PROTEINS: Structure, Function, and Genetics Suppl 3:199–203 (1999)

Prediction of Protein Structure: The Problem of Fold Multiplicity

Andrei L. Lomize,* Irina D. Pogozheva, and Henry I. Mosberg
College of Pharmacy, University of Michigan, Ann Arbor, Michigan

ABSTRACT **Three-dimensional (3D) models of four CASP3 targets were calculated using a simple modeling procedure that includes prediction of regular secondary structure, analysis of possible β-sheet topologies, assembly of amphiphilic helices and β-sheets to bury their nonpolar surfaces, and adjustment of side-chain conformers and loops to provide close packing and saturation of the "hydrogen bond potential" (exposure of all polar groups to water or their involvement in intramolecular hydrogen bonds). It has been found that this approach allows construction of 3D models that, in some cases, properly reproduce the structural class of the protein (such as β-barrel or β-sandwich of definite shape and size) and details of tertiary structure (such as pairing of β-strands), although all four models were more or less incorrect. Remarkably, some models had fewer water-exposed nonpolar side-chains, more hydrogen bonds, and smaller holes than the corresponding native structures (although the models had a larger water-accessible nonpolar surface). The results obtained indicate that hydrophobicity patterns do not unequivocally determine protein folds, and that any ab initio or fold recognition methods that operate with imprecise potential energy functions, or use crude geometrical approximations of the peptide chain, will probably produce many different nonnative structures. Proteins Suppl 1999;3:199–203.** © 1999 Wiley-Liss, Inc.

Key words: CASP; secondary structure; modeling; hydrophobic interactions; hydrogen bonds; beta-sheet

INTRODUCTION

One of the most important questions in the protein structure prediction field is which energy contributions must be taken into account in the modeling procedure. Hydrophobic interactions represent one of the "dominant forces" in protein folding.[1] Therefore, some simplified lattice simulations take into account only burial of nonpolar residues from water. It is also possible to simultaneously optimize hydrophobic interactions and hydrogen bonding.[2] In general, there are three most important criteria of "good protein structure":

1. Burial of nonpolar side-chains.
2. Saturation of "hydrogen bonding potential."[3]
3. Stereochemical quality, i.e., close packing with no hindrances or holes.[4]

However, the question arises: how many different structures, or folds can simultaneously satisfy all three criteria? If the native "fold" (an approximate arrangement of regular secondary structures) is unequivocally determined by quantities that can be calculated easily (such as the numbers of buried nonpolar residues or hydrogen bonds), the fold could be identified by using crude geometrical models of the peptide chain and imprecise potentials, and refined later at the atomic level with better potential energy functions. If this is not the case, some additional energy contributions must be included or alternative strategies must be applied. To study this question, we have constructed precise full-atomic models that satisfy all criteria of "good structures" for four CASP3 targets and compared them with an experiment.

METHODS AND RESULTS

Our goal was a study of factors that must be taken into account in the modeling of protein structure, rather than the design of automated methods. The modeling procedure was kept as simple as possible and included the following steps.

1. Identification of tentative regular secondary structures as continuous segments with high content of nonpolar residues and conserved hydrophobicity patterns through sequence alignments;
2. Choice of an amphiphilic α-helix or a β-strand for each segment based on the hydrophobicity pattern and secondary structure propensities.
3. Assembly of the β-strands into β-sheet(s) to maximize continuous areas formed by nonpolar side-chains.
4. "Manual" (using QUANTA) docking of the amphiphilic α-helices and β-sheets to bury their nonpolar surfaces and provide "knobs into holes" packing of side-chains.
5. Adjustment of side-chain conformers and loops to provide close packing and maximize hydrogen bonding in the model using distance geometry refinement with H bond and other constraints (with the program DIANA[5]), as described for the modeling of rhodopsin

Grant sponsor: National Institutes of Health; Grant numbers: DA03910 and DA09989.
*Correspondence to: Andrei L. Lomize, College of Pharmacy, University of Michigan, Ann Arbor, MI 48109. E-mail: almz@umich.edu
Received 1 February 1999; Accepted 10 May 1999

and other G-protein coupled receptors[6] (http://www-personal.umich.edu /~him/modeling.htm). Some details of the refinement are described in our coordinate files available through the Protein Structure Prediction Center (http://predictioncenter.llnl.gov).

