TOPICAL REVIEW —Statistical physics and complex systems

Proteins: From sequence to structure*

Zheng Wei-Mou(郑伟谋)†

Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, *China*

(Received 2 March 2014; published online 13 June 2014)

Protein sequences as special heterogeneous sequences are rare in the amino acid sequence space. The specific sequential order of amino acids of a protein is essential to its 3D structure. On the whole, the correlation between sequence and structure of a protein is not so strong. How well would a protein sequence contain its structural information? How does a sequence determine its native structure? Keeping the globular proteins in mind, we discuss several problems from sequence to structure.

Keywords: proteins, protein sequence, protein structures, protein folding

PACS: 87.14.E, 87.15.B, 87.15.Cc **DOI:** 10.1088/1674-1056/23/7/078705

1. Introduction

A protein is a linear chain formed by joining amino acids end to end. Proteins are essential parts of organisms and participate in virtually every process within cells. Usually, a protein performs its various function only when it folds into certain conformations. Thus, understanding the protein structure is important to comprehending the protein function. Protein sequences as special heterogeneous sequences are rare in the sequence space of amino acids. Such heterogeneity guides proteins to sample only a limited set of conformations in the folding process.

The correlation between protein sequence and structure is the starting point for protein structure prediction. One long-standing question is how well and in what way a protein sequence carries its structural information, and determine its structure. Here we examine the globular proteins to briefly review some important problems from sequence to structure.

2. Basics of proteins

Amino acid is the consisting unit of protein.[1] There are 20 different amino acids; each has a central C_α carbon atom, to which a side chain, an amino group and a carboxyl group are attached. What distinguishes them one from another is the side chain. (Only proline differs from this basic structure as it contains an unusual ring to the amine group.) The carboxyl group of the first amino acid condenses with the amino group of the next to eliminate water and forms a peptide bond. Proteins are chains of amino acids held together by peptide bonds. The starting end of the protein with a free amino group is known as the N-terminus or amino terminus, whereas the ending end with a free carboxyl group is known as the C-terminus or carboxyl terminus. Once linked in the protein chain, an individual amino acid is called a residue. Properties of residues are mainly characterized by the properties of their side chains: electricity, hydrophobicity, size, etc. Proteins have to fold into specific spatial conformations in order to be able to perform their biological function. Interactions among different side chains in water are the main source for protein to form the stable native structure in aqueous solution. The folding is mainly driven by the non-specific hydrophobic interactions to form a structural core, but the structure is stable only when the parts of a protein are locked into place by specific tertiary interactions. Hydrogen-bond can occur between the carbonyl oxygen atom of a residue and the amino nitrogen of another. The patterns of hydrogen bonds between backbone amino and carboxyl groups in proteins define the secondary structure. The most common secondary structures are alpha helices and beta sheets, which further consist of several extended strands. The first refined version of alpha helices and beta sheets as hydrogen bonding patterns in proteins was proposed by Pauling *et al.*[2,3] Tight turns and loose, flexible loops link the more 'regular' secondary structure elements of helices or strands. In three-dimensional (3D) space secondary structure elements of a protein are packed into a rather compact structure.

The two main types of secondary structure, the helix and strand, have a regular geometry. The backbone geometry of a protein is often described by its dihedral angles ϕ and ψ, which are dihedral angles with the bonds before and after C_α being the rotational axes, respectively. The geometry of helix or strand is characterized by specific values of ϕ and ψ on the Ramachandran plot, which is the space spanned by these two dihedral angles.[4] The main factor to determine the available values of ϕ and ψ is the stereochemical constraint of backbone atoms (and C_β atom). Glycine, which has a hydrogen

*Project supported by the National Natural Science Foundation of China (Grant Nos. 11175224 and 11121403).

†Corresponding author. E-mail: zheng@itp.ac.cn

© 2014 Chinese Physical Society and IOP Publishing Ltd

atom as a side chain, experiences the least constraint, compared with all other amino acids. Different amino acids have different propensities to various secondary structures. It is not simple at all to tell in what extent a local sequence, as a specific combination of nearby amino acids, determines its local conformations.

