RESEARCH PAPER

PCCP
www.rsc.org/pccp

A computational pathway for bracketing native-like structures for small alpha helical globular proteins†

Pooja Narang, Kumkum Bhushan, Surojit Bose and B. Jayaram*

Department of Chemistry, Indian Institute of Technology, Hauz Khas, New Delhi, 110016, India. E-mail: bjayaram@chemistry.iitd.ac.in; Fax: +91-11-2658 2037; Tel: +91-11-2659 1505, 91-11-2659 6786

Received 14th February 2005, Accepted 5th April 2005
First published as an Advance Article on the web 29th April 2005

Impressive advances in the applications of bioinformatics for protein structure prediction coupled with growing structural databases on one hand and the insurmountable time-scale problem with *ab initio* computational methods on the other continue to raise doubts whether a computational solution to the protein folding problem—categorized as an NP-hard problem—is within reach in the near future. Combining some specially designed biophysical filters and vector algebra tools with *ab initio* methods, we present here a promising computational pathway for bracketing native-like structures of small alpha helical globular proteins departing from secondary structural information. The automated protocol is initiated by generating multiple structures around the loops between secondary structural elements. A set of knowledge-based biophysical filters namely persistence length and radius of gyration, developed and calibrated on approximately 1000 globular proteins, is introduced to screen the trial structures to filter out improbable candidates for the native and reduce the size of the library of probable structures. The ensemble so generated encompasses a few structures with native-like topology. Monte Carlo optimizations of the loop dihedrals are then carried out to remove steric clashes. The resultant structures are energy minimized and ranked according to a scoring function tested previously on a series of decoy sets *vis-à-vis* their corresponding natives. We find that the 100 lowest energy structures culled from the ensemble of energy optimized trial structures comprise at least a few to within 3–5 Å of the native. Thus the formidable "needle in a haystack" problem is narrowed down to finding an optimal solution amongst a computationally tractable number of alternatives. Encouraging results obtained on twelve small alpha helical globular proteins with the above outlined pathway are presented and discussed.

1 Introduction

The growing genome knowledge[1,2] and the pressing necessity for the discovery of new drug targets for life threatening diseases bring the protein folding problem[3–29] to center stage with the expectation of an early solution either *via* theory or experiment or both.[30,31] Investigations on protein folding pathways[32–35] have helped in redesigning proteins[36–38] and in understanding how mis-folding can lead to disorders such as Alzheimer's and Parkinson's diseases.[39–46] Protein structure prediction techniques, range from 2D lattice models to atomic level *ab initio* methods to comparative modeling. The ultimate scientific challenge addressed by all these studies is to understand at a molecular level as to how proteins fold to their unique 3D conformations, starting from sequence information alone. Technology beckons to the development of predictive tools for biocatalyst design and nanobiomachines in some of the most significant impending contributions of protein folding to mankind.

The current computational strategies for protein structure prediction are either database dependent such as comparative modeling which includes homology modeling[47] and fold recognition techniques[48] or *ab initio*[49] in nature. Comparative modeling approaches aim to propose plausible structures utilizing *a priori* sequence and structural knowledge of related proteins. With large amounts of genome and proteome data accumulating *via* sequencing projects, comparative modeling has become the method of choice to characterize sequences where related representatives of a family exist.[50–55] *Ab initio* protein folding endeavors, on the other hand start with the amino acid sequence information and attempt to attain the fully folded native form consistent with the global free energy minimum.[56–60] Simulating a multi-dimensional surface and locating the global minimum on it, is a task common to a wide array of optimization problems in physics, chemistry, biology, atmospheric sciences and economics among others. The main focus in *ab initio* methodologies applied to proteins had been on understanding the protein structure and dynamics[21,57,61–73] and the various pathways/mechanisms involved in the folding of proteins.[18,32,35,74–84]