The modeling procedure can be illustrated using the Sm D3 protein (T0059).[7] This protein has six continuous nonpolar segments (residues 1–10, 18–23, 28–32, 39–48, 56–63, and 68–73) with hydrophobicity patterns that are conserved throughout the amino acid sequence alignment of 13 proteins from the Sm D3 family constructed by the gapped BLASTP program. Using "HP" (H-hydrophobic, P-polar residue) language, the patterns of segments 1 to 6 can be summarized as HPHPHPHPHH, HPHPHP, HP-Gly-PH, HPHPHPPHPH, HPHPPHPH, and HPHHPH, respectively. All segments were assigned as β-strands, because they have typical β-sheet (*i, i*+2) hydrophobicity patterns and high β-sheet propensity (a significant content of Ser, Thr, Val, Ile, and aromatic residues). The PP and HH irregularities in the patterns of the last three strands were interpreted as β-bulges. Next, we constructed a β-sheet with maximum buried nonpolar surface from the β-strands. The maximal nonpolar surface of the single β-sheet can be provided if the two longest β-strands, 1 and 4, are brought together, which is possible with a Greek key connection between the third and fourth strands. This produces a β-sheet with ladder-like shift of β-strands, as is necessary for its cyclization to a β-barrel. The share number of the β-barrel can be identified as eight by taking into account the length and number of the participating β-strands.[8] Importantly, all three β-bulges are located in proper places to provide the curvature of the β-barrel. Next, the geometry of the β-barrel and conformers of side-chains and loops were refined by distance geometry calculations using the cyclic system of backbone hydrogen bonds in the β-barrel and other constraints.

The second target, cyanovirin-N (T0052),[9] consists of two structural repeats (residues 1–50 and 51–101) with clear sequence homology to each other. We found that each repeat can form an amphiphilic β-sheet with the simplest "β-meander" ("up and down") topology of four strands and a large, continuous, nonpolar surface interrupted by a single polar spot: Ser^{20} and Ser^{38} residues in the first repeat and Asp^{89} in the second. After packing these two β-sheets in a "sandwich" that provides burial of their nonpolar surfaces, OγH groups of the $Ser^{20, 38}$ residues form H-bonds with the O_γ oxygen of $Asp.^{89}$ Thus, it was suggested that cyanovirin-N forms a β-sandwich similar to that in artificially constructed betabellin.[10] The third target, γ-adaptin ear domain (T0046), had hydrophobicity patterns characteristic for alternating amphiphilic α-helices and β-strands, so it was suggested that it forms an α/β fold with a mixed β-sheet (its N-terminal portion forms a β-hairpin). For the fourth protein, HdeA (T0061),[11] we constructed an antiparallel β-sheet from three β-strands with several continuous nonpolar surface spots that matched precisely hydrophobic arcs of three surrounding amphiphilic α-helices (exactly as in the ear domain), which made possible the burial of nearly all nonpolar side-chains.

DISCUSSION

A comparison of our "blind" predictions with experimental structures shows that in two cases (Sm D3 protein and cyanovirin-N), the structural class of the protein (β-barrel and β-sandwich, respectively) and its shape were predicted correctly (Fig. 1), whereas in two others (ear domain and HdeA) we calculated nonnative folds (α/β instead of β-sandwich, and α+β instead of all-α, respectively), with nearly all nonpolar residues buried. The most precisely calculated "local" element of tertiary structure was β-hairpin 2–18 of the ear domain, where the pairing of β-strands, location of a β-bulge, the type of β-turn, and even conformers of many side-chains were identified correctly (Fig. 2). In some other cases (β-hairpins 40–64 of Sm D3 and 8–22 and 58–73 of cyanovirin-N), the pairing of β-strands was predicted correctly; however, the "narrow" and "wide" H-bond pairs were improperly exchanged, or the "narrow" and "wide" pairs were identified correctly, but two adjacent β-strands were shifted "in register" by two residues (β-hairpin 14–34 of Sm D3).

The model of the Sm D3 domain looks especially similar to the experimental structure (Fig. 1). Five β-strands, all "turning points" between regular secondary structures, and all three β-bulges were detected correctly. However, the topology of the β-sheet is wrong: the β-sheet actually has no Greek key connection and forms an "up and down" β-meander from four β-strands, whereas its fifth β-strand passes from one side of the β-sheet to the other. It is noteworthy that the sequence alignment of proteins from the Sm D3 family did not help to determine the β-sheet topology. The alignment helps to identify buried and water-exposed residues; however, the sets of buried and exposed residues in both topologies are almost the same.