The secondary structure representation of protein is a useful tool for analyzing and comparing structures and functions of proteins. However, this description of the protein 3D geometry, especially geometry of loops, is inaccurate, and then lack of a direct connection to the 3D structure. There are many refined ways for discretized description of local 3D geometry of fragments, roughly in three categories: based on the phase space partition, based on representative points, and based on a distribution in the phase space. Among them the third one is most precise. The distance between successive C_α atoms is almost constant 3.8 Å (with an exception less than 1%). By connecting these atoms with 'pseudobonds', the backbone geometry of a protein is conveniently described by bending and torsion angles of the pseudobonds. Such a representation with angles is independent of the choice of reference coordinate frame, and hence intrinsic. The minimal unit of local geometry which has one-to-one correspondence with the coordinates is the three angles of tetrapeptide (two bending and one torsion angles). Clustering based on the distribution of these three angles provides an alphabet of 17 conformational letters.[5]

3. Propensities of amino acids to secondary structures

The most important database of protein structures is PDB.[6] A subset of representative structures with identity 25% is PDB_select_25. By taking more than one thousand structures of this subset and counting amino acids in different secondary structures, the statistics of amino acids is listed in Table 1, where amino acids are indicated with the IUPAC one-letter codes.

Table 1. Frequencies of amino acids (×20).

Amino acid	A	V	C	D	E	F	G	H	I	W
All	1.56	1.39	0.33	1.18	1.30	0.81	1.44	0.46	1.14	0.28
Coil[a)]	1.23	0.91	0.34	1.57	1.12	0.63	2.30	0.50	0.67	0.22
Sheet	1.19	2.66	0.43	0.61	0.91	1.17	0.96	0.48	1.99	0.37
Helix	2.20	1.19	0.26	1.05	1.76	0.82	0.66	0.42	1.17	0.31
Amino acid	K	L	M	N	Y	P	Q	R	S	T
All	1.25	1.77	0.42	0.92	0.71	0.90	0.79	0.97	1.25	1.13
Coil	1.25	1.23	0.34	1.27	0.56	1.55	0.70	0.88	1.50	1.22
Sheet	0.97	1.96	0.42	0.55	1.05	0.37	0.60	0.88	1.05	1.37
Helix	1.43	2.32	0.51	0.70	0.67	0.44	1.02	1.15	1.05	0.87

[a)] Here 'coil' stands for all irregular states other than helix and strand.

In the table, except the first and fifth rows, each row corresponds to a conditional probability. The ratio of an entry in a row other than the first or fifth to that in the first or fifth can be taken as a measure of the propensity of an amino acid to either coil, or helix, or sheet. A ratio greater than one means that the amino acid is favored by the secondary structure, otherwise is disfavored. For example, valine (V) is favored by sheet, but disfavored by helix. For coil, the propensities of amino acids in descending order are PGNDS-HTC|KRQE-MYAWF-LVI, while for sheet are VIYFWCT-LH|MR-SKAQ-EGNDP and for helix are AELQM-RKWIF|YHDVSCTN-PG (where symbol '|' separates favored ones from disfavored). Roughly speaking, coils are mostly located on surface, so prefer polar or charged small residues, the smallest amino acid glycine (G) or proline (P) whose conformational rigidity fits a turn; helices, whose crowdedly staggered side chains are prone to steric interference, dislike the amino acids with branched C_β or large aromatic ones while sheets have the opposite tendency.

4. Comparison of protein sequences

Sequence alignment is the most effective means for comparing various biological sequences, and often used for investigating evolutionary relationships between sequences, especially protein sequences. In an alignment, a specific site of a given sequence, with respect to another sequence, may correspond to an identical match, or a mismatch (explained as a point mutation), or a gap (insertion or deletion). It is possible to introduce a scheme to score the event of identical match, mismatch, insertion or deletion, or to evaluate the possibility of their occurrence. Once a scoring scheme is given, the total score may be taken as the objective function to determine the optimal sequence alignment. Once the optimal sequence alignment is given, a scoring scheme can also be derived from the alignment. However, we know neither the scoring scheme, nor the alignment at the beginning. In principle, we have to determine the both self-consistently. Almost every textbook of bioinformatics would mention the sequence alignment. An

excellent book is Ref. [7]. Here we briefly discuss the topic.

