FEBS Letters 584 (2010) 1059–1065

FEBS Letters
journal homepage: www.FEBSLetters.org

Generating artificial homologous proteins according to the representative family character in *molecular mechanics properties* – an attempt in validating an underlying rule of protein evolution

Xin Liu, Ya-Pu Zhao *

The State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 5 January 2010
Revised 1 February 2010
Accepted 2 February 2010
Available online 9 February 2010

Edited by Takashi Gojobori

Keywords:
Protein evolution
Protein design
Hydrophobic interaction
Hydration
Binding assay

ABSTRACT

The *molecular mechanics property* is the foundation of many characters of proteins. Based on intramolecular hydrophobic force network, the representative family character underlying a protein's *mechanics property* is described by a simple two-letter scheme. The tendency of a sequence to become a member of a protein family is scored according to this mathematical representation. *Remote homologs of the WW-domain family* could be easily designed using such a mechanistic signature of protein homology. Experimental *validation* showed that nearly *all* artificial homologs have the representative folding and bioactivity of their assigned family. Since the *molecular mechanics property* is the only consideration in this study, *the results indicate its possible role in the generation* of new members of a protein family during evolution.

1. Introduction

The amino acid sequence is believed to specify a protein's atomic structure and biological function [1]. Proteins are diversiform due to differences in residue sequence. Although their compositions are quite different, some proteins share common biological properties with one another. For instance, some remotely homologous proteins can have less than 30% identical residues. However, the reason for such functional uniformity, which arises from the diversity of intramolecular details, is still unknown.

Two levels of studies are related to protein homology research: investigations of a single physical system and those of the uniformity of multiple systems.

(i) The first type refers to studies focusing on the properties of a biomolecule-solution physical system, including the native fold, function, and conformational motion. Since only one system is investigated, the basic and universal physical principles, quantities, and methods are applicable in this type of studies. For example, the free energy of the physical system per protein is believed to play a vital role in protein folding [2,3].

(ii) The second type refers to studies focusing on the reasons behind the occurrence of homology. This includes the present study on why the folding that generates the representative family biological properties, but not other decoy folds, is specified as the native structure of a protein family member. This type of studies usually focuses on something common within a homolog set, and embodies the selection pressure during the process of choosing the eligible molecules from the outcome of the basic physical principle.

As each protein corresponds to a physical system, a set of systems must be jointly investigated so that some common mechanisms within these systems can be identified. Since multiple systems are simultaneously focused on, the methodology will be different from that of a single physical system. For example, as compositions differ across homologs, the residue interactions that contribute free energies should also vary in their corresponding physical systems, especially among those of remote homologs. Consequently, the similarity in free energy is not a necessary condition in protein homology, and the importance of free energy is ultimately decreased. Therefore, it is rational that the fundamental physical principle focusing on the homology of protein evolution is based on, but not limited to, those at single-system level. At present, due to such a shift in the object of research, there is still a gap between the physical principles of the two levels. In this paper, we present a novel level of description for a multiple system, and at-

* Corresponding author. Address: Beisihuanxi Road #15, Beijing, China.
E-mail addresses: liuxin@lnm.imech.ac.cn (X. Liu), yzhao@imech.ac.cn (Y.-P. Zhao).

0014-5793/$36.00 © 2010 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2010.02.010