Structure prediction *via ab initio* attempts broadly evolved along two lines. The first strategy involves generating a multitude of possible structures at the atomic level sampling the vast configurational space either stochastically or deterministically and ranking all these conformations according to free energy and locating the global minimum, which corresponds to the native.[56–60] Conformational searches can be performed using systematic approaches[85] or random search methods in Cartesian[86,87] or dihedrals space,[88,89] genetic algorithms,[90] orthogonal latin square[91,92] and various other methods.[93–107] Empirical and free energy functions in this regard help in locating the most preferred conformation under the prescribed external constraints. Structure prediction attempts using distance geometry approach[108,109] incorporating experimental information in the form of inter-residue distances in minimization procedures,[110,111] metric matrix distance geometry[112] for generating native-like folds, and scoring functions such as residue pair potentials, hydrophobicity function[113] have shown to select

† Electronic supplementary information (ESI) available: PDB codes of proteins used in the study. See http://www.rsc.org/suppdata/cp/b5/b502226f/

DOI: 10.1039/b502226f

native-like structures. Hierarchical approaches,[114,115] build up procedures[93,116–119] and simulated annealing techniques[120–123] as well as Monte Carlo[124,125] methods have been used extensively for conformational sampling of proteins. A few other initiatives in this area are also reported in the literature.[126,127]

The second strategy for *ab initio* protein structure determination involves simulating the folding pathway of the polypeptide chain by solving Newton's equation of motion. Early simulation techniques evolved from reduced representations of protein molecule,[128–133] along with continuum models for the solvent.[134] Increase in computational power and efficiency has made possible explicit all-atom treatment of protein with solvent effects taken implicitly or explicitly.[135–137] It is surmised that the last stages in protein folding involve side chain ordering into well defined and closely packed structures and that molecular dynamics simulations with explicit solvent could be used as an end game. *Ab initio* structure prediction of small proteins using database driven Rosetta server[138] for generating 3D constructs of polypeptide chains and explicit solvent MD simulations on selected structures vividly demonstrates the successful marriage of bioinformatics with *ab initio* methods. Although, limited only by the computational expediencies for millisecond to second long molecular dynamics simulations, *ab initio* methods are the choice for a rigorous solution to the problem, the need for faster and smarter methods for structure prediction cannot be overemphasized. Since protein structure prediction is a problem largely dealing with main chain and side chain dihedrals, prediction accuracies critically depend upon generation of proper rotamer orientations and quality of the force-fields. Attaining atomic level accuracy with correct native-like topology is a challenge till date.

Of particular interest in this regard are methods based on a combination of bioinformatics tools and conformational sampling procedures for prediction of native-like topologies, molecular dynamics simulations for side-chain packing and refinement, and free energy analyses for accurate quantitative estimates, which have the potential for automation and implementation in cluster/grid computing modes. We have adopted a combination approach *i.e.* generate a linear chain from sequence, pre-build the secondary structures, and reduce the search space of the tertiary structures *via* usage of knowledge-based biophysical filters. In the design of these biophysical filters too, either one employs the available data directly to set bounds for rejecting structures, or one may convert the existing structural data to a well established physico-chemical model and extract limits for acceptance/rejection of structures. We have chosen the latter strategy. Since the overall goal of the protein structure prediction attempts is to arrive at the tertiary structure starting from the amino acid sequence, the protein structure prediction problem is subdivided for computational convenience into secondary structure prediction from the sequence, overall tertiary fold prediction from the secondary structure and finally side chain packing. At this point, our emphasis is to go from the secondary structure to the fully folded tertiary structure or at least native-like decoys. In this article, we propose a computational pathway (Fig. 1) from secondary structure to arrive at tertiary fold for alpha helical globular proteins comprising less than 100 amino acids. Results obtained on twelve proteins investigated to elicit a proof of concept are extremely encouraging.