With a scoring scheme given, the alignment between sequences $A = a_1 a_2 \cdots a_m$ and $B = b_1 b_2 \cdots b_n$, whose lengths are respectively m and n, may be denoted by a sequence written in letters $\{M, D, I\}$ (meaning match, deletion, and insertion) whose length is at most $m+n$. For example,

$$\begin{array}{l} A = \texttt{prote--in} \\ B = \texttt{prot-amin} \end{array} \quad \rightarrow \quad \text{MMMMIDDMM}.$$

Here, when interpreting a gap as an insertion or deletion sequence, A has been taken as the 'template'. In contrast with the given sequences A and B, the sequence of alignment states is unseen and 'hidden'. The total score of such a hidden alignment sequence can be calculated once the scoring scheme is known. The optimal alignment or the hidden sequence with the highest score can be obtained efficiently by recursion relations or the so-called dynamic programming approach.

Multiple sequence alignment is an extension of above pairwise alignment. There are also many algorithms. Here we mention only an efficient approach called the 'profile hidden Markov model'. The simplest multiple sequence alignment model for signals is the model of gapless aligned blocks. The mathematical description of such a block is the so-called position specific weight matrix (PSWM) where an aligned position is a column whose elements represent frequencies for each symbols to occur at that position. For proteins the matrix has 20 rows for 20 amino acids. This model can be extended to aligning multiple sequences with gaps allowed. Now each sequence is aligned to a given common template. If a letter of the sequence is aligned to a letter of the template, the state of alignment is M (match); if a gap is added to align to the template, the state is D (deletion); if a letter of the sequence is aligned to a gap in the template, the state is I (insertion). States M, D, and I are hidden variables. The profile hidden Markov model assumes that the sequence of hidden states maps to a Markov chain, which is described by an initial probability distribution of states, $\{\pi_\alpha\}, \alpha \in \{M, I, D\}$, and a set of transition rates between states, $T_{\alpha\beta}$. At each state $\sigma_i = M$, a letter $s_i = a$ is 'emitted' with probability $P(a|M)$. This is a 'generating model' according to which the aligned multiple sequences are 'generated'. Thus, it can be evaluated how well a sequence matches the template by determining the alignment with the highest probability. When interpreting probabilities as scores, a gap corresponds to a negative penalty. A high total score of alignment implies that the alignment very unlikely comes by chance.

Multiple sequence alignment is often used to explore conserved sequence sites and point mutations, to extract motifs, and to aid in establishing evolutionary relationships by constructing phylogenetic trees. It is a useful tool for analyzing protein structure and function.

5. Protein secondary structure prediction

Methods for protein secondary structure prediction are mainly based on the propensity of amino acids to various secondary structure states. Among helix, sheet, or coil, which state a specific residue will take also depends on the context of the amino acid sequence around the given residue. This context dependency can be verified by examining the mutual information between the secondary structure state of the given residue and the amino acid identity of residues at different up- or downstream positions. It is found that the secondary structure state of a center residue is affected by residues at up to eight positions on both sides of the center. This dependency can be described with the PSWM mentioned above. The advantage of this weight matrix model is the small number of its parameters while its disadvantage is its neglect of correlations. A way to keep the correlation between amino acid pairs is to consider 20 weight matrices according to the amino acid identity of the central residue.[8] An alternative way to keep correlation at a small number of parameters is to reduce the size of amino acid alphabet, say from 20 to 5 by clustering similar amino acids. By selecting representative sequences with know secondary states to construct a training set, parameters for a model can be obtained, and the secondary structure of any sequence can then be predicted based only on the given single sequences.[9] The prediction accuracy based on a single sequence is about 65%. This also roughly measures the extent at which the local structure depends on local sequence.

The methods for protein secondary structure prediction based on a single sequence are inevitably affected by large background noises. By extracting only those sequences which are closely related to the sequence whose secondary structure is to be predicted, the prediction accuracy can be raised greatly to above 80%. The multiple sequence alignment contains the evolutionary information of sequences, and reflects the sequential conservation of secondary structure, so helps to improve the prediction. However, it is required to search sequence database in the prediction stage, so the computational cost is generally higher. There are many softwares, among which the most popular one is PSIPRED.[10,11]

Most methods for protein secondary structure prediction first ignore the correlation between secondary structure states of nearby residues, and predict the state of each residue separately. Post-prediction filtering rules then have to be applied to meet the correlation requirements (such as at least three contiguous 'H's and two 'E's) by adjusting artificially. However, a reasonable way is to include such fundamental correlations in the model at the very beginning when constructing a model. For example, instead of three states (H, E, and C) for a single residue, we may consider states of pentapeptides like CCCHH. The total number of pentapeptide states is far below 3^5 since strings like CCHHC are illegal with only two contiguous 'H'.