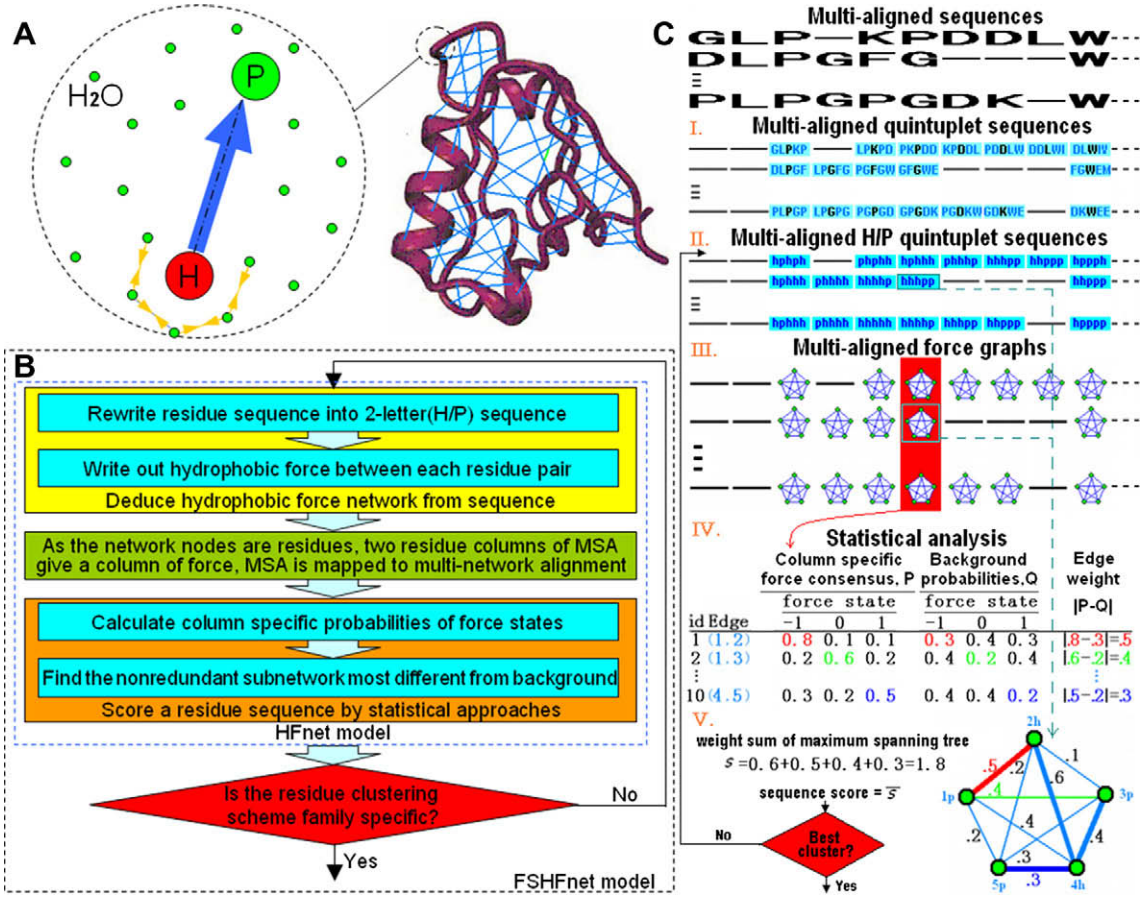

Fig. 1. Illustrations of hydrophobic force and the FSHFnet algorithm. The hydrophobic force along the virtual line of an H–P residue pair is shown in (A), with the sketch map of the hydrophobic force's origin wherein the attractions between water molecules are denoted in gold. As indicated in the flowchart (B), we tried each kind of clustering scheme and evaluated the performance of HFnet. The clustering scheme with the maximum counts of correctly identified samples in the learning set was selected as the family specific amino acid classification scheme in FSHFnet. Some details of FSHFnet are shown by examples in (C), including: (I) rewriting protein sequence into successive overlapping 5-residue units; (II) rewriting quintuplet sequence into H/P quintuplet sequence; (III) drawing a force graph of residue-to-residue interaction in each H/P quintuplet; (IV) calculating edge-specific probabilities of the occurrences of force states for each column of aligned graphs, and those of the background that are evaluated by all graphs in a background sequence set; and (V) finding the maximum spanning tree for each graph (the tree is shown in bold line, with the edge weight reflecting the difference between the occurrence of a force state and that of the background), scoring a sequence, and evaluating the residue clustering scheme.

tempt to make a step in closing up the two with a simple, empirical, but physics-based, mathematical representation.

Evolution has been the focus of protein science for a long time. Many efforts have been devoted to the study of sequence [4,5], structure [6,7] and function [8], which are biological properties that are suitable for a direct comprehension and are relatively easy to observe. But some evolutionary events can not be fully investigated without the analysis of physical mechanism. For instance, the native-structure absent homologs of a disorder protein family can carry out their biological functions by dynamic conformational changes. Since structure or distance is no longer important in these cases, the biological properties of these homologs should be determined by an physical quantity that is responsible for the change of movement state of polypeptide or the change of movement tendency, that is, the force. The molecular mechanics property may be more conserved than the structure [9]. Therefore, there is a requirement to investigate protein evolution in an aspect of physics. In particular, the molecular mechanics property is the basis of side-chain fluctuations, movement of active site loops, structural exchanges and rearrangements, and other processes that are vital to protein biological properties. The investigation of such property has been regarded as a new hotspot of protein evolution [10].

There are vast complexities of interactions in the protein that can be coped with quantum mechanical, molecular mechanical, or other treatments. As multiple systems are jointly investigated, the complexity increases drastically in the study of protein homology. To reduce the difficulty involved, a feasible option is to adopt a coarse-grained scheme that focuses on significant items but still monitors the secondary factors.

Hydrophobic interactions have been suggested as the driving force of protein folding [3], and play an important role in protein function [11]. In an aqueous solution, a hydration shell is formed on a protein surface by at least two layers of water molecules [12]. The water molecules that surround a hydrophobic (H) residue attract one another, resulting in a radial compressive stress on the amino acid. No such force is loaded on a polar (P) residue. As shown in Fig. 1A, this results in a force between each residue pair, and subsequently, a complicated force network in each protein molecule. This network is a representation of the consequence of hydration in a corresponding physical system.

In agreement with Frauenfelder's observation that internal protein motions or dynamic properties are controlled by the hydration shell, we suggest that there are some common and representative family characters in the inbuilt force networks of homologous proteins, which eventually govern the conservation of biological properties during protein evolution [13]. The maintenance of these characters would serve as the fundamental physical principle that potentially governs protein homology. We believe that if this

theoretical basis is correct, it should be possible to build the artificial members of a family accordingly.

The protein design of the remote homologs of a family needs an efficient process of identifying the eligible non-redundant candidate from a huge (20^N) sequence space. Therefore, algorithms based on common basic principles are required. Thus, the design of remote homologs is suitable for illustrating the validity of a theory that focuses on the underlying rule of protein evolution.