2 Methodology

The overall strategy consists of nine steps. The first two steps involve the formation of a representative structure for the polypeptide chain from amino acid sequence with the secondary elements in place. The third step involves generation of a large number ($\sim 10^5$ to 10^6) of trial structures with a systematic sampling of the conformational space of loop dihedrals. These structures are screened through persistence length and radius of gyration filters, developed for the purpose of reducing the number of improbable candidates in the fourth step. At this stage, the ensemble of trial structures is brought down to a manageable number. The resultant structures are refined by a Monte Carlo sampling in dihedral space to remove steric clashes and overlaps involving atoms of main chain and side chains in the fifth step. In the next two steps the structures are energy minimized and ranked using an empirical scoring function and the lowest 100 energy structures are selected. It is noticed that, in all the systems studied, native-like structures to within an RMSD of 3–5 Å of the native are bracketed within these 100 structures. Metropolis Monte Carlo simulations conducted on each of the 100 structures in the final step are noticed to improve the energy ranking of the native-like structures further. Each of these steps is explained in detail below.

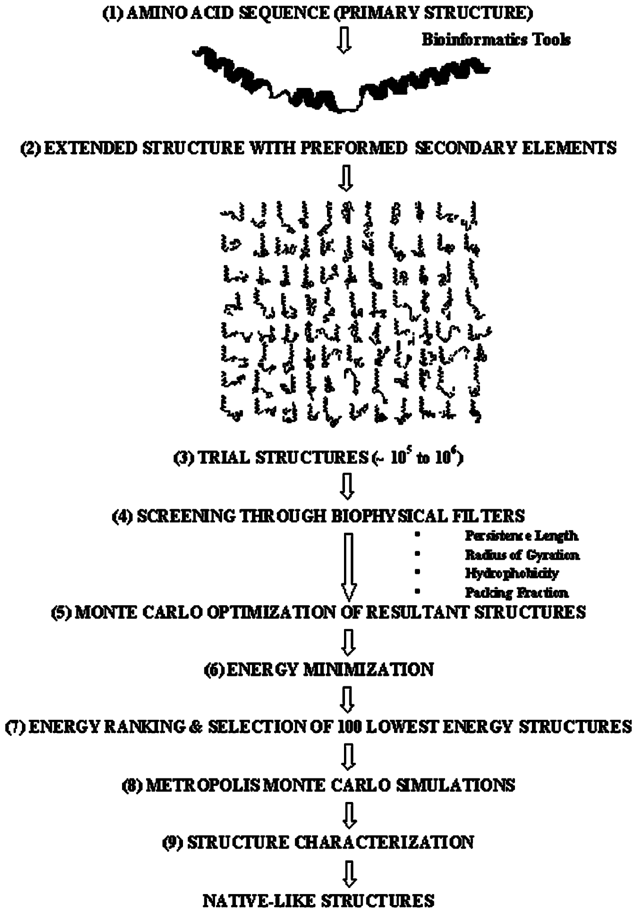
(1) AMINO ACID SEQUENCE (PRIMARY STRUCTURE)
Bioinformatics Tools
(2) EXTENDED STRUCTURE WITH PREFORMED SECONDARY ELEMENTS
(3) TRIAL STRUCTURES (~ 10⁵ to 10⁶)
(4) SCREENING THROUGH BIOPHYSICAL FILTERS
Persistence Length
Radius of Gyration
Hydrophobicity
Packing Fraction
(5) MONTE CARLO OPTIMIZATION OF RESULTANT STRUCTURES
(6) ENERGY MINIMIZATION
(7) ENERGY RANKING & SELECTION OF 100 LOWEST ENERGY STRUCTURES
(8) METROPOLIS MONTE CARLO SIMULATIONS
(9) STRUCTURE CHARACTERIZATION
NATIVE-LIKE STRUCTURES

Fig. 1 A computational pathway for bracketing native-like structures of small alpha helical globular proteins.

(1) From sequence to generation of 3D representation of a linear polypeptide chain

Based on the sequence information provided for a given protein, templates of amino acids[139] are joined with each other to generate an extended polypeptide chain. The primary goal here is to convert sequence information into a 3D structure with a specification of the Cartesian coordinates of all the atoms of the main chain and side chains.