The total number of legal states reduces to only 75. Furthermore, by considering the hidden secondary states as a Markov chain, a hidden Markov model can be introduced.[12]

6. Protein 3D structure prediction

Most proteins fold into unique compact three-dimensional structures under the physiological condition. The shape into which a protein naturally folds is known as its native conformation, which is often dependent on the specific order of amino acids in the sequence. The hydrophobic interactions cause the burial of hydrophobic residues from water. The native structure is stabilized by specific tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds when the parts of a protein are locked into place. In fact, the hydrophobic interaction as an effective interaction is a difference of free energy including water environment.[13] Protein structure prediction is the prediction of the three-dimensional structure of a protein, its backbone geometry or even spacial coordinates of side-chain atoms, from its amino acid sequence. Methods of protein 3D structure prediction are roughly in three categories: homology modeling, protein threading, and *ab initio* modelling or *de novo* prediction.

6.1. Homology modeling

'Homology' here need not be in a strict sense of that in the biological molecular evolution. It is often true that if two proteins are very similar in their sequences, then they also similar in their 3D structures. Generally speaking, in the biological molecular evolution, structures are more conservative than sequences, being under more constraints. Thus, for a given query sequence, if a 'template', whose structure is known and whose sequence is similar to the query sequence, can be found by sequence alignment the structure of the template can be taken as the prediction of structure for the query sequence. The key issue in homology modeling is sequence alignment. If profiles of multiple sequence alignment are obtained first for both the query and template sequences, and then their profiles are aligned, the quality of alignment can be improved.[14,15] Now it is a consensus that no reliable sequence alignment is available if their sequence identity is below 30%; homology modeling then fails.

6.2. Protein threading

Only sequence alignment is considered in the homology modeling, so its limitation is obvious. Taking the template sequence found by sequence alignment as a 3D model, the protein threading further examines how well the query sequence fits the model structure, or conducts a sequence-structure alignment.[16] The protein threading utilizes the structural information of template databases: interactions between specific residue pairs, burial states of certain residues, secondary structure states of residues, etc. The prediction accuracy is then improved. The conditions for the success of protein threading are: the selected templates are correct, the alignment between the query and template sequences is reliable, and the features to judge sequence-structure matching are effective.

6.3. *Ab initio* modelling

The above two categories of methods strongly depend on the selection of a proper model from structure databases. If any relevant template does not exist or is hardly found, such comparison based methods will fail. In this situation, methods without referring any templates are demanded. *Ab initio* modeling designs a potential function as the objective function, and regard the native conformation as the state with the lowest potential energy. Many different 3D conformations are generated, and their total potentials are calculated. The one with the lowest potential is then taken for prediction. The starting point of such methods does not depend on any templates. However, the number of possible 3D conformations is huge, and the design of a proper potential function is by no means simple. A coarse-grained approach to reduce the search space is to divide the query protein into fragments, sample possible fragmental conformations according to structure databases of fragments, and then assemble fragments into whole conformations. There are many softwares in the category of *ab initio* modeling, for example, FRAGFOLD,[17] ROSETTA,[18] TASSER,[19] FB5-HMM[20] and FALCON.[21]

As for design of model potential functions, usually some empirical statistical potentials are constructed from a training set of representative native structures. Since we have only limited knowledge about the folding process, and we have few non-native conformations there are unsolved basic problems related to potential functions. No potential functions so far can handle the problem about the topology of 3D packing of secondary structure elements.

7. Folding of peptide fragments

If the conformations of peptide fragments agree with those they take in native states, the problem of the protein native structure prediction can be solved by the 'divide and conquer' strategy. The 65% accuracy of protein secondary structure prediction based on single sequence implies that local sequences contain information of local conformations, and the latter are also affected by 3D structure as a whole. Is the protein folding hierarchical? Does folding start with certain local conformations? Study of peptide folding is the first step, which is not only practical, but also meaningful.

