A Contact Energy Function Considering Residue Hydrophobic Environment and Its Application in Protein Fold Recognition

Mo-Jie Duan and Yan-Hong Zhou*

Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074, China.

The three-dimensional (3D) structure prediction of proteins is an important task in bioinformatics. Finding energy functions that can better represent residue-residue and residue-solvent interactions is a crucial way to improve the prediction accuracy. The widely used contact energy functions mostly only consider the contact frequency between different types of residues; however, we find that the contact frequency also relates to the residue hydrophobic environment. Accordingly, we present an improved contact energy function to integrate the two factors, which can reflect the influence of hydrophobic interaction on the stabilization of protein 3D structure more effectively. Furthermore, a fold recognition (threading) approach based on this energy function is developed. The testing results obtained with 20 randomly selected proteins demonstrate that, compared with common contact energy functions, the proposed energy function can improve the accuracy of the fold template prediction from 20% to 50%, and can also improve the accuracy of the sequence-template alignment from 35% to 65%.

Key words: protein structure prediction, fold recognition, contact energy, hydrophobic environment

Introduction

The knowledge of protein structures plays a very important role in understanding protein functions, studying protein-protein interactions (*1*), reconstructing protein structures (*2*), and performing rational drug design (*3*). Protein structures can be determined by both experimental and computational methods. Experimental methods such as x-ray crystallography and nuclear magnetic resonance can determine the three-dimensional (3D) structure of proteins precisely; however, currently these methods are still inefficient and can only be applied to a small part of proteins (*1*). On the other hand, computational methods can, in principle, not only overcome the shortages of experimental methods, but also assist in understanding the mechanism of protein folding (*4*). As a result, computational methods have been studied extensively and become an effective way to analyze protein structures (*5*).

Methods for predicting the protein 3D structure can be divided into three main categories: homology modeling (*6*, *7*), fold recognition (*8*, *9*), and *ab initio* prediction (*10*). The homology modeling methods first search for homological proteins of the target protein in a structure-known protein database, and then use the structures of the homological proteins as templates to build a structure model for the target protein. The fold recognition methods try to find a fold template for the target protein from a template library, and then construct a full structure model for the target protein based on the selected fold template. The *ab initio* prediction methods, which predict the structure of the target protein only based on its sequence information, calculate the energy for all possible conformations that the target sequence may fold into, and select the conformation with the lowest energy as the native conformation of the target protein.

In addition, according to the information used, the protein structure prediction methods can also be classified into two classes. The first one uses the information of known protein structures and evolution, which searches for homologous proteins of the target protein first, and then builds the structural model for the target protein based on the structural information of the homological proteins. This class includes the above mentioned homology modeling and fold recognition

*** Corresponding author.**
E-mail: yhzhou@hust.edu.cn

methods based on PSI-BLAST (*7*) or hidden Markov model (*11*, *12*). The second class makes use of the information about residue-residue and residue-solvent interactions, which creates energy functions by statistical or theoretical analysis first, and then uses the functions to search for the optimal structure from a structure template library or all possible conformations of the target protein. The threading (*13*) and *ab initio* prediction methods (*10*) belong to this class. Obviously, for the second class, it is crucial that the energy functions should be able to describe residue-residue or residue-solvent interactions efficiently (*14*).

In threading methods, the widely used energy functions are obtained from statistical analysis (*15*), and most of them are based on the contact energy between residues (*16*, *17*). These energy functions define pairwise residue contact energy scores according to the residue-residue contact frequency occurred in known protein structures. The basic idea of these energy functions was first proposed by Tanaka and Scheraga (*18*), and various improvements have been made by subsequent researchers, such as combing the hydrophobic property of residues (*19*, *20*), considering the orientation anisotropy of side chains (*21*), applying atom-level functions (*22*), and using more complicated models (multi-body models) (*23*). In addition, the performance of applying these energy functions to protein fold recognition has also been evaluated by previous researches (*24–26*). The results indicate that the energy functions merely based on the residue contact frequency are inexact, and one of the reasons might be that the influence of the hydrophobic interaction on contact energy has not been considered. In the process that protein sequences fold into advanced structures, the hydrophobic interaction is believed to be the dominant driving force (*27*), which makes hydrophobic residues come into the core and makes hydrophilic residues tend to exist on the surface. This phenomenon indicates that the contact preference between residues relates to solvent molecules to some extent, and therefore the influence of the solution environment (hydrophobic environment) should be considered while analyzing the pairwise residue contact energy. However, some of the existing pairwise energy functions take no account of the influence of the hydrophobic interaction on the structure stability at all (*16*), others only consider the solution influence in terms of the hydrophobic property of residues (*19*). For threading-based protein fold recognition, these functions are not able to reflect the influence of the hydrophobic environment on the residue contact energy effectively.

