SHORT COMMUNICATION

Relationship Between Protein Structures and Disulfide-Bonding Patterns

Chao-Chun Chuang,[1,2] Chun-Yin Chen,[2] Jinn-Moon Yang,[2] Ping-Chiang Lyu,[1*] and Jenn-Kang Hwang[2*]
[1]*Department of Life Sciences, Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsin Chu, Taiwan*
[2]*Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin Chu, Taiwan*

ABSTRACT **We found that that disulfide-bonding patterns can be used to discriminate structure similarity. Our method, based on the hierarchical clustering scheme, is applicable to proteins with two or more disulfide bonds and is able to detect the structural similarities of proteins of low sequence identities (<25%). Our results show the surprisingly close relationship between disulfide-bonding patterns and proteins structures. Our findings should be useful in protein structure modeling. Proteins 2003;53:1–5.** © 2003 Wiley-Liss, Inc.

Key words: disulfide-bonding patterns; the hierarchical clustering method; structure classification

INTRODUCTION

Disulfide bonds are common to many proteins and are known to play a key role in stabilizing protein structures.[1–5] Disulfides bonds help stabilize the folded states by increasing favorable enthalpy interactions in the folded states and by lowering the entropy of the unfolded states.[6] Protein folding simulations[2,7,8] show that inclusion of disulfide-bond constraints helps reduce the search of protein conformations. Because disulfide bonds impose distance and angular constraints on the protein backbones, one would expect that disulfide bonds should exert significant constraints on the overall three-dimensional (3D) protein structures. Harrison and Sternberg[9] reported that, although the small disulfide-rich protein folds are problematic in protein structure taxonomy and prediction, the regularities in disulfide-bridged β-sheets and in cystine clusters can be used to classify their folds. Recently, Mas et al.[10] developed an approach KNOT-MATCH to superimpose protein structures that contain three or more disulfide bonds by means of 3D disulfide bridge topology. Using this approach, they are able to find relationships among proteins that are hidden to the current alignment methods based on sequence or main-chain topology.

However, because the number of protein structures is far less than that of protein sequences, it will be of great value if one can detect structural similarity directly from protein sequences. A lot of work has been done to develop approaches to detect structural similarity directly from protein sequences by using sequence profiles[11,12] or hidden Markov models (HMM).[13,14] For example, PDB-BLAST[15] uses PSI-BLAST[11] to generate sequence profiles for specific protein families, and these profiles are then used to scan protein structure databases. 3D-PSSM[16] uses 1D and 3D profiles coupled with secondary structure and solvation potentials to predict protein folds. *prof_sim*[17] is a profile–profile comparison method to detect structural similarity of remote homologues. SAM-T99[18] builds a multiple-sequence alignment by iterated search using HMM. There are other approaches based on various algorithms such as the support vector machine,[19] threading techniques,[20,21] or the multistrategy approach,[22] which combines several methods to use sequence and structure information in different ways to generate one consensus structure. In this work, we report that it is possible to use disulfide-bonding patterns instead of the complete protein sequences to discriminate protein folds. This idea is analogous to that of Mas et al.,[10] who use disulfide bridge topology instead of the complete main-chain topology to superimpose structures.

MATERIALS AND METHODS

We first define the terms used in this work: for two disulfide proteins A and B, each having n disulfide bonds, we denote their disulfide-bonding pairs by $(x_1 - x_{n+1}, x_2 - x_{n+2}, \ldots, x_n - x_{2n})$ and $(y_1 - y_{n+1}, y_2 - y_{n+2}, \ldots, y_n - y_{2n})$, respectively, where $x_i - x_{n+i}$ and $y_i - y_{n+i}$ are the sequence numbers of the cystine pair forming the $i$thdisul-

Grant sponsor: National Science Council in Taiwan, Republic of China.