7.1. Folding of peptides: wet experiments

Progresses in experimental techniques, including the solid-phase synthesis techniques that earned Bruce Merrifield the 1984 Nobel Prize in Chemistry, worth writing home about. Here, only the results most relevant to our discussion are briefly mentioned. For some time, general consensus was that the conformational ensemble of peptides would show no preferences for any particular structure since the hydrogen bonding potential of the bulk water would trump any feeble attempts at intramolecular hydrogen bonding in the peptide. However, the NMR experimental data of peptide YPYDVPDYA show a conformational preference of tetrapeptide YPYD for a reverse turn conformation.[22] The sequential dependency of turn conformation is quite clear. For example, tetrapeptide VPDY on the same fragment peptide has no preference for a turn. Later detailed point mutation experiments indicated that tetrapeptide APGD shows the strongest propensity to a turn, and the tendency still exists even after the second proline is replaced.[23]

A natural picture about the starting of protein folding process is an early formation of a few secondary structures such as a helical piece or a turn.[24] Although there is an opinion about early hydrophobic collapse it is hard to explain how the collapse leads to a specific 3D structure.[25] Myohemerythrin is a typical all-helix protein, consisting of several helices. As an isolated peptide, the peptide of its third helix does not form helical structure in solution. However, the populated states of the peptide show the formation of multiple, interlinked reverse turns in part of the peptide as 'nascent helix'. After adding the helix-promoting solvent trifluoroethanol, helical structure is formed.[26] Experiments on point mutations of this protein indicated that a point mutation at one helix would affect the stability of another helix which is far away in sequence but close in space. Plastocyanin represents an all-beta structure. The peptides representing β-strands appears to have no propensity for helix or turn formation, but remains extended for the vast majority of the population of states.[27,28]

7.2. Folding of peptides: computer simulations

Besides the above wet experiments, computer simulations are also conducted to investigate in what extent local conformations of peptides determine the native state. Peptide simulation requires relatively low computational cost, and has been conducted in several labs.[29] We shall not discuss simulation techniques and algorithms, and mention only some results most relevant to our discussion here. A typical example is the simulation of Ho and Dill.[30] They selected six proteins from various types of structures, and extracted a peptide of eight residues every three residues from every protein. Totally 133 8-mer fragments were obtained for simulation. Dihedral angles (ϕ, ψ) were divided into three regions: helix [a], sheet [b] and the other [l]. A conformation of 8-mer peptide corresponds to a 'mesostring' written in these three letters while a mesostring corresponds to a set of conformations (due to the coarse-grained nature of mesostring). Thus, a free energy can be assigned to a mesostring. The mesostring with the lowest free energy is defined as the 'ground mesostring'. If the ground mesostring of a peptide has a high enough population and the free energy gap or difference between the ground mesostring and others is large enough in a simulation, the peptide is identified as 'structured'. (Reasonable values for the two parameters could be: population $> 45\%$ and gap > 0.6 kCal/Mol.) Among the above 133 8-mer peptides 48 peptides were found as structured and for 41 of these structured their conformations of ground mesostrings agree with their native conformations with small root mean square deviations (RMSD).

7.3. Folding of peptides: bioinformations

The known structures of proteins are about 90 thousands, and the known sequences of proteins are much more. Some fragmental conformations frequently appear in structures of proteins. Helices, extended strands and turns may be regarded as the minimal units of frequently appearing structural motifs. Two secondary structure elements which are sequentially close can form the so-called 'supersecondary structures' as structural motifs of a higher level. Protein sequences also have frequently appearing local sequence motifs. Generally, structures are more conservative than sequences in evolution. There might be multiple ways for the realization of a given structural motif. For example, two helical segments often have very similar local geometry, but their local sequences could differ significantly. For motifs in the joint structure-sequence space the correlation between their sequences and structures should be strong by definition. Peptides which correspond to such motifs should have a high chance to be structured in simulation, and their ground conformations also have a high chance to agree with their native conformations.