In this study, the preference of residues to the hydrophobic environment is analyzed by a statistical method, and an improved contact energy function that considers both the residue contact frequency and the residue hydrophobic environment is proposed, which can reflect the influence of hydrophobic interaction on protein structure stability more effectively. Furthermore, a fold recognition (threading) approach based on this energy function is developed, and the testing results demonstrate that, compared with common contact energy functions, the proposed one can improve the accuracy of protein fold recognition more effectively.

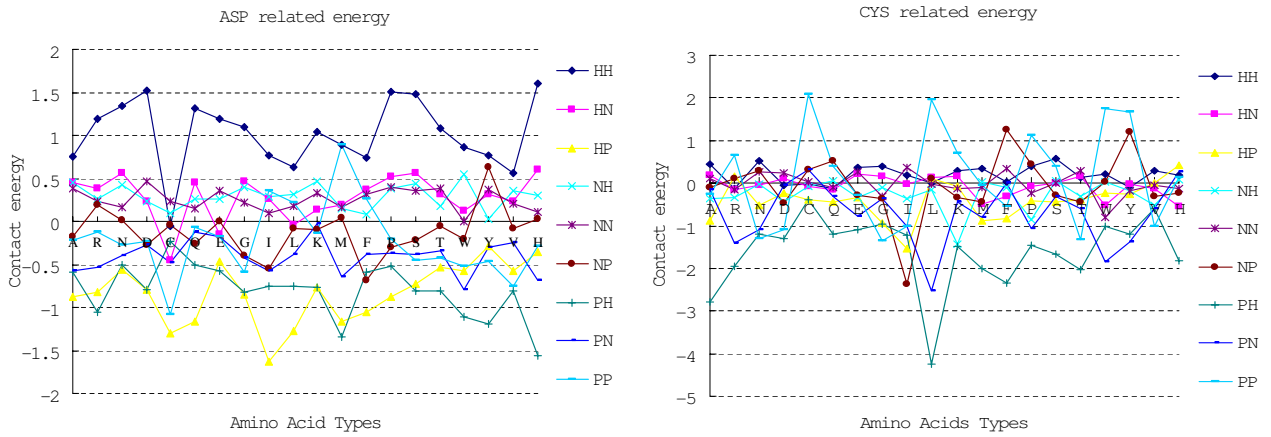
Results

The contact energy considering residue hydrophobic environment

First, a dataset consisting of the structural information of 525 proteins was collected from the Protein Data Bank (PDB) database (*28*) for analyzing the residue contact energy and developing the energy function (see Materials and Methods). According to this dataset, the coordinates of C_β atoms of all residues (C_α atoms for glycine) were obtained and used to evaluate the distance between residues. This distance was further used to determine whether two residues were contacted with each other. Then, the solvent accessible surface areas (SASAs) of residues (*29*) were obtained by the POPS program (*30*) to determine the hydrophobic environment for each of the residues. Based on SASA, the hydrophobic environment was classified into three types, hydrophobic, hydrophilic, and neutral (uncertain). Finally, the information of residue distance and hydrophobic environment was integrated to determine the residue contact energy (see Materials and Methods).

The contact energy was defined for each of the 400 possible residue pairs between 20 types of amino acids. Furthermore, for each residue pair, there were nine possible combinations in terms of residue hydrophobic environment. Totally, 3,600 items of residue contact energy were determined. The contact energy related to asparagines and cysteines is shown in Figure 1.