*Correspondence to: Jenn-Kang Hwang, Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin Chu, Tawiwan. E-mail: jkhwang@cc.nctu.edu.tw or, Ping-Chiang Lyu, Department of Life Sciences, Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsin Chu, Taiwan, E-mail: lslpc@life.nthu.edu.tw

Received 25 December 2002; Accepted 28 April 2003

fide bond. In a similar way, for proteins A and B, we denote their disulfide-bonding connectivity by $(N_1 - N_{n+1}, N_2 - N_{n+2}, \ldots N_n - N_{2n})$ and $(M_1 - M_{n+1}, M_2 - M_{n+2}, \ldots, Mn - M_{2n})$, respectively, where $N_i - N_{n+i}$ and $M_i - M_{n+i}$ are the relative orders of the cystine pair forming the $i$thdisulfide bond. For instance, the notation [1-3,2-4] means that the first and the third cysteines form the first disulfide bond, and the second and fourth cysteines form the second disulfide bond. Using these notations, we cluster the disulfide-boding patterns by the following equations:

$$\alpha = \sum_{i=1}^{2n} (x_i - \bar{x})\ (y_i - \bar{y}) / \sqrt{\sum_{i=1}^{2n} (x_i - \bar{x})^2 \sum_{i=1}^{2n} (y_i - \bar{y})^2}, \quad (1)$$

$$\beta = \sum_{i=1}^{n} |\Delta N_i - \Delta M_i| / n, \quad (2)$$

where $\bar{x} = 1/2n \sum_i^{2n} x_i$ and $\bar{y} = 1/2n \sum_i^{2n} y_i$, and $\Delta N_i = N_{i+n} - N_i$ and $\Delta M_i = M_{i+n} - Mi$. If $\alpha \geq \alpha_0$ and $\beta \leq \beta_0$, both proteins are defined as having the same disulfide-bonding pattern. We set the values of α_0 and β_0 to 0.996 and 3.0.

Data Sets

We collect all disulfide proteins with two or more disulfide bonds from Protein Data Bank (PDB),[23] and the data set is composed of 3134 disulfide chains that are defined in the PDB file records. Each chain is treated as a separate unit, and the interchain disulfide linkages are not considered. Disulfide chains are classified hierarchically in three levels: disulfide-bonding numbers, disulfide-bonding connectivity, and disulfide-bonding patterns. The hierarchical classification is shown schematically in Figure 1. In this work, all pairwise sequence comparisons and structure alignments are computed by ALIGN[24] and CE,[25] respectively. The root-mean-square deviation (RMSD) values reported are for C_α atoms.

RESULTS

The protein pairs in the same cluster group are shown in Figures 2–4. Figure 2 shows the structures of (a) the tick anticoagulant peptide (1tap[26]), a serine protease inhibitor, and (b) cacicludine (1bf0[27]), a calcium channel blocker. These proteins are clustered in the same disulfide-bonding patterns, which have the disulfide-bonding connectivity [1-6,2-3,4-5]. Their RMSD value of C_α atoms is 3.6 Å, but their sequence identity is only 18.2%. In this cluster group, we found a total of 92 disulfide chains, all of which are classified in the BPTI-like *superfamily* in SCOP.[28] The complete list can be accessed from the SSDB website.[29] Figure 3 shows (a) thionin (1gps[30]), a plant toxin , and (b) brazzein (1brz[31]), a sweet protein. Their RMSD value is 2.3 Å and their sequence identity is 18.8%. All proteins in this cluster group have the scorpion toxin-like structures.[28] Figure 4 shows (a) tetranectin (1tn3[32]) and (b) the α-monomer of flavocetin-A (1c3a:a[33]). These proteins have 17.7% sequence identity and an RMSD value of 1.5 Å. Despite the different orientations of their loops, both proteins have a C-type lectin fold.[28] Automatic structure alignment programs such as VAST,[34] FSSP,[35] or CE[25] are not able to detect their structure similarities from the database, although both proteins are classified in the C-type lectin domain *family* in SCOP, which is based on extensive expert knowledge. Further analysis shows that the proteins of this cluster group are classified into five SCOP *domains*[28]: 1) snake coaggultinin, 2) the asialoglycoprotein receptor, 3) CD69, macrophage mannose receptor CRD4, 4) tetranectin, and 5) lithostathine. Figure 5 shows the RMSD values versus sequence identities of the proteins in this cluster group. The pairwise sequence identities of these proteins vary in wide ranges, but their 3D structures are similar.