The β-sheet topology of Sm D3 was incorrectly predicted because the suggested N-terminal " β-strand," with typical β-structural propensity and hydrophobicity pattern, is partially disordered in the crystal (residues 1–4) and partially forms an α-helix (residues 6–13). However, an even more important cause of the incorrect topology prediction was the erroneous structure of the β-hairpin 40–-64 that probably initiates formation of the entire β-sheet in the Sm D3 protein. In the model, two strands in this β-hairpin have correct alignment; however, their left-to-right positions (looking from the inner side of the β-barrel) were erroneously exchanged, which automatically resulted in exchange of narrow and wide H-bond pairs and incorrect arrangement of other β-strands. The modeled structure of this β-hairpin is obviously less energetically preferred than the experimental one: for example, it cannot form the standard $\alpha_R\gamma_R\alpha_L$ β-turn[12] that is present in the experimental structure and was shown to be important for stabilization of the β-hairpins in model peptides.[13] In addition, this exchange produced poorer hydrophobic contacts between nonpolar side-chains in the β-hairpin, especially those arranged in diagonal[14] positions.

Apparently, the maximization of nonpolar β-sheet surface must be considered to be an incomplete strategy. Instead, it is necessary to calculate stabilities of all possible initiating β-hairpins (taking into account energies of β-turn motifs and side-chain interactions), and to con-

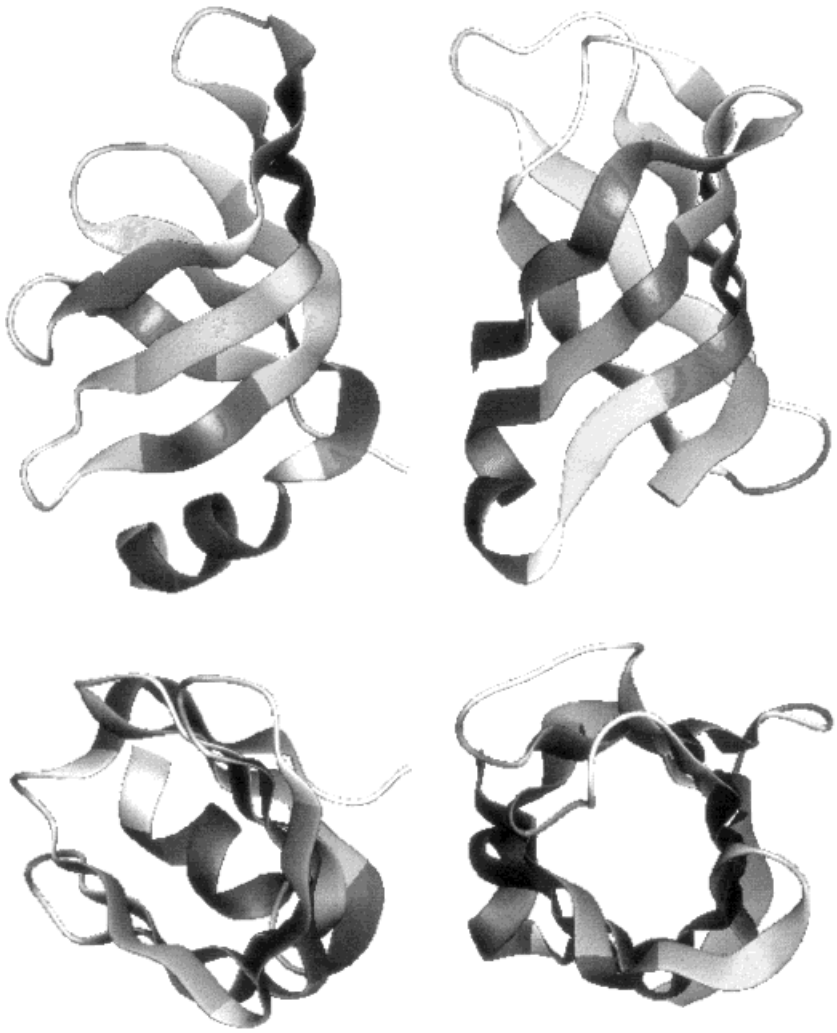

Fig. 1. Side and top views of experimental structure (**left**) and predicted model (**right**) of Sm D3 protein.

sider all possible ways of their association in the β-sheet. Then, the erroneously suggested N-terminal β-strand of Sm D3 could be excluded as a redundant element from the β-barrel and could form an α-helix.