(2) From linear chain to structure with preformed secondary structures

In the proposed pathway, we start from the secondary structure information from the native PDB structure, as the experimental structure in each case is available. Bioinformatics tools can be utilized where experimental information is not available. The linear segments of polypeptide chain are converted to the corresponding secondary structures using database dihedrals. The main-chain Ramachandran angle (Φ,Ψ) analysis and the side-chain angle (χ_1, χ_2, χ_3) analysis for each of the twenty amino acids carried out for this purpose are given in Tables 1 and 2. This analysis was done on $\sim$1000 monomeric globular proteins taken from the PDB,[140] for helix, sheet and loop regions. The criteria for obtaining the dataset include absence of metal ions and disulfide linkages and a resolution below 2.0 Å. The proteins are non-homologous and selected from various functional classes including hydrolases (286), isomerases (31), lyases (44), ligases (16), oxidoreductases (108), transferases (122) and non-enzymes (406), the number in parenthesis indicates the number of proteins in each class. According to structural classification, the proteins have been divided into three classes namely alpha helical (96), beta sheet containing (22) and mixed type (895). The PDB codes of all the proteins are provided as supplementary information. The averaged main-chain dihedral values along with standard deviations are given in Table 1(a). It is observed that the average values of dihedrals in the helix region do not differ much from those depicted in the original Ramachandran plots[141] *i.e.* −57° for Φ and −47° for Ψ for the right handed α-helix. Comparison of Φ and Ψ values for the sheet region shows that the values obtained from the database are more widely spread, which was expected because these values include both parallel as well as antiparallel β-sheets. The Φ and Ψ values for the loop regions showed a large standard deviation. This is expected because of the flexible nature of the loop regions. Average dihedral values for the loop regions are therefore less reliable. So frequency distributions for the Φ and Ψ for this region were generated and most frequently occurring values were used for the loop region (given in Table 1(b). Though many rotamer libraries[142–150] are available in the public domain, we have developed a backbone independent but secondary structure dependent library for side chain dihedrals with a large dataset of proteins and resolution less than 2 Å. For both the helix and loop regions most frequently occurring values were incorporated (Table 2(a) and 2(b)).

(3) Trial structure generation

In the present study, we have opted for a systematic exploration of the conformational space of each loop dihedral following the grid method for trial structure generation. In our preliminary investigations, the structures were made to span the orthogonal space, by alignment of one secondary structural element on a particular axis as a reference and the next secondary element on the other Cartesian axes or their bisector. About 26 orientations are possible for the second helix in relation to the first helix without invoking symmetry. The algorithm generates different possible conformations by positioning the secondary elements in 3D space starting with the first and adding each secondary element in the reference frame of the previous one step-by-step. Total number of structures generated with this method is $(26)^{(n-1)}$, where n is the number of helices. Considering that the secondary structural elements can occur with a reversal of polarity in the known proteins, a total of $52^{(n-1)}$ structures are generated. It was found with this method of trial structure generation that a large strain was introduced in the loop regions during the placement of secondary structural elements along the Cartesian axis. In a helix-turn-helix protein (1FLX), out of 52 structures generated with this method, 17 structures had an RMSD $<$10 Å, with the lowest RMSD at 4.2 Å. We explored a second option of utilizing the dihedral space for trial structure generation in which the main chain dihedrals (Φ and Ψ) were incremented from 0 to 360° in steps of 90° each. Thus, the conformations generated corresponded to 0, 90, 180 and 270° for each dihedral selected for rotation. We selected two amino acids from each loop region, adding upto four rotatable dihedrals per loop. A total of $(256)^{(n-1)}$ structures were generated with