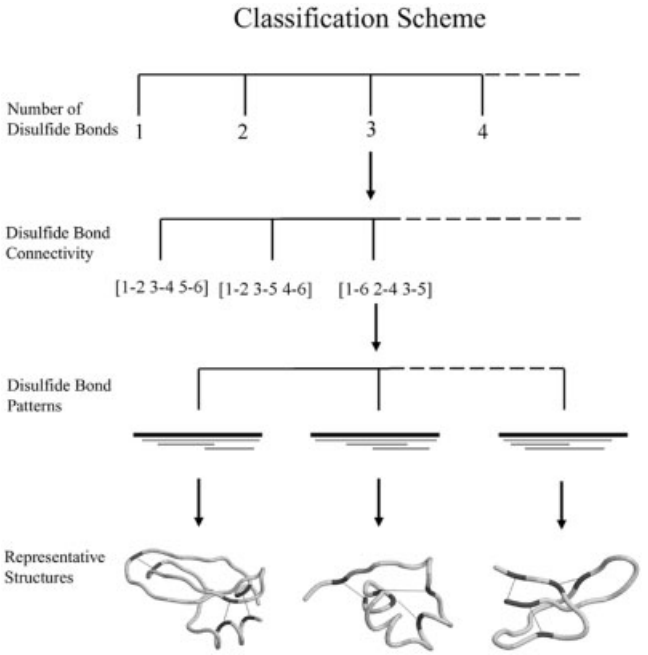

Fig. 1. The hierarchical classification of disulfide proteins, starting from the disulfide-bonding numbers, the disulfide-bonding connectivity, and to the disulfide-bonding patterns. In the schematics of the disulfide-bonding patterns, the first thick line represents the total protein lengths, and the thin lines represent the cystine bridges.

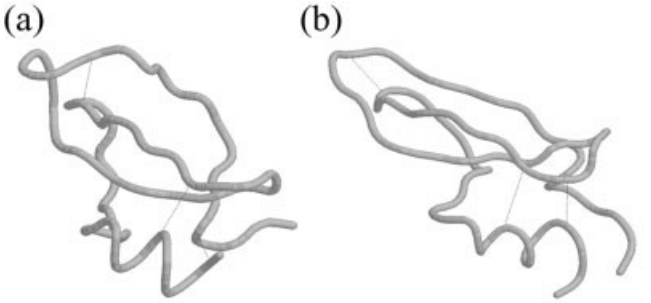

Fig. 2. 1tap, an anticoagulant protein (**a**) and 1bf0, a calcium channel blocker (**b**), each having four disulfide bonds [1-6, 2-3, 4-5]. Both proteins have a BPTI-like structure and a sequence identity of 18.2%. The protein images are rendered by Rasmol[41] in the trace model. The disulfide bonds are indicated by dotted lines.

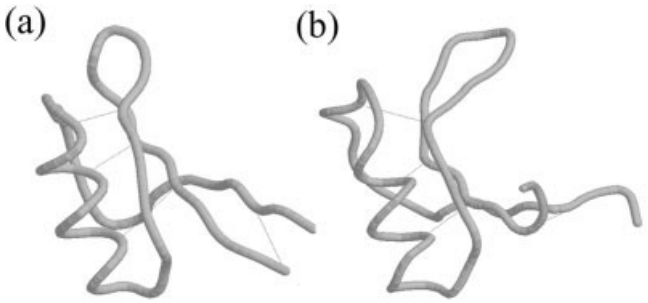

Fig. 3. 1gps, a plant toxin (**a**) and 1brz, a sweet protein called brazzein (**b**), each having four disulfide bonds [1-8, 2-5, 3-4, 6-7]. Both proteins have a scorpion toxin-like structure and 18.8% sequence identity.

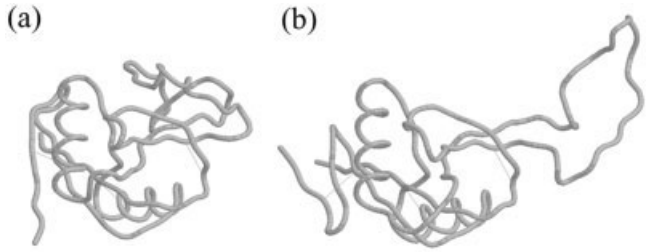

Fig. 4. 1tn3, tetranectin (**a**) and 1c3a:a, the α-monomer of flavocetin-A (**b**). Both proteins have a disulfide-bonding connectivity [1-2, 3-6, 4-5]. Both proteins have C-type lectin folds, despite the different orientations of their loops. Their RMSD value and sequence identity are 1.5 Å and 17.7%, respectively.