In this study, we introduced the representative family characters of a hydrophobic force network to the protein design, and attempted to produce the remote artificial members of a protein family using the family specific hydrophobic force network model (FSHFnet), which is a fully computational approach. We then confirmed the bioactivities of the new members with ligand-binding experiments. The WW domain, a computationally and experimentally simple model system, was used to test the feasibility of this approach. Since all of the artificial members share similar function and folding with their natural counterparts, our scheme was proven effective in catching the evolutionary information that governs the conservation of family-representative biological properties such as structure and natural function.

2. Materials and methods

2.1. Hydrophobic Force network model (HFnet)

The mathematical representation of the representative family characters of the hydrophobic force network is deduced from multiple sequence alignment (MSA).

First, the hydrophobic force network is deduced from the sequence. As illustrated in Fig. 1A, we can determine the coarse-grained hydrophobic force between a residue pair after the residue sequence is rewritten into H/P sequence. The state of the resultant hydrophobic force, along the virtual line between residues i and j ($j > i$), can be written as (0, +1, −1, 0), wherein the types of ij include HH, HP, PH or PP, respectively [13]. A positive sign means that the resultant force is pointing towards the C terminal. 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 HH case [14]. Then considering each residue pair, we obtained a representation of the inbuilt network per protein from sequence information, wherein the residues are treated as the nodes of the network. The residue-residue virtual lines are network edges.

Second, a multiple network alignment is conducted from multiple sequence alignment. As two joint residue columns of MSA correspond to a column of forces, these forces also become aligned, column by column. This results in the creation of a map from MSA to the multiple network alignment.

Finally, a residue sequence is scored using statistical approaches. We characterize the representative feature of a set of force-networks/proteins by the consensus of its set members. Since the force network of a biologically significant protein should be extremely different from that of an unrelated family, we scored the tendency of a sequence to be a member of a protein family by the deviation of its inbuilt network vis-à-vis its background (see Appendix for details). The sequences were ranked in decreasing order of their score. A high-scoring sequence was likely to be a member of the corresponding protein family.

Several applications showed that the HFnet algorithm can perform very well even with an unoptimized residue classification scheme [15]. For instance, it has been successfully used in uncovering the detailed donut-shaped topological feature of the polypeptide relationship [16], identifying the significant sites responsible for the initial pathogenic structural changes in conformational disease [17], and boosting up the capability of existing tools in multiple sequence alignment [14].

2.2. Family Specific Hydrophobic Force network model (FSHFnet)

The goal of this study is to design new members of a protein family according to the HFnet. While there are different functional groups on a residue, a different exposed/buried state results in a different biochemical property. The scheme of residue classification (H/P) should be family specific in HFnet. Therefore, this study attempts to determine a clustering scheme specific for the WW domain. The 42 natural sequences of the WW domain collected and aligned by Socolich et al. were selected as the basis of our learning set [18], 28 of which are natively folded, true samples (TS). Since the inverse transformation of an H/P classification scheme is unresponsive to the mathematical approach HFnet, there are a total of $2^{19} = 524\,288$ kinds of non-redundant clustering schemes. We tried all these schemes and different selections for the threshold T, and found a scheme with the maximum score of success $Z = \sum_n \delta(if(S^n > T), if(n \in \{TS\}))$ in identifying true and false signals. In this formula, S^n is the HFnet score for sequence n, and $if(\text{True}) = 1$ and $if(\text{False}) = 0$, $\delta(x,y)$ are the step functions with $\delta(x,y) = 1$ for $x = y$ and $\delta(x,y) = 0$ otherwise. With each classification scheme η, we carried out the HFnet and used different cut-off data to identify a threshold T_η with the maximum success counts $Z^{T\eta}_{\eta\_Max}$. After each of the clustering schemes was scanned, we identified the best residue clustering scheme for the WW domain (P = {M, I, L, W, C, N, Q, E, K}, H = {F, V, A, P, D, R, G, H, S, T, Y}, with Max $(Z^{T\eta}_{\eta\_Max}) = 39$ and $T = 1.347$). Thirty-nine out of the 42 sequences could be correctly identified.

According to Livingstone and Barton [19] (shown in Fig. 2 insertion), the typical hydrophobic residues {F, V, A, P, G, H, T, Y} and polar/charged residues {W, C, N, Q, E, K} are properly clustered in our classification scheme. A further consideration showed that the substitutability among residues is also significant for the clustering. To display the consensus sequences of the WW domain, the sequence logos of the learning set [20] is shown in Fig. 2. At positions 14 and 12, residues I and M are most abundant and have high propensity in substitution with residues K and E, respectively. As such, I and M are grouped with K and E, that is, into polar class. Similarly, according to the abundant substitutions R ↔ A, D ↔ R, S ↔ T in positions 13, 35 and 15, residues Y, R, and D are classified into hydrophobic class. When a 5-residue window is sliding along a sequence, the residues at the N- and C-terminals are covered at a lesser time than that of the other sites. Therefore, residue L at position 2 has a small weight and contributes weakly to residue classification. Residue L may be clustered into the polar class owing to the abundant substitution L ↔ I at position 14.

2.3. Homolog generation

We assumed that an artificial sequence is aligned with the sequences in the MSA of the learning set. Before assigning a residue, however, each site is in blank. For each blank, we assigned a residue selected randomly from those that appeared in the corresponding residue column of MSA. Since all candidate residue types in the column are treated equally, no type is considered to be the key residue according to our algorithm. If the sequence designed has a pairwise sequence identity (SI) of no less than 30% with any of the learning set or with any of the artificial proteins that have been accepted in the former process, it is rejected. Otherwise, we further evaluated its significance using the FSHFnet algorithm. It is accepted as a candidate of the remote homolog of the WW domain if the score is more than the threshold T_θ; otherwise, the sequence was rejected. New sequences are then written further.