The statistical result indicates that the residue contact energy is distinct for different residue pairs or different combinations of residue hydrophobic environment. The analysis of this result reveals that:



A **B**

Fig. 1 Examples of the relationship between the residue contact energy and the residue hydrophobic environment. The x-axis represents twenty types of amino acids, and the y-axis represents the residue contact energy; different curves represent different hydrophobic environment. The letters H, N, and P denote hydrophobic, neutral (uncertain), and polar (hydrophilic), respectively. **A**. Residue contact energy related to asparagine (ASP). It can be seen that the contact energy of residue pairs containing asparagine tends to be high when the residue pairs are in the HH state (that is, both residues are in the hydrophobic position; the only exception is the pair with cysteine), but tends to be low in the HP or PH state. **B**. Residue contact energy related to cysteine (CYS). For some residue pairs consisting of cysteine and another residue such as cysteine, leucine, tryosin, or tyrosine, the contact energy is high when the residue pairs are in the PP state, which is distinct from that related to asparagine.

(1) For most residue pairs, their contact energy scores tend to be large when both residues are in the hydrophobic position, but small when one is in the hydrophobic position and the other in the hydrophilic position. (2) The contact energy scores for different residue pairs are quite different from each other. (3) The contact energy scores of some residue pairs are special. For example, the distributions of contact frequency between cysteine and other residues tend to be random, but the contact probability is high when two cysteine residues appear in the hydrophilic environment simultaneously (Figure 1B).

The prediction accuracy of applying the contact energy to threading

In this study, the proposed contact energy was applied to protein fold recognition using the threading method, and a dataset consisting of 20 randomly selected proteins was used to test the prediction accuracy. The PDB identifiers of these proteins are shown in Table 1. For the purpose of comparison, the prediction accuracy of commonly used contact energy was also tested. Three measures, including self-template prediction accuracy, sequence-template alignment accuracy, and native alignment score, were used to evaluate the prediction performance.

Self-template prediction accuracy

To evaluate the accuracy of self-template prediction, z-scores (*31*) of the alignments between the target protein sequence and each template in a template library were calculated and used to rank the templates, then the position of the target template in the ranked templates can reflect the accuracy of self-template prediction. The testing results indicate that, compared with the common energy function, the improved energy function can perform better for 14 out of the 20 test proteins (70%). The percentage of testing proteins whose z-scores are ranked within the top 10%, 25%, and 50% of all library templates are given in Table 2.

Sequence-structure alignment accuracy

The accuracy of the alignment between target sequences and their own structures, which judges whether the optimal alignments are consistent with actual situations, was also used to evaluate the fold recognition effect. As there were certain alignment errors, shifts from the exact alignment within four residues were counted as correct alignments (*16*). In this test, 7 out of 20 proteins (35%) were aligned correctly using the common energy function, while the percentage was 65% for the improved energy function.

Table 1 Features of the Twenty Testing Proteins

PDB ID	Secondary structure		No. of	Group
	No. of α-helix	No. of β-sheet	S–S bonds	
1ahu	3	0	0	Better
1cuk	3	0	0	Better
1hry	3	0	0	Better
1ahl	0	3	3	Better
1aw6	2	0	0	Better
1fre	1	2	0	Better
1r2a	2	0	0	Better
1lea	3	0	0	Better
1cbn	2	2	2	Better
1co4	2	2	0	Better
1auu	0	3	0	Better
1sei	1	3	0	Better
1nkl	5	0	3	Better
1neq	5	0	0	Better
1tpm	0	3	1	Equal
1ehs	2	0	2	Worse
1fd4	1	3	3	Worse
1mkn	0	3	3	Worse
1hyk	0	2	2	Worse
1chc	1	3	0	Worse

Table 2 Percentages of the Testing Proteins with Different z-score Ranks

Energy function	Top 10%	Top 25%	Top 50%
Common energy function	20%	55%	85%
Improved energy function	50%	80%	95%

Native alignment score

The effect of energy functions can also be evaluated by the difference of energy scores between the sequence-template native alignments and random alignments. The energy scores of the sequence-template native alignments were obtained by aligning the residues of target sequences to their own positions in the templates, and the energy scores of random alignments were calculated by randomly aligning the residues of target sequences to the templates. For each target sequence, 1,000 random alignments were made, and the average score of these alignments was used for this target sequence. In this test, when the improved energy function was used, the native alignment scores of 75% of proteins were higher than their average scores of random alignments, and this figure was only 50% for the common energy function.