Table 1 (a) Average values for Ramachandran angles (Φ,Ψ) for the helix, sheet and loop region obtained from a database analysis of $\sim$1000 globular proteins. (b) The most probable values for Ramachandran angles (Φ,Ψ) for the loop region obtained from a database analysis of $\sim$1000 globular proteins (listed in order of preponderance for each dihedral)

	(a)												(b)	
	Helix				Sheet				Loop					
	Φ (Phi)		Ψ (Psi)		Φ (Phi)		Ψ (Psi)		Φ (Phi)		Ψ (Psi)			
Amino acid	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Φ (Phi)	Ψ (Psi)
ALA	−63.6	7.8	−39.3	9.3	−121.3	33.6	132.9	45.3	−77.1	41.6	62.1	86.0	−60, −140, 0, 60	140, −20
ARG	−64.4	9.7	−39.5	10.2	−116.7	28.7	129.7	39.1	−83.5	47.9	63.5	82.6	−60, −120, 60, 0	160, −20
ASN	−66.6	14.2	−36.0	13.4	−107.0	37.7	115.6	60.3	−69.1	67.9	52.3	77.0	−80, 60, 0	20, 40, 120, −180
ASP	−64.8	9.9	−38.6	11.7	−105.9	35.6	107.4	69.3	−79.2	50.6	47.6	84.0	−80, 60, 0	20, 120, −20, 160, −180
CYS	−66.9	12.9	−39.2	10.8	−119.4	27.3	128.4	54.8	−88.5	50.4	66.3	87.4	−60, −120, 60, 0	140, 0, −180
GLN	−65.3	9.3	−38.9	9.4	−114.6	29.2	127.6	41.0	−79.9	50.3	60.7	81.0	−60, 60, 0	160, 0, −180
GLU	−64.8	9.1	−39.2	9.1	−113.8	30.7	127.3	42.1	−78.2	44.0	53.3	83.9	−60, 60, 0	140, −20
GLY	−60.9	20.9	−40.2	21.6	−44.4	127.6	36.0	149.2	33.6	88.5	0.3	101.4	100, −60, 0	20, −180, 180
HIS	−67.5	15.0	−38.3	14.4	−118.5	30.9	129.0	44.0	−83.8	56.5	66.1	80.0	−60, −120, 60, 0	160, 60, 0, −180
ILE	−65.1	9.2	−42.1	8.4	−114.5	19.0	123.8	34.4	−93.5	31.1	79.5	76.4	−100, −60, 0	140, −40
LEU	−65.3	9.3	−39.2	10.3	−108.8	22.8	124.0	38.8	−84.0	35.5	69.0	81.8	−60, 0	160, −20
LYS	−64.7	10.5	−39.4	10.0	−113.0	29.0	128.4	41.4	−78.7	49.3	57.3	83.1	−60, 60, 0	160, −20, 60, −180
MET	−65.6	8.9	−38.4	10.7	−119.0	25.8	131.1	38.0	−85.9	41.8	67.4	82.4	−60, 0, −120	160, −10, 80
PHE	−66.0	12.6	−40.5	13.0	−119.5	25.8	133.4	36.5	−91.2	44.4	74.4	78.9	−60, −120, 70, 0	160, 0, −20, 80
PRO	−58.0	6.4	−36.4	11.2	−70.7	10.4	133.2	39.2	−65.2	12.4	83.3	84.8	−60, 0	160, −20, 60, −180
SER	−65.8	12.2	−36.5	13.8	−122.5	29.4	128.9	64.4	−86.4	46.4	67.4	91.5	−60, −120, 0, 60	160, 0, −180
THR	−67.4	13.7	−39.3	12.8	−118.3	22.8	123.7	66.3	−96.1	35.3	68.0	92.5	−80, −120, 0	180, 140, 0, −180
TRP	−64.7	12.6	−40.1	12.1	−118.0	27.6	129.4	50.4	−87.5	42.1	66.7	81.1	−60, −120, 60	160, 0, 80, −180
TYR	−65.9	14.2	−40.3	13.4	−121.0	25.8	136.7	30.4	−89.0	47.5	71.8	80.3	−60, −120, 60, 0	160, 0, 80, −180
VAL	−65.5	9.8	−41.7	9.8	−116.9	19.3	125.2	37.8	−95.6	33.0	82.5	79.6	−120, −60, 0	140, −30, −180