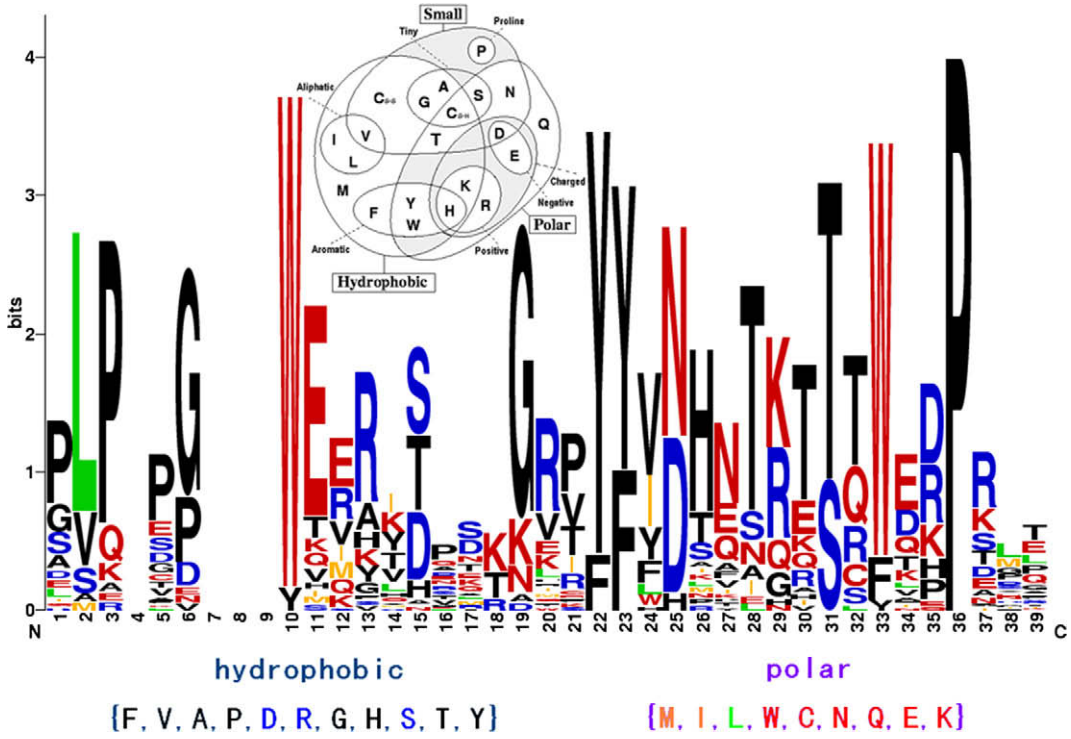

Fig. 2. The consensus sequences of the WW domain shown by sequence logos [20], a graphical representation of multiple sequence alignment. The overall height of the stack indicates the sequence conservation at that position, while the height of the symbols within the stack indicates the relative frequency of each amino acid at that position. Insertion: Amino acid classification according to physical-chemical properties. This has been proposed by Livingstone and Barton [19], and is probably a scheme to which most people would agree.

Through this method, each protein designed has a low pairwise SI (<30%), either with each other or with each of the proteins in the learning set. Since our preparative research has shown that the structural deviations of the proteins designed decrease linearly with the increase in FSHFnet score as compared to the wild-type (WT) protein, we chose an increased threshold $T_\theta = 1.4$ to improve the success rate. Due to the limitation in the search scope of sampling, only five candidates could be produced by each initial randomization for such a strict scheme. These five candidates of remote homologs could characterize the maximum scope of non-redundant sampling.

2.4. Structure prediction

Three-dimensional structures of the five candidates of artificial remote homologs are predicted by the I-TASSER algorithm [21]. This threading algorithm is developed by Zhang and is deemed to be the most successful in protein structure prediction [22,23]. For simple cases such as the WW domain, it has been reported that as compared to the real structure, the algorithm has an accuracy of less than 3 Å in root mean squared deviation (RMSD). Therefore, I-TASSER is considered suitable for predicting the overall fold of the present five candidates. In this study, I-TASSER was only used after the candidates had been generated and not during protein design.

2.5. Binding assays

The natural proteins of the WW domain recognize proline-rich-target peptides to carry out their function. This recognition is a family-representative biological property of the members of the WW domain's. Based on the target peptide sequence motifs, WW domains members are classified into four groups: I, PPxY; II, PPLP; III, PPR; and IV, pS/pT-P, where x stands for any amino acid, and pS/pT-P refers to a phosphoserine/phosphothreonine-proline containing peptide [18]. The protein function of the artificial homolog was tested using experimental approaches. Since the proteins in the learning set are mostly of group-I and group-III types, we examined the binding affinities of the proteins designed to group-I (GTPPPPYTVG) and group-III peptides (PPGPPPRGPPP).

During the sample preparation, all proteins and peptides were synthesized with solid phase peptide synthesis using Fmoc-Chemistry from SBS Genetech Co. Ltd. The high-purity (>95%) samples are dissolved into aqueous solutions.