Discussion

The above testing results demonstrate that the contact energy function combining with the hydrophobic environment is superior to the common energy function, indicating that the hydrophobic environment not only relates to the residue contact energy, but also influences the accuracy of fold recognition.

In order to analyze whether the prediction accuracy was correlated with protein structure features, we divided the 20 testing proteins into three groups, namely "Better", "Equal", and "Worse", according to the template prediction accuracy. The "Better" group means that the template prediction accuracy using the improved energy function was better than using the common energy function, the "Worse" group refers to the contrary situations, and the "Equal" group represents that the prediction accuracy was

equal. Then, the secondary structures and the number of disulfide bonds in these proteins were analyzed. The results are given in Table 1.

As can be seen from Table 1, most of the proteins in the "Better" group belong to the α class or α/β class, while most of the proteins in the "Worse" group belong to the β class. In addition, the prediction accuracy also relates to the number of disulfide bonds in proteins. The reasons of the above phenomenon might be: when α-helix forms its compact conformation, it will be driven by outside forces like the repulsion and attraction of solvent molecules, consequently, the influence of hydrophobic interaction is more significant for proteins consisting of α-helix; on the contrary, as β-sheet is not so tight as α-helix, it may be stabilized by residue interaction, therefore it is unnecessary to consider the hydrophobic interaction.

However, the optimal alignment between target sequences and their own templates may not be exact, which might be caused by the following reasons: (1) The contact energy is derived from statistical analysis, which only reflects the statistically possible interactions between residues. (2) The optimal alignment is obtained by global alignment, which cannot guarantee that all local alignments are optimal. (3) For a specific residue, the residues around it may have similar characters with it.

Materials and Methods

Dataset

A total of 525 proteins were selected from the PDB database for analyzing the contact energy function, which meet the following criteria: (1) None of the identity between each other is less than 30%. (2) The structure is determined by x-ray crystallography. (3) The resolution is better than 2.0 Å. (4) The sequence consists of 30–750 amino acids.

Residue hydrophobic environment

The residue hydrophobic environment can be determined by the SASA of residue (*29*), which is defined as the center area traced out by solvent molecules as they roll over the exposure surface of residues in the solvent (*32*). A small value of SASA means that the residue tends to be in the hydrophobic environment, otherwise, it tends to be in the hydrophilic environment. There are many programs for calculating SASA (*30*, *33*), most of them are based on the atom coordinates submitted by users. In this research, a freely available program POPS (*30*) was used. Based on the SASA value, we classified the environment of the residues into three types: hydrophobic, hydrophilic, and neutral (uncertain).

Contact energy function

Common energy function

The common contact energy function is based on the contact preference of residues. The contact energy between residues a_i and a_j is defined as

$$e_c(a_i, a_j) = \log \frac{p\,(a_i, a_j; r_c)}{p^{\,0}(a_i, a_j; r_c)} \qquad (1)$$

where $p\,(a_i, a_j; r_c)$ is the probability that the distance between a_i and a_j is less than the designated cutoff value r_c, and $p^{\,0}(a_i, a_j; r_c)$ is the expected probability correspondingly.

During fold recognition, the residues of the target protein are first placed onto templates by some alignment methods, then the score of the alignment can be determined by

$$E_{Align} = -\sum_{i<j} e_c(a_i, a_j)\,\sigma(r_{ij} - r_c) \qquad (2)$$

where (i, j) is a site pair of the template; a_i and a_j are residue types on sites i and j; $\sigma(x)$ is equal to 0 when $x > 0$ and is equal to 1 when $x \leq 0$; and r_{ij} is the distance between sites i and j. The energy functions based on the above idea are still broadly used in fold recognition methods.