Baker's lab has constructed a database of joint structure-sequence motifs called I-sites.[31] Each motif of I-sites is described by a weight matrix whose 20 rows correspond to 20 amino acids, and whose columns correspond to aligned positions or sites. The matrix of a motif assigns a score to a peptide to evaluate the extent for the peptide to belong to the motif; if the score is high enough the peptide may be annotated as a member of the motif. For a given protein sequence, one may identify the highly scored peptides using matrices of I-sites motifs. Simulations of those highly scored peptides showed that these peptides were mostly structured, and their conformations in simulations agreed with their native conformations.[32]

As mentioned above, the conformational alphabet of 17

letters for local geometry of protein backbone is much finer than the secondary structure description especially for loops. A proper scoring scheme may be designed to evaluate the propensity of local amino acid sequences to certain conformational letters. Typical motifs of local structures usually correspond to definite strings of conformational letters. The peptides whose scores for certain conformational strings are high will have a high chance for their conformations in simulations to agree with their native conformations.

On the whole, the correlation between protein sequence and structure seems not so strong. However, a few sites of a sequence as key sites may carry strong signals of local geometry. The problem is then how to discover such sequence signals.

8. Spatial arrangement of secondary structure elements

The interactions between secondary structure elements which are far in sequence but close in space are essential to the stability of protein 3D structures. Such interactions may affect local structures of peptide fragments, and make a compromise between local and global advantages to reach a global optimal stability. The potential functions used so far, either those atomic potential functions derived from quantum chemistry or empirical statistical ones derived from databases, cannot describe the interactions between secondary structure elements, and originally are not designed for this purpose. In fact, we have no effective means to handle the spatial arrangement of secondary structure elements, or the rough topology of a protein. It is an important task to establish an interacting model at this scale of secondary structure elements.

The case of a super-secondary structure connecting by a turn, especially a short turn, is relatively simple. The turn then often carries a strong sequence signal. For interacting secondary structure elements which do not belong to super-secondary structures, the available means for analysis are mainly the multiple sequence alignment obtained from the profile hidden Markov model. The protein folding core consists of regular secondary structure elements. Since improper mutations in these core elements might make the structure to become unstable they tends to be conservative in sequence under this stability constraint. Thus, it is expected to see such conservative patterns in the profile of multiple sequence alignment. Among the databases of multiple sequence alignments of proteins, a famous one is Pfam.[33,34] It is important to infer interacting secondary structure elements based on multiple sequence alignments and then to determine the rough topology of a protein. There have been discussions on correlation decoupling analysis for discovering residue contacts.[35,36] These discussions are tempting in their mathematical formalism, but are also problematic. The analysis requires a large size (number of sequences) of the multiple sequence alignments, and the relevant sites should be conservative but at the same time show also variety, so it is not so ideal. A recent review[37] which cover many topics in correlation analysis is worth reading.

9. An ending remark: protein folding — "kidnap by elites on a way with no return"

Amino acid sequences of proteins are extremely rare among all possible amino acid sequences. The specific order of amino acids in a protein sequence is essential for its 3D structure. As a whole, the correlation between protein sequence and structure is not so strong, but a few individual helical rings and turns may have strong sequence signals for local geometry. At the starting stage of folding these very short fragments may be locked at their local conformations governed by sequence signals. The remaining residues of the sequence would show some tendency for accepting their own native conformations, but the sequence signals are only marginal. Only in this way can easily be adjusted the local structures to finally fit the native state. Such rings and turns with strong sequence signals may be regarded as 'elites' at the minimal scale which 'kidnap' the whole protein. Similarly, a few pairs of secondary structure elements may carry strong sequence signals to indicate that they form contacts by some specific strong interactions to lock partial local topology. They are 'elites' at a larger scale. The process of protein folding is far from ergodic. Once some 'elites' take their native conformations, any configurations which conflict with these elite conformations are excluded, and then are never visited. By forming their conformations, elites at successive scales progressively reduce the space of configurations to be visited and drive the protein finally to its native state. The traditional equilibrium statistical mechanics and the concepts like the free energy landscape are not so useful, and are often misleading. Understanding proteins from sequence to structure is an inviting challenge.