Isothermal titration calorimetry (ITC) measurements were made using a NANOITC^{2G} calorimeter (TA) at 25 °C, starting with 240–720 μM of a WW domain protein in the sample cell and titrating 4–8 mM of the group-I or group-III peptide. After detecting the heat effects of protein binding, the data gathered were fit using the NaNoAnalysis software provided by the manufacturer, such that the binding constants can be calculated. We measured the binding constant K_D of a protein to each type of ligand or peptide.

3. Results

3.1. Homologous level of the protein designed

According to the methods of homolog generation, the maximum sequence identity is 29.4% as a protein designed is compared with any of the training proteins. Compared with the proteins of the learning set, the mean sequence identities range from 21.6% to 23.7% (Table 1). The lowest homology to any of the training structures is 8.8%. Therefore, the proteins designed have a distant homology with those of the training set.

3.2. Structure of the proteins designed

According to the I-TASSER algorithm, the deviation in structural prediction is at most 2.4 ± 1.8 Å in RMSD (as seen in the "I-TASSER

Table 1
Binding constants (K_D) of the artificial remote homologs of the WW domain for each type of ligand. The mean and minimum sequence identities of the artificial proteins are reported compared with those of the learning set. The maximum identity is 29.4% for each artificial protein. The accuracies of the structures predicted by the I-TASSER algorithm [21] are listed, together with the RMDS deviations of these structures compared with that of a wild-type (WT) protein (PDBID 1WR4). For the positive control, we measured the affinities of the WT protein 1WR4 to be 75.2 ± 0.4 μM for the group-I peptide, and weak (not available, N/A) for the group-III peptide. As shown in this table, most artificial remote homologs belong to the group-III type. It is reasonable. Since most native-folded proteins in the learning set are in the group-I type, the search space for type-I protein is smaller than that for group-III type during the generation of artificial residue sequence. Thus, more type-III proteins are produced.

Protein sequences	Mean SI (%)	Min SI (%)	RMSD (Å)		K_D (μM) of group-I peptide	K_D (μM) of group-III peptide
			I-TASSER accuracy	Compared with WT		
1:LSAPPWSVFMTPAAHVFFYNSQEQQTTWQPPTSE	23.0	14.7	2.3 ± 1.8	1.8	232.6 ± 29.6	1.4 ± 0.2
2:EVRPDWQMHISDSALPFFLNKKANRSQWKDPTSK	21.6	11.8	2.0 ± 1.6	1.4	223.9 ± 6.4	N/A
3:GMKVPWEQVKHSKKKRFFVHMKTQKSSWQRPRLQ	23.7	14.7	1.9 ± 1.6	1.4	N/A	43.2 ± 5.4
4:IMQSDYEEHLTHMDVVFYHDSQIGTSTWIRPNTE	21.8	8.8	2.3 ± 1.8	1.8	37.7 ± 2.6	42.9 ± 2.7
5:SVASPYQQGIDRNGKPYFYHTNNRRSSWKRPGEH	22.6	14.7	2.4 ± 1.8	1.5	N/A	37.0 ± 0.9

Accuracy" column of Table 1). The structure prediction is therefore basically accurate. Moreover, when the structures predicted are compared with those of the WT protein, the maximum deviation (RMSD) is small at 1.8 Å. Since the binding assays further confirm the representative biological function of these artificial proteins, it can be deduced that the proteins designed fold into a typical structure of the WW domain.

3.3. Biological function of the protein designed

The ITC results of the first artificial protein are shown in Fig. 3. In this example, the binding constants $K_D = 1/K$ with group-I and group-III peptides are 217.8 μM and 1.37 μM, respectively. Therefore, this protein recognizes the group-III peptide in a high affinity, and exhibits modest affinity to the group-I peptide. The results of the other artificial proteins are listed in Table 1. All of the artificial proteins exhibit the ligand-binding affinities in a similar level to the WT protein. Four of them manifest high affinities to proline-rich target peptides. Therefore, all artificial remote homologs have the representative family biological properties of the WW domain.

In this study, we learned the representative mechanistic characteristics of a family from the learning set, and then extended this physical knowledge to new and non-redundant unknown samples in the protein phase space. The feasibility of such knowledge extension is verified by the experiments of protein binding. In 2005, Socolich et al. has designed some artificial proteins using the identical learning set of WW domain. In their work, the underlying sampling of sequence space has been investigated by designing sequence based on site-independent residue propensity, that is, randomly selecting residue at each position only from those that are present in the WW-domain family at that position. Experimental validation showed that none of the protein designed using underlying sampling is a natively folded protein, or belongs to WW-domain family. Since the site-independent residue propensity in the MSA is weak in determining the family specific folding of the WW protein [18], the success of the present work is a result of the underlying physical theory and algorithm.

4. Discussion

In biophysics, one fundamental viewpoint is that protein folding is driven by the hydrophobic interactions in the protein. There have been several works that made great contributions to the development of this theory. These include studies by Dill [2] and Li et al. [3], in which the interaction energies and different parameters are introduced to develop a class of coarse-grained model named the "HP model." Due to the significance of hydrophobic interaction, it is reasonable to deduce that there are certain representative family characters within the hydrophobic interaction systems of different homologous proteins. There is, however, a lack of theoretical description for such common features. This study converts this rational opinion into a practical mathematical approach, and links the hydrophobic interaction system per molecule and protein evolution physically.