Improved energy function

In this study, an improved energy function is proposed, which concerns not only the residue type, but also the residue hydrophobic environment. This energy function is given by

$$e_c^i(a_i, a_j; env_i, env_j) = \log \frac{p_c(a_i, a_j; env_i, env_j)}{p^{\,0}(a_i, a_j; env_i, env_j)} \qquad (3)$$

where $p_c(a_i, a_j; env_i, env_j)$ is the probability that a_i and a_j contact with each other in the hydrophobic environment env_i and env_j, respectively, and $p^{\,0}(a_i, a_j; env_i, env_j)$ is the expected probability. The distance between residues is measured by the distance between C_β atoms (C_α atoms for glycine), and the cutoff r_c is set as 7.5 Å.

Similar to Equation (2), the energy score of the sequence-template alignment is defined as

$$E_{Align} = -\sum_{i<j} e_c(a_i, a_j; T_{env_i}, T_{env_j})\,\sigma(r_{ij} - r_c) \quad (4)$$

where (T_{env_i}, T_{env_j}) denotes the hydrophobic environment of sites i and j in the template.

Protein fold recognition based on threading

Threading is an efficient method for protein fold recognition, which can be used to evaluate the structural similarity among proteins with low sequence identity. The threading process realized in this study is shown in Figure 2, which consists of the following steps:

1. Build the fold template library. Based on the classification of the CATH database (*34*), 438 fold families and their representative structures were selected, resulting in a fold template library containing 438 structures.

2. Determine residue contact energy. The residue contact energy scores were utilized to calculate the alignment score between target sequences and structure templates, which was determined by Equations (2) and (4).

3. Align target sequences to structure templates. Because the sequence lengths of the target protein and template proteins are usually different and gaps are allowed in the sequence-structure alignment, it would be an NP-problem to search the optimal alignment. To solve this problem, the divide-and-conquer algorithm (*16*), an approximate global optimal searching algorithm, was adopted in this study.