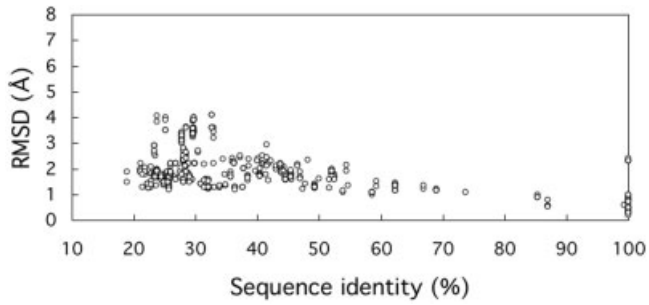

Fig. 5. The plot of sequence identities versus RMSD values of 32 disulfide proteins[16] in the same cluster group, including 1tn3 and 1c3a:a shown in Figure 4. All these proteins have C-type lectin folds and similar disulfide-bonding patterns.

We performed exhaustive pairwise comparisons of both sequence similarities and structure similarities of all 3134 disulfide-bonding chains in the PDB. Figure 6 shows the plot of the RMSD values of C_α atoms versus sequence identities of every pair of disulfide proteins whose sequence length ratios are >70%. The trends of the RMSD values are a familiar one: the structural deviations remain relatively flat and then rise sharply at around 25–30% sequence identities, which are the usual lower bounds of sequence identity set by the homology modeling methods of protein structures. For comparison, we also performed pairwise comparisons of the structure similarities in the same cluster groups. The results are shown in Figure 7. The RMSD values remain flat throughout the range of sequence identities, and there is no sharp rising of RMSD

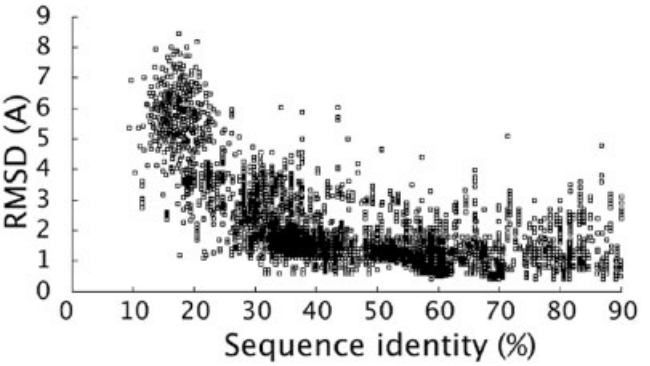

Fig. 6. The RMSD values of C_α against sequence identities of all disulfide chains in PDB. Only protein pairs whose length ratios are ≥70% are computed.

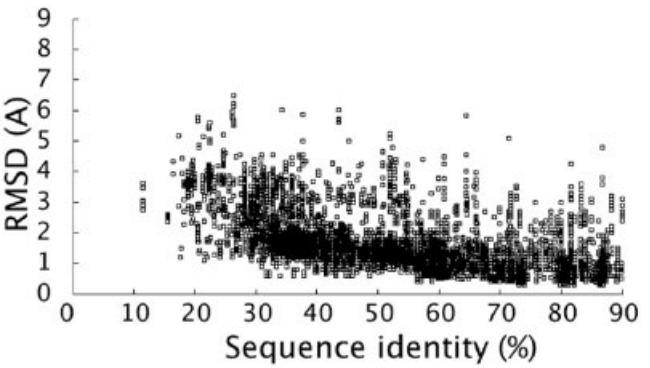

Fig. 7. The RMSD values of C_α against sequence identities of the disulfide chains in the same cluster group of disulfide-bonding patterns.

values. There are some scattering points with relatively higher RMSD values, which, under visual inspection, do in fact have similar structures; 90% of the proteins that are in the same cluster groups are also classified in the same SCOP *families*, which comprise proteins of sequence identities of 30% or greater, or proteins of lower sequence identities but of similar structures and functions.[28] Other proteins, although not belonging to the same SCOP *families*, are found to be in the same SCOP *superfamilies*, which share a common evolutionary origin[36,37] due to functional similarities or common features unlikely to have occurred randomly.