As compared to the HP model, this study has two features. We used only two residue symbols, which is similar to the HP model. The focus of this study is to uncover the mechanism that the representative family biological properties are endowed to a candidate protein. The second aspect is about the physical quantity under study. For most HP models, the central physical quantity is free energy or interaction energy. Therefore, distance or the distance threshold must be introduced to evaluate the energetic property. Since the present algorithm focuses on the hydrophobic force vector, distance is no longer a necessary physical quantity here. Consequently, although the current approach uses just indices for HH/HP/PH/PP interaction, the physical quantity used to characterize the interaction is the force vector and not the energetic item as that of HP models. Moreover, due to the absence of distance, the indices of 0/+1/−1/0 are not meant to be meaningful in an energetic manner.

Protein design is an approach of inverse folding that requires an understanding of molecular interactions, which stabilize the proteins in a specific native fold. From a physical point of view, the sequence identified from an enormous sequence space must have the chosen structure as a free energy minimum. This can be done by the computational protein design algorithm, which uses energy functions to evaluate how mutations would affect a protein's structure and function [24–26]. A combination of molecular mechanic, knowledge-based, and empirical terms is generally used in typical energy functions [27]. A fast and accurate energy function is still a challenge in computational protein design. The present approach provides a new scheme that relies not on energy function but on the empirical force network of intramolecular interactions trained from the multi-alignment of homologous sequences. In this scheme, the only requirement is the use of some native homologous sequences of a family and other decoy sequences. Since structural information is not necessary, the requirement of protein design is not as restrictive as other methods. For example, the 42 natural sequences of the WW domain are clustered by sequence alignment, and only seven of them have structural data. The clustering of a training set can be accomplished quite well by sequence alignment tools. Therefore, the present scheme is a handy and powerful tool in protein design, and in candidate filtration. To the best of our knowledge, this is the first two-letter algorithm that can produce a very distant homolog with a high success rate of conserving the representative biological properties of a family.

Since only two letters are used in rewriting residue sequence, the present algorithm is extremely simple. An explicit and effective approach that represents the fundamental physical principle in