4. Determine the optimal template. The fittest template of target sequences was determined by *z*-score (*31*).

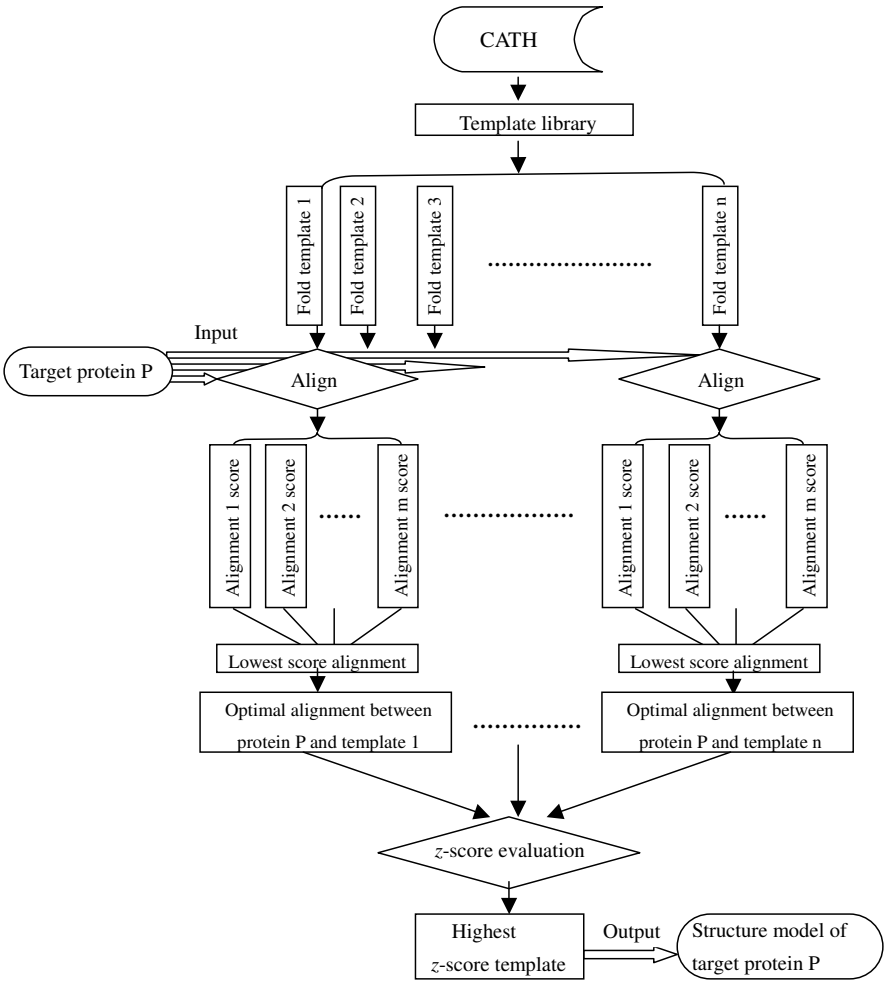


Fig. 2 The flow chart of the threading process in this study.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 90203011 and 30370354) and the Ministry of Education of China (No. 505010 and CG2003-GA002).

References

1. Vitkup, D., *et al.* 2001. Completeness in structural genomics. *Nat. Struct. Biol.* 8: 559-566.
2. Russell, R.B., *et al.* 2004. A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.* 14: 313-324.
3. Jacobson, M. and Sali, A. 2004. Comparative protein structure modeling and its applications to drug discovery. *Annu. Rep. Med. Chem.* 39: 259-276.
4. Dobson, C.M. 2003. Protein folding and misfolding. *Nature* 426: 884-890.
5. Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* 294: 93-96.
6. Blundell, T.L., *et al.* 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326: 347-352.
7. Marti-Renom, M.A., *et al.* 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29: 291-325.
8. Bowie, J.U., *et al.* 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253: 164-170.
9. Godzik, A., *et al.* 1992. Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* 227: 227-238.
10. Bonneau, R. and Baker, D. 2001. *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* 30: 173-189.
11. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
12. Park, J., *et al.* 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284: 1201-1210.
13. Jones, D.T., *et al.* 1992. A new approach to protein fold recognition. *Nature* 358: 86-89.
14. Sippl, M.J. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5: 229-235.
15. Jones, D.T. and Thornton, J.M. 1996. Potential energy functions for threading. *Curr. Opin. Struct. Biol.* 6: 210-216.
16. Kim, D., *et al.* 2003. PROSPECT II: protein structure prediction program for genome-scale applications. *Protein Eng.* 16: 641-650.
17. McGuffin, L.J. and Jones, D.T. 2003. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 19: 874-881.
18. Tanaka, S. and Scheraga, H.A. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 9: 945-950.
19. Miyazawa, S. and Jernigan, R.L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256: 623-644.
20. Huang, E.S., *et al.* 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252: 709-720.
21. Buchete, N.V., *et al.* 2004. Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.* 14: 225-232.
22. Fang, Q. and Shortle, D. 2005. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. *Proteins* 60: 90-96.
23. Mayewski, S. 2005. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. *Proteins* 59: 152-169.
24. Park, B.H., *et al.* 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 266: 831-846.
25. Vendruscolo, M., *et al.* 2000. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins* 38: 134-148.
26. Khatun, J., *et al.* 2004. Can contact potentials reliably predict stability of proteins? *J. Mol. Biol.* 336: 1223-1238.
27. Anfinsen, C.B., *et al.* 1954. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J. Biol. Chem.* 207: 201-210.
28. Berman, H.M., *et al.* 2000. The protein data bank. *Nucleic Acids Res.* 28: 235-242.
29. Eisenberg, D. and McLachlan, A.D. 1986. Solvation energy in protein folding and binding. *Nature* 319: 199-203.
30. Cavallo, L., *et al.* 2003. POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.* 31: 3364-3366.
31. Bryant, S.H. and Altschul, S.F. 1995. Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* 5: 236-244.
32. Lee, B. and Richards, F.M. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55: 379-400.
33. Hayryan, S., *et al.* 2005. A new analytical method for computing solvent-accessible surface area of macromolecules and its gradients. *J. Comput. Chem.* 26: 334-343.
34. Pearl, F.M., *et al.* 2003. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* 31: 451-455.