The result obtained for cyanovirin-N is similar to that for Sm D3: in the experimental structure, each repeat contains two independent β-sheets (β-hairpin and β-meander from three β-strands), not a single β-meander from four strands, as was suggested in the model. Again, use of a continuous nonpolar surface, as a single criterion, was unable to reproduce an exact β-sheet structure, because this structure depends on the balance of many interactions. At the same time, the general architecture of the protein (a two-layer β-sandwich from several β-sheets with "up and down" topology) still has some resemblance to the native structure. On the other hand, the calculated folds of the γ-adaptin ear domain and HdeA, were completely nonnative. This resulted because the amino acid sequences of these proteins form atypical hydrophobicity patterns, with buried polar and/or exposed nonpolar residues (the proteins form significant exposed nonpolar surfaces, because HdeA is a dimer, and ear domain is a part of a larger protein). In the ear domain, several adjacent β-strands have hydrophobicity patterns of amphiphilic helices. The incorrect assignment of β-strands as α-helices yielded an α/ β fold instead of a β-sandwich. An opposite situation occurs in HdeA: an amphiphilic β-sheet was constructed instead of α-helices. In the model, the surface of this β-sheet forms several nonpolar spots that match nonpolar arcs of three surrounding α-helices. However, the first of these helices actually corresponds to a disordered protein fragment in the crystal. In these examples, a combination of specific secondary and tertiary interactions overrides the local hydrophobicity pattern.

The most interesting feature of the calculated models is that they meet the structure quality criteria mentioned above better than do the native structures. Indeed, some models have fewer water-exposed nonpolar side-chains, smaller holes, and significantly more intramolecular backbone and side-chain H-bonds (Table I). Inspection of the models also shows that all of their polar groups are accessible to water, or form intramolecular H bonds. The models have no significant interatomic hindrances, and all of their side-chain and backbone conformers are sterically allowed, because this is automatically provided by the van der Waals and dihedral angle constraints incorporated in