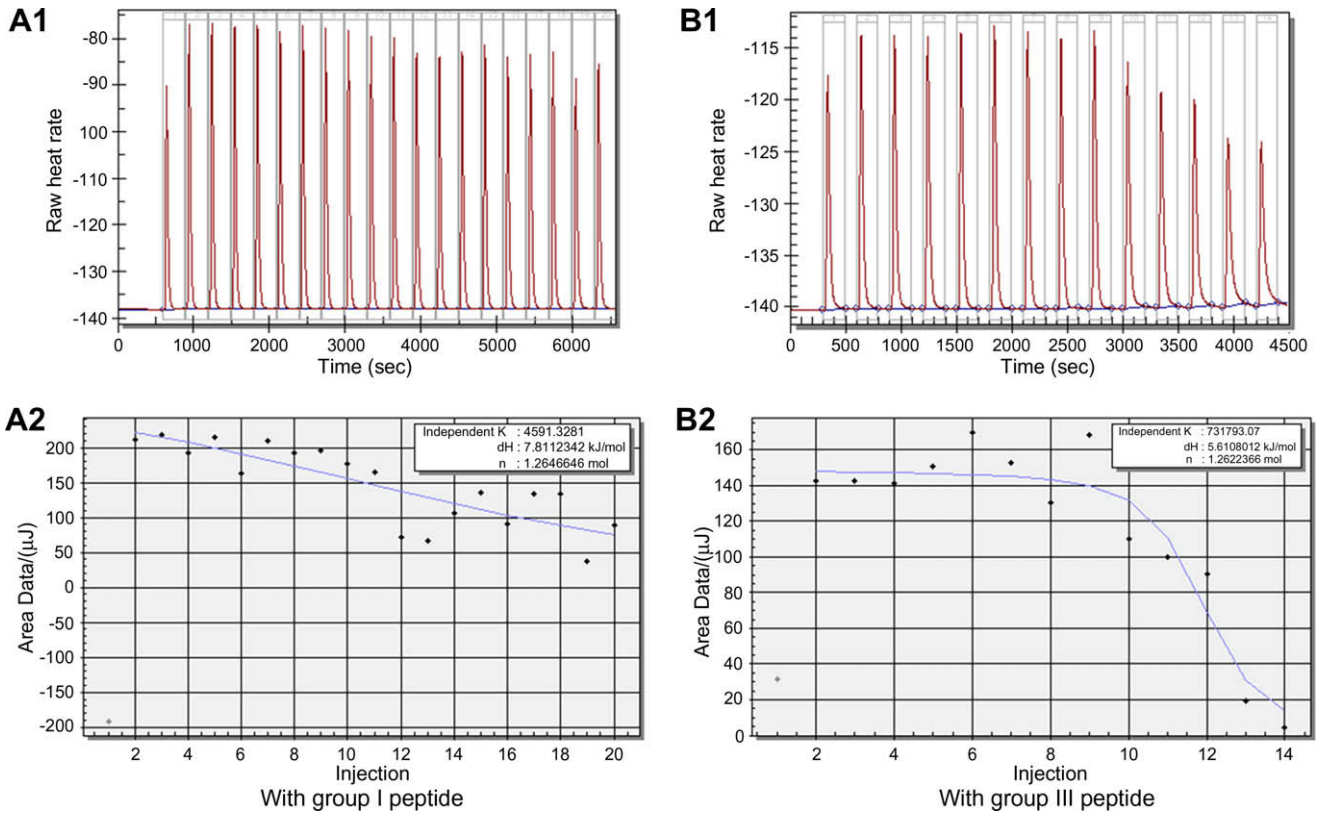

Fig. 3. Results of heat changes and data fitting made by the isothermal titration calorimeter NANOITC^{2G} and the software NaNoAnalysis, respectively. The binding assays of artificial protein 1 with group-I and group-III peptides are shown as examples. The binding constants can be calculated by $K_D = 1/K$. (A) The 500 μM artificial protein 1 was prepared in the sample cell. The 8 mM group-I peptide was injected into the cell with 5 μl per time, every 300 s. Data on heat changes are shown in (A1) and are fitted as reported in (A2). Data on the first injection were omitted because they are not accurate due to the initialization of the device. The results reported in (B1) and (B2) are of group-III peptide (5.3 mM) titrating into artificial protein 1 (250 μM), using a similar approach.

evolution is required for the success of such a simple model. Therefore, a simple approach using the H/P score of individual sites, but not the network, may not achieve similar results. Since the residue-residue interactions are most prominent within a window of five residues [28], the algorithm works well with quintuplet. Protein design is accomplished for the WW domain where long distance interaction, such as the S–S bond, is not conserved. Although more factors may be considered for the conservation of long distance interaction, we believe that the residue quintuplet should contribute the dominant information.

In this study, the molecular mechanics property contributed by hydrophobic interaction is the only factor we considered. The results of our approach showed that the representative family character of the molecular mechanics property is sufficient for the generation of new members of the WW domain. We propose that the mechanistic feature, especially that contributed by hydrophobic effects, is important in governing protein evolution.

There are two aspects of protein evolution: variability and robustness. Variability, also called "protein evolvability," is the capability of proteins to rapidly adopt a new function within existing folds, or even adopt entirely new folds. Meanwhile, robustness is the capability of proteins to conserve representative family biological properties and thus some native biological functions can be performed. The scientific viewpoints are inconsistent: some evidences show that protein structures are changeable in sequence; while others claim that few mutations can lead to big changes in protein structure [29–31]. It is a widely discussed problem how structural robustness and innovation can exist at the same time. Recently, scientists tend to override the puzzling problem related to the variation of structures exhibited in native state, but reveal the common cause of these phenomena in a mechanistic view so that an in-depth uniform perspective can be achieved [32,10]. Focusing on variability, Tokuriki and Tawfik suggested an "avant-garde view" in 2009 that protein dynamism, which mediates alternate folds and functions, is the foundation of "protein evolvability." The present study suggests that the conservation of the molecular mechanics property is a central descriptor of protein robustness. It indicates that the mechanistic feature is the foundation of protein evolution and is thus worthy of further research.

Acknowledgements

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.

Appendix. Details of the Hydrophobic Force network model (HFnet)

Since the cost of a complete consideration of the forces in an N residue protein is extremely high (C_N^2 forces), we decided to focus on a localized network. A protein sequence is treated as successive overlapping five-residue units. As shown in Fig. 1C, we assigned the first quintuplet to the third residue, the second to the fourth, and so on, until finally, the last quintuplet was assigned to the last residue but two. A quintuplet was assigned to a position corresponding to its central residue. If there is a gap in the MSA, we also insert a gap in the quintuplet sequence alignment. Therefore, the

quintuplet sequences are aligned with the same gap positions as the original MSA, but with four additional gaps on the first two and last two sites in each sequence. A column of residues in MSA corresponds to a column of aligned quintuplets, with the central residues identical to those of the residue column.

In each unit, the $C_5^2 = 10$ residue-residue virtual lines (Edges, E) and the five residues (Vertices, V) form a complete graph G = (V, E). In this way, the representation of the whole force network is simplified to a sequence of successive graphs with a definite force state along each edge. Since these force graphs are also aligned due to multiple network alignment, for edge (i, j) of graph column k, the probability of the occurrence of force state l can be calculated as

$$P_{ij}^{k}(l) = \frac{\sum_{n=1, G^{nk} \notin gap} \delta(l, f_{ij}^{nk})}{\sum_{n=1, G^{nk} \notin gap} 1} \quad (1)$$

wherein n is the sequence index, the step function $\delta(x,y)$ equals 1 for x = y and $\delta(x,y) = 0$ otherwise, and f_{ij}^{nk} is the force state in force graph G^{nk}. For a sequence set, there are specific statistical features that can be considered as the commonness of its members, that is, the background feature. To describe such trivial feature, a model of background force graph can be constructed using the consensus of all graphs contained in a background set B. The probability of force state l contributed by the background can be calculated as

$$Q_{ij}(l) = \frac{\sum_{G \in B} \delta(l, f_{ij}^{G})}{\sum_{G \in B} 1} \quad (2)$$