References

[1] Branden C and Tooze J 1991 *Introduction to Protein Structure* (New York: Garland)

[2] Pauling L, Corey R B and Branson H R 1951 *Proc. Natl. Acad. Sci. Wash.* **37** 205

[3] Pauling L and Corey R B 1951 *Proc. Natl. Acad. Sci. Wash.* **37** 729

[4] Ramachandran G N, Ramakrishnan C and Sasisekharan V 1963 *J. Mol. Biol.* **7** 95

[5] Zheng W M and Liu X A 2005 *Protein Structural Alphabet and Its Substitution Matrix CLESUM*, in Lecture Notes in Bioinformatics 3680, edited by Priami C and Zelikovsky A (Berlin: Springer) pp. 59–67

[6] Berman H M and Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucl. Acids Res.* **28** 235

[7] Durbin R, Eddy S R, Krogh A and Mitchison G 1998 *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge: Cambridge University Press)

[8] Garnier J, Osguthorpe D J and Robson B 1978 *J. Mol. Biol.* **120** 97

[9] Zheng W M 2004 *J. Bioinfor. Comp. Biol.* **2** 333

[10] Buchan D W A, Minneci F, Nugent T C O, Bryson K and Jones D T 2013 *Nucl. Acids Res.* **41** W340
[11] McGuffin L J, Bryson K and Jones D T 2000 *Bioinformatics* **16** 404
[12] Zheng W M 2005 *Int. J. Bioinfor. Res. Appl.* **1** 420
[13] Finkelstein A V and Ptitsyn O 2002 *Protein Physics: A Course of Lectures* (Academic)
[14] Söding J 2005 *Bioinformatics* **21** 951
[15] Söding J, Biegert A and Lupas A N 2005 *Nucl. Acids Res.* **33** W244
[16] Bowie J U, Luthy R and Eisenberg D 1991 *Science* **253** 164
[17] Jones D T 2001 *Proteins* **45** 127
[18] Das R and Baker D 2008 *Ann. Rev. Biochem.* **77** 363
[19] Zhang Y, Arakaki A K and Skolnick J 2005 *Proteins* **61** 91
[20] Hamelryck T, Kent J T and Krogh A 2006 *PLoS Comput. Biol.* **2** e131
[21] Li S C, Bu D, Xu J and Li M 2009 *Protein Sci.* **17** 1925
[22] Dyson H J, Cross K J, Houghten R A, Wilson I A Wright P E and Lerner R A 1985 *Nature* **318** 480
[23] Dyson H J, Rance M, Houghten R A, Wright P E and Lerner R A 1988 *J. Mol. Biol.* **201** 201
[24] Karplus M and Weaver D L 1979 *Biopolymers* **18** 1421
[25] Chan H S and Dill K A 1990 *Proc. Natl. Acad. Sci. USA* **87** 6388
[26] Dyson H J, Merutka G, Waltho J P, Lerner R A and Wright P E 1992 *J. Mol. Biol.* **226** 795
[27] Dyson H J, Merutka G, Waltho J P, Lerner R A and Wright P E 1992 *J. Mol. Biol.* **226** 819
[28] Rotondi K S and Gierasch L M 2003 *Biochemistry* **42** 7976
[29] Avbelj F and Moult J 1995 *Proteins* **23** 129
[30] Bosco K H and Ken A Dill 2006 *PLoS Comput. Biol.* **2** e27
[31] Bystroff C and Baker D 1998 *J. Mol. Biol.* **281** 565
[32] Bystroff C and Garde S 2003 *Proteins* **50** 552
[33] Bateman A, Coin L, Durbin R, Finn R D, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E L L, Studholme D J, Yeats C and Eddy S R 2004 *Nucl. Acids Res.* **32** D138
[34] Finn R D, Tate J, Mistry J, Coggill P C, Sammut S J, Hotz H, Ceric G, Forslund K, Eddy S R, Sonnhammer E L L and Bateman A 2008 *Nucl. Acids Res.* **36** D281
[35] Morcos F, Pagnanib A, Lunta B, Bertolinoc A, Marksd D S, Sandere C, Zecchinab R, Onuchica J N, Hwaa T and Weigtb M 2011 *Proc. Natl. Acad. Sci. USA* **108** E1293
[36] Jones D T, Buchan D W A, Cozzetto D and Pontil M 2012 *Bioinformatics* **28** 184
[37] de Juan D, Pazos F and Valencia A 2013 *Nat. Rev. Genet.* **14** 249