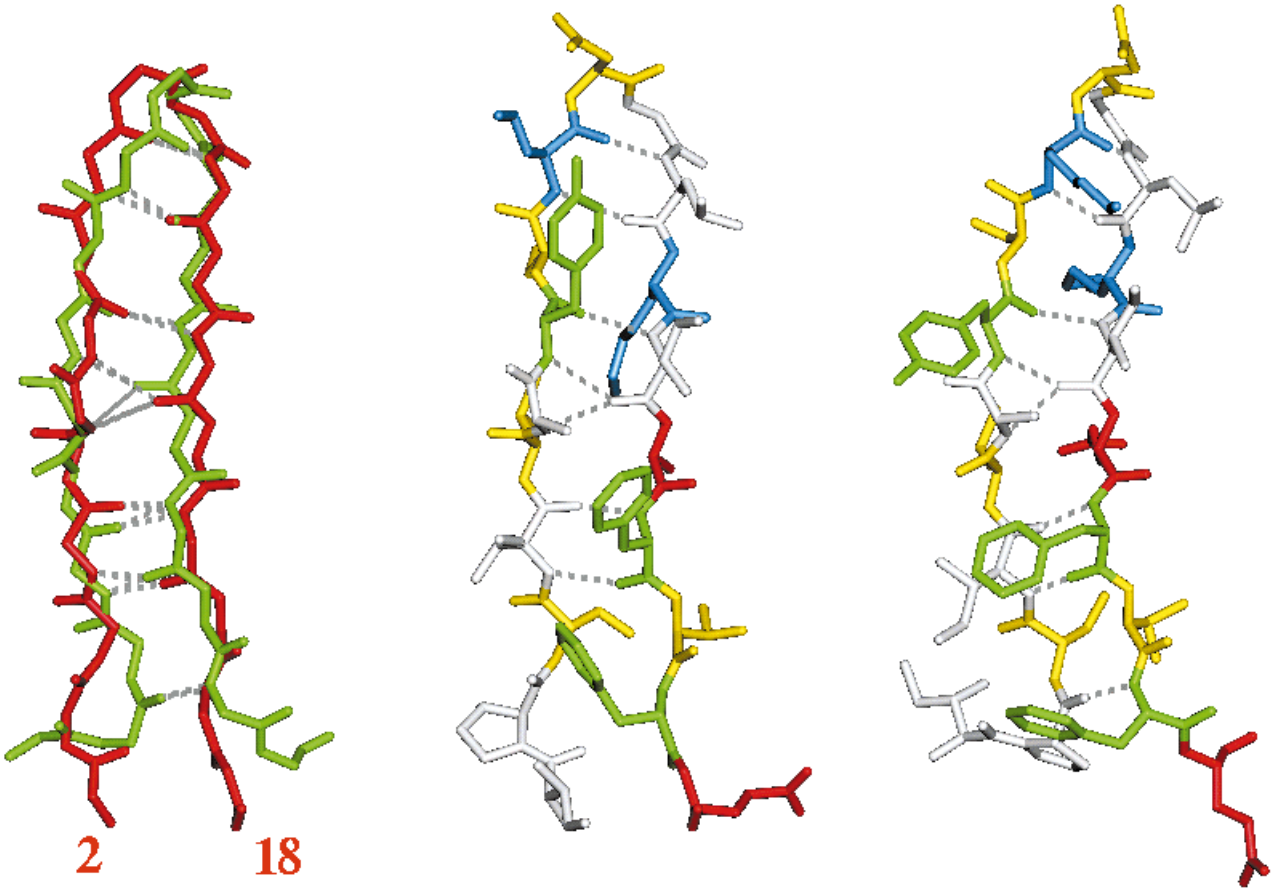

Fig. 2. Comparison of experimental structure (**middle**) and model (**right**) of β-hairpin 2–18 from γ-adaptin ear domain. The residues are colored by polarity: aliphatic—gray, aromatic—green, polar (Ser, Thr, Asn, and Gln)—yellow, negatively charged (Asp and Glu)—red, and positively charged (Lys and Arg)—blue. **Left**: Backbone superposition of model (green) and crystal structure (red).

TABLE I. Parameters of Models (Model) and Experimental Structures (Exper.): The Number of Water-Exposed Nonpolar Side-Chains ($N_{exposed}$), the Number of Total and Backbone ($N_{Hb,tot}$ and $N_{Hb,bb}$, Respectively) Intramolecular H Bonds, the Number of Holes (N_{holes}) and Their Water-Accessible Volume (V_{holes}, A^3), and the Differences in Nonpolar Water-Accessible Surface (ΔS_{nonp}, $Å^2$, and ΔG, kcal/mol)†

	T0046		T0052		T0059		T0061	
	Model	Exper.	Model	Exper.	Model	Exper.	Model	Exper.
$N_{exposed}$	12	14	10	8	8	11	14	12
$N_{Hb,tot}$	146	109	101	76	75	75	99	90
$N_{Hb,bb}$	89	62	61	49	49	48	66	62
N_{holes}	5[a]	2	1	1	None	1	1	2
V_{holes}	4.2[a]	0.0	0.1	2.4		0.1	0.0	1.3
ΔS (ΔG)	349 (7.0)		332 (6.6)		128 (2.6)		348 (7.0)	

†The calculations of accessible surface area and holes were done using the QUANTA Protein Design module. The maximum acceptable accessibility for hydrophobic (M, I, L, V, P, F) side-chains was 0.3. Conformers of water-exposed side-chains in experimental structures were adjusted to maximize the number of their hydrogen bonds. Nonpolar surfaces, ΔS, were calculated for all carbon atoms excluding those in peptide, carboxyl, and guanidine groups. The equivalent transfer free energy differences were estimated using solvation constant 0.02 kcal/mol/A^2.

[a]The larger holes in T0046 model are present because of incomplete refinement.

the distance geometry calculations.[6] Most probably, the simple modeling procedure described above can generate a multitude of very different folds that, judged by these criteria, may be better than the native structure. The fold multiplicity arises because each step of the modeling procedure includes multiple choices: there are different

ways to choose the regular secondary structures, to combine the β-strands into β-sheets, and to build different three-dimensional (3D) folds. It appears that the incorrectly predicted amphiphilic α-helices and β-sheets can usually be assembled to bury nearly all of their nonpolar surfaces and simultaneously provide an outside orientation of their polar side-chains. Moreover, conformers of side-chains can probably be adjusted without hindrances and holes in almost any model. Therefore, the docking and refinement steps of the modeling procedure cannot eliminate the incorrectly chosen secondary structures.

It must be emphasized that it is possible to distinguish the misfolded structures from correct ones: for example, all experimental structures actually have smaller water-accessible nonpolar surfaces than the corresponding models (Table I). However, the differences in nonpolar surfaces are visible only at the atomic level from comparison of precisely calculated structures.