To evaluate the difference between the occurrence of force state f_{ij}^{nk} and that contributed by the trivial background, $D_{ij}^{nk} = |P_{ij}^{k}(f_{ij}^{nk}) - Q_{ij}(f_{ij}^{nk})|$ is introduced as the weight of edge (i, j) in graph G^{nk}. We believe that the force network of a biologically significant protein should be remarkably different from that of the background. Therefore, we find the most significant non-redundant interactions by identifying the maximum spanning tree (MST) [33] in each force graph G^{nk}. The weight sum $s(G^{nk})$ of MST is introduced as a description of the deviation. Then the significance of sequence n is scored by the mean of deviation as

$$S^{n} = \frac{\sum_{k, G^{nk} \notin gap} s(G^{nk})}{\sum_{k, G^{nk} \notin gap} 1} \quad (3)$$

In evaluating Q, the whole aligned sequences are first used as background data set. The protein sequences are then arrayed by score S^n in a descending order. The top-ranked sequences are deemed to be more significant than those at the end. We update Q by using the sequences at the bottom as the new background set. Such background set can be deemed unrelated to the family interested. Then we score and rank each sequence again until convergence is achieved. As shown in Eq. (1), there are $(3-1) \times C_5^2 = 20$ probabilities to be calculated in each graph. Since there are only about $20 \times N$ parameters estimated for an N-residue fold, the present scheme is extremely simple. Although only sequence information is used, based on the complete graph, the present scheme is a model of a three-dimensional network.

References

[1] Anfinsen, C.B., Haber, E., Sela, M. and White Jr., F.H. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. USA 47, 1309–1314.

[2] Dill, K.A. (1990) Dominant forces in protein folding. Biochemistry 29, 7133–7155.

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

[4] Trifonov, E.N., Kirzhner, A., Kirzhner, V.M. and Berezovsky, I.N. (2001) Distinct stages of protein evolution as suggested by protein sequence analysis. J. Mol. Evol. 53, 394–401.

[5] Qi, J., Wang, B. and Hao, B.L. (2004) Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. J. Mol. Evol. 58, 1–11.

[6] Orengo, C.A. and Thornton, J.M. (2005) Protein families and their evolution-A structure perspective. Annu. Rev. Biochem. 74, 867–900.

[7] Zeldovich, K.B., Berezovsky, I.N. and Shakhnovich, E.I. (2006) Physical origins of protein superfamilies. J. Mol. Biol. 357, 1335–1343.

[8] Kinch, L.N. and Grishin, N.V. (2002) Evolution of protein structures and functions. Curr. Opin. Struct. Biol. 12, 400–408.

[9] Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.

[10] Tokuriki, N. and Tawfik, D.S. (2009) Protein dynamism and evolvability. Science 324, 203–207.

[11] Jones, S. and Thomton, J.M. (1996) Principles of protein-protein interactions. Proc. Natl. Acad. Sci. 93, 13–20.

[12] Frauenfelder, H., Chen, G., Berendzen, J., Fenimore, P.W., Jansson, H., McMahon, B.H., Stroe, I.R., Swenson, J. and Young, R.D. (2009) A unified model of protein dynamics. Proc. Natl. Acad. Sci. 106, 5129–5134.

[13] Liu, X., Zhang, L.M., Yin, J. and Zhao, Y.P. (2008) Major factors of protein evolution revealed by eigenvalue decomposition analysis in Proceeding of the International Conference on Bioinformatics and Computational Biology BIOCOMP'08, Las Vegas, CSREA Press, USA. pp. 91–97.

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

[15] 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 66, 021906.

[16] Liu, X. and 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.

[17] Liu, X. and Zhao, Y.P. (2010) Switch region for pathogenic structural change in conformational disease and its prediction. PLoS One 5, e8441.

[18] Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H. and Ranganathan, R. (2005) Evolutionary information for specifying a protein fold. Nature 437, 512–518.

[19] Livingstone, C.D. and Barton, G.C. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biol. Sci. 9, 745–756.

[20] Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

[21] Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9, 40.

[22] Zhang, Y. (2009) I-TASSER: fully automated protein structure prediction in CASP8. Proteins 77, 100–113.

[23] Zhang, Y. (2008) Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18, 342–348.

[24] Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368.

[25] Looger, L.L., Dwyer, M.A., Smith, J.J. and Hellinga, H.W. (2003) Computational design of receptor and sensor proteins with novel functions. Nature 423, 185–190.

[26] Liu, S., Liu, S.Y., Zhu, X.L., Liang, H.H., Cao, A.N., Chang, Z.J. and Lai, L.H. (2007) Nonnatural protein-protein interaction-pair design by key residues grafting. Proc. Natl. Acad. Sci. 104, 5330–5335.

[27] Boas, F.E. and Harbury, P.B. (2007) Potential energy functions for protein design. Curr. Opin. Struct. Biol. 17, 199–204.

[28] 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 67, 051927.

[29] Cordes, M.H., Walsh, N.P., McKnight, C.J. and Sauer, R.T. (1999) Evolution of a protein fold in vitro. Science 284, 325–328.

[30] Cordes, M.H., Burton, R.E., Walsh, N.P., McKnight, C.J. and Sauer, R.T. (2000) An evolutionary bridge to a new protein fold. Nat. Struct. Biol. 7, 1129–1132.

[31] Glykos, N.M., Cesareni, G. and Kokkinidis, M. (1999) Protein plasticity to the extreme: changing the topology of a 4-alpha-helical bundle with a single amino acid substitution. Structure 7, 597–603.

[32] James, L.C. and Tawfik, D.S. (2003) Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. Trends Biochem. Sci. 28, 361–368.

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