Use of Disulfide-Bonding Patterns in Structure Prediction

We can exploit the relationship between the disulfide-bonding patterns and structures to predict protein folds directly from disulfide-bonding patterns without the need of complete sequences. An example is the nonspecific lipid transfer protein (nsLTP2) from rice, whose structure[38] (1l6h) was recently solved after we completed the library of the disulfide-bonding patterns. NsLTPs are divided into two families, nsLTP1 and nsLTP2. Many structures of nsLTP1 have been solved,[39] but 1l6h is the only nsLTP2 whose structure is solved. Rice nsLTP2 has <30% sequence identity with nsLTP1s, and its cysteine-pairing

pattern is different from nsLTP1. However, using Eq. 1, we find one protein that has the same disulfide-bonding pattern as rice nsLTP2. This protein, soybean hydrophobic protein[40] (1hyp), has 16.1% sequence identity with rice nsLTP2. Our approach predicts that these two proteins should have a similar fold, and this is indeed the case, because they have an RMSD value of 4.2 Å. Because our approach does not need complete sequences, it has the advantage of finding structural templates of little sequence similarities to the query sequence. However, if the disulfide-bridge pattern does not exist in the library, then our approach will not work. Such limitations also exist in other structural template-based approaches.

DISCUSSION

In this work, we demonstrate for the first time that disulfide-bonding patterns can be effectively used to discriminate structure similarities between proteins. For the homologous sequences, one would expect that their disulfide-bonding patterns are similar. However, we show that there is a very close relationship between the disulfide-bonding patterns and the protein structures and that such relationship holds in the case of low sequence similarity (sequence identities $< 25\%$). An interesting question arises as to whether the relationships found by our approach are due to purely geometrical constraints, which, allowing only a few possibilities in protein conformations, force the structures to conserve; or whether the relationships are due to sequence divergence with conserved structures. In general, the presence of only a structure similarity does not allow us to clearly distinguish between these two possibilities. However, according to Russell et al.,[37] homologs and analogs can be distinguished by means of SCOP data set based on extensive expert knowledge. Proteins within the same SCOP *superfamily* are taken to be homologous due to obvious functional similarities or common characteristics unlikely to have occurred randomly, even though these proteins often lack sequence similarity. Analogs are defined as proteins with similar 3D structures but generally with different functions and little evidence of a common ancestor (within the same *fold* but in different *superfamilies*). We found in the Results section that the proteins of each cluster group always belong to the same *families* or *superfamilies*, and never in different *folds*. Our results seem to suggest that the relationship between disulfide-bonding patterns and protein structures comes from sequence divergence. This conclusion is also consistent with the observation[9] that many of the similarities in the disulfide-bridge topology may have diverged from a common ancestor, such as the α/β scorpion toxins. However, it is obvious that further investigations are needed to draw a definite conclusion.

ACKNOWLEDGMENT

This work was supported by grants to J.K.H. and P.C.L. by National Science Council in Taiwan, Republic of China.