The number of possible folds is expected to increase with methods that are based only on some criteria of "good" structures, such as lattice simulations that optimize only the number of nonpolar contacts, or methods that do not explicitly use secondary structure. Apparently, any ab initio prediction methods that operate with approximate potential energy functions, or use crude geometric approximations of the peptide chain, will produce many different nonnative folds, which can be separated from the native one only by constructing full-atomic models with potentials that precisely account for the small energies of all specific interactions in the protein structure. It would be easier to calculate protein structure if the folds were determined, for example, by the hydrophobicity pattern, while the precise structure was dependent on weaker interactions. However, the results obtained here indicate that this is not the case.

This situation means that ab initio modeling methods must automatically generate the set of alternative folds and at the same time restrict this set of folds as much as possible by using sufficiently precise potential energy functions. For example, the secondary structure-based approach applied here must be implemented in a more quantitative manner. This can be done by using recent mutagenesis and model peptide studies to quantify specific interactions in α-helices and β-sheet. The corresponding theoretical model would quantitatively describe formation and competition of different secondary and supersecondary structures in a hypothetical hydrophobically collapsed state during protein folding, instead of any qualitative considerations (such as steps (1) to (3) described in Methods and Results). Then, all specific interactions within the secondary structures would be immediately taken into account to reduce, from the beginning of the modeling, the number of alternative folds. The first part of this approach that describes formation of α-helices in peptides and proteins has been recently designed and tested.[15] A similar model of β-sheet formation (currently under development in our laboratory) considers all possible β-sheet topologies, and represents free energy of a β-sheet as a combination of backbone interactions, secondary structure propensities, pairwise interactions between side chains, specific β-turn and β-bulge terms, and transfer energy of the β-sheet from water to the protein "droplet" interface. Design of such a quantitative model could be important, judging from results obtained for CASP3 targets. Indeed, two major problems were:

1. Determination of correct β-sheet topology and structure (for Sm D3 domain and cyanovirin-N).
2. The choice between amphiphilic α-helices and β-sheets (for N-terminal β-strand of Sm D3, and several regions in γ-adaptin ear domain and HdeA).

Both problems could be solved using the quantitative model of β-sheet formation, even without calculations of 3D structures. However, the thermodynamic model of secondary structure cannot resolve completely the "fold multiplicity" problem. The next required step is an automatic generation of possible folds in three dimensions that could be organized as a stepwise growth of protein core, starting from alternative initiating sets of most strongly interacting α-helices and β-sheets identified by the quantitative model of secondary structure.

REFERENCES

1. Dill KA. Dominant forces in protein folding. Biochemistry 1990; 29:7133–7155.
2. Yue K, Dill K. Folding proteins with a simple energy function and extensive conformational searching. Protein Sci 1996;5:254–261.
3. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.
4. Richards FM, Lim WA. An analysis of packing in the protein folding problem. Q Rev Biophys 1994;26:423–498.
5. Güntert P, Braun W, Wüthrich K. Efficient computation of three-dimensional protein structure in solution from NMR data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J Mol Biol 1991;217:517–530.
6. Pogozheva ID, Lomize AL, Mosberg HI. The transmembrane 7-α-bundle of rhodopsin: distance geometry calculation with hydrogen bonding constraints. Biophys J 1997;72:1963–1985.
7. Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K. Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. Cell 1999;96:375–387.
8. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of β-sheet barrels in proteins. II. The observed structures. J Mol Biol 1994;236:1382–1400.
9. Bewley CA, Gustafson KR, Boyd MR, Covell DG, Bax A, Clore GM. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. Nature Struct Biol 1998;5:571–578.
10. Richardson JS, Richardson DC. The *de novo* design of protein structures. Trends Biochem Sci 1989;14:304–309,.
11. Yang F, Gustafson KR, Boyd MR, Wlodawer A. Crystal structure of Escherichia coli HdeA. Nature Struct Biol 1998;5:763–764
12. Sibanda BL, Blundell TL, Thornton JM. Conformations of β-hairpins in protein structures. J Mol Biol 1989;206:759–777.
13. deAlba E, Jiménez MA, Rico M, Nieto J. Conformational investigation of designed short linear peptides able to fold into β-hairpin structures in aqueous solution. Folding Design 1996;1:133–144,.
14. Richardson JS, Richardson DC. Principles and patterns of protein conformation. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York and London: Plenum Press; 1989. p 2–97.
15. Lomize AL, Mosberg HI. Thermodynamic model of secondary structure for α-helical peptides and proteins. Biopolymers 1997;42: 239–269.