Y.P. (2009b) 'Donut-shaped fingerprint in homologous polypeptide relationships – a topological feature related to pathogenic structural conversion of conformational disease', *J. Theor. Biol.*, Vol. 258, pp.294–301.

Liu, X. and Zhao, Y.P. (2010a) 'Substitution matrices of residue triplets derived from protein blocks', *J. Comput. Biol.*, Vol. 17, pp.1679–1687.

Liu, X. and Zhao, Y.P. (2010b) 'Switch region for pathogenic structural change in conformational disease and its prediction', *Plos one*, Vol. 5, pp.e8441.

Liu, X. and Zhao, Y.P. (2010c) 'Simulated pathogenic conformational switch regions matched well with the biochemical findings', *J. Biomed. Inform.*, Vol. 43, pp.365–375.

Liu, X. and Zhao, Y.P. (2010d) 'Generating artificial homologous proteins according to the representative family character in molecular mechanics properties – an attempt in validating an underlying rule of protein evolution', *FEBS Lett.*, Vol. 584, pp.1059–1065.

Liu, X., Zhao, Y.P. and Zheng, W.M. (2008) 'CLeMAPS: multiple alignment of protein structures based on conformational letters', *Proteins*, Vol. 71, pp.728–736.

Liu, X., Zhang, L.M., Guan, S. and Zheng, W.M. (2003) 'Distances and classification of amino acids for different protein secondary structures', *Phys. Rev. E*, Vol. 67, pp.051927.

Taverna, D.M. and Goldstein, R.A. (2002) 'Why are proteins marginally stable?', *Proteins*, Vol. 46, pp.105–109.

Tokuriki, N. and Tawfik, D.S. (2009) 'Protein dynamism and evolvability', *Science*, Vol. 324, pp.203–207.

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

Zhang, Y. (2008) 'I-TASSER server for protein 3D structure prediction', *BMC Bioinformatics*, Vol. 9, p.40.