REFERENCES

1. Clark J, Fersht A. Engineered disulfide bonds as probes of the folding pathway of branase–increasing the stability of proteins against the rate of denaturation. Biochemustry 1993;32:4322–4329.
2. Abkevich VI, Shaknovich EI. What can disulfide bond tells us about protein energetics, function and folding: simulations and bioinformatics analysis. J Mol Biol 2000;300:975–985.
3. Clarke J, Hounslow AH, Bond CJ, Fersht AR, Dagget V. The effects of disulfide bonds on the denatured state of barnase. Protein Sci 2000;9:2394–2404.
4. Wedemeyer WJ, Welker E, Narayan M, Scheraga HA. Disulfide bonds and protein folding. Biochemistry 2000;39:4207–4215.
5. Yokota A, Izutani K, Takai M, Kubo Y, Noda Y, Koumoto Y, Tachibana H, Segawas S. The transition state in the folding-unfolding reaction of four species of three-disulfide variant of hen lysozyme: the role of each disulfide bridge. J Mol Biol 2000;295:1275–1288.
6. Anfinsen C, Scheraga HA. Principles that govern the folding of protein chains. Adv Protein Chem 1975;29:205–299.
7. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
8. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267–281.
9. Harrison PM, Sternberg MJE. The disulfide beta-cross:from cystine geometry and clustering to classification of small disulfide-rich protein folds. J Mol Biol 1996;264:603–623.
10. Mas JM, Aloy P, Marti-Renom MA, Oliva B, Blanco-Aparicio C, Molina MA, de Llorens R, Quero lE, Aviles FX. Protein similarities beyond disulphide bridge topology. J Mol Biol 1998;284:541–548.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
12. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. MPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999;15:1000–1011.
13. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235:1501–1531.
14. Eddy SR. Hidden Markov models. Curr Opin Struct Biol 1996;6:361–365.
15. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
16. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
17. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
18. Karplus K, Barret C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins 1999;37:121–125.
19. Yu C-S, Wang J-Y, Yang J-M, Lin CH, Hwang J-K. Fine-grained SCOP protein fold assignment by support vector machines using generalized n-peptide coding schemes. Proteins 2003;50:531–536.
20. Jones DT. THREADER: protein sequence threading by double dynamic programming. In: Salzberg SL, Searls DB, Kasif S, editors. Computational methods in molecular biology. Amsterdam: Elsevier; 1998. p 285–311.
21. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
22. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
24. Myers EW, Miller W. Optimal alignments in linear space. Comput Appl Biosci 1989;4:11–17.
25. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
26. Antuch W, Guntert P, Billeter M, Hawthorne T, Grossenbacher H, Wuthrich K. NMR solution structure of the recombinant tick

anticoagulant protein (rTAP), a factor Xa inhibitor from the tick Ornithodoros moubata. FEBS Lett 1994;352:251–257.
27. Gilquin B, Lecoq A, Desne F, Guenneugues M, Zinn-Justin S, Menez A. Conformational and functional variability supported by the BPTI fold: solution structure of the Ca2+ channel blocker calcicludine. Proteins 1999;34:520–532.
28. Lo Conte L, E BS, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
29. Chuang CC, Hwang J-K. The structural classification of disulfide protein database: SSDB. http://e106.life.nctu.edu.tw/~ssbond
30. Bruix M, Jimenez MA, Santoro J, Gonzalez C, Colilla FJ, Mendez E, Rico M. Solution structure of gamma 1-H and gamma 1-P thionins from barley and wheat endosperm determined by 1H-NMR: a structural motif common to toxic arthropod proteins. Biochemistry 1993;32:715–724.
31. Caldwell JE, Abildgaard F, Dzakula Z, Ming D, Hellekant G, Markley JL. Solution structure of the thermostable sweet-tasting protein brazzein. Nat Struct Biol 1998;5:427–431.
32. Nielsen BB, Kastrup JS, Rasmussen H, Holtet TL, Graversen JH, Etzerodt M, Thogersen HC, Larsen IK. Crystal structure of tetranectin, a trimeric plasminogen-binding protein with an alpha-helical coiled coil. FEBS Lett. 1997;412:388–396.
33. Fukuda K, Mizuno H, Atoda H, Morita T. Crystal structure of flavocetin-A, a platelet glycoprotein Ib-binding protein, reveals a novel cyclic tetramer of C-type lectin-like heterodimers. Biochemistry 2000;39:1915–1923.
34. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
35. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–560.
36. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
37. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
38. Samuel D, Liu YJ, Cheng CS, Lyu PC. Solution structure of plant nonspecific lipid transfer protein-2 from rice (Oryza sativa). J Biol Chem 2002;277:35267–35273.
39. Sodano P, Caille A, Sy D, de Person G, Marion D, Ptak M. 1H NMR and fluorescence studies of the complexation of DMPG by wheat non-specific lipid transfer protein. Global fold of the complex. FEBS Lett 1997;416:130–134.
40. Baud F, Pebay-Peyroula E, Cohen-Addad C, Odani S, Lehmann MS. Crystal structure of hydrophobic protein from soybean; a member of a new cysteine-rich family. J Mol Biol 1993;231:877–878.
41. Sayle R, Bissel A. RasMol: a program for fast realistic rendering of molecular structures with shadows. Proceedings of the 10th Eurographics UK'92 Conference, 1992.