Table 2 (a) The most probable values for the side chain dihedrals in the helix region obtained from a database analysis of ~1000 globular proteins (listed in order of preponderance for each dihedral). (b) The most probable values for the side chain dihedrals in the loop regions obtained from a database analysis of ~1000 globular proteins (listed in order of preponderance for each dihedral)

(a)

Amino acid	χ_1 (Chi1)	χ_2 (Chi2)	χ_3 (Chi3)
ARG	−60, −180, 180, −40, −80, 80	180, −180, −60	−180, 180, −60, 80
ASN	−60, −180, −80, −40, 180, 80, 60	−40, 40, 120, −180	
ASP	−60, −180, −80, 180, −40, 60, 80	0, 180, −180	
CYS	−60, −40, −180, 180, 80, 60, −80		
GLN	−60, −180, −40, −80, 180, 80, 60	−180, 180, −60, 80	−40, 60, 20, 140, −180
GLU	−60, −180, 180, −40, −80, 80, 60	−180, 180, −60, 80	0, 60, −60, 180, −180
HIS	−60, −180, 180, −40, −80, 80, 60	−60, 90, 180, −180	
ILE	−60, −40, −180, 80, 60, −80	−140, 160	
LEU	−60, 180, −180, −40, −80	180, −180, 60	
LYS	−60, −180, 180, −40, −80, 80, 60	−180, 180, −60, 60	180, −180, 60, −60
MET	−60, −180, 180, −40, −80, 80	−180, 180, −60, 60	−80, 100, 160, −160, 0
PHE	−60, 180, −180, −80, −60, 80	60, −60	
PRO	−20, 40, 0, 20	−40, 40	
SER	80, −60, 60, 180, −40, −180, 100		
THR	180, −180, −40, −60, 80, 60		
TRP	−60, −180, 180, −80, −40, 80, 60	100, −100, 0	
TYR	−60, 180, −180, −40, −80, 80, 60	−60, 60	
VAL	180, −180, −60, −40, 80		

(b)

Amino acid	χ_1 (Chi1)	χ_2 (Chi2)	χ_3 (Chi3)
ARG	−60, −40, −180, 80, 0	180, −180, −60	−180, 180, −60, 80
ASN	−60, −180, −40, 80, 0, 180	−40, 40, 120, −180	
ASP	−60, −180, 80, 60, −40, 180, 0	0, 180, −180	
CYS	−60, −40, −180, 80, 180, 0		
GLN	−60, −40, −180, 0, 80, 180	−180, 180, −60, 80	−40, 60, 20, 140, −180
GLU	−60, −40, −180, 180, 0, 80	−180, 180, −60, 80	0, 60, −60, 180, −180
HIS	−60, −40, −180, 80, 60, 0, 180	−60, 90, 180, −180	
ILE	−60, −40, 80, 60, −180, 0	−140, 160	
LEU	−60, −40, −180, 180, 0	180, −180, 60	
LYS	−60, −40, −180, 0, 180, 80	−180, 180, −60, 60	180, −180, 60, −60
MET	−60, −40, −180, 180, 80, 0	−180, 180, −60, 60	−80, 100, 160, −160, 0
PHE	−60, −40, −180, 180, 0, 60, 80	60, −60	
PRO	40, −20, 0, 20,	−40, 40	
SER	80, −60, 180, −40, −180, 0		
THR	80, 60, −40, −60, 0, −180		
TRP	−60, −40, −180, 80, 60, 180, 0	100, −100, 0	
TYR	−60, −40, −180, 180, 80, 60, 0	−60, 60	
VAL	180, −180, −60, −40, 0, 80		

this protocol, where n is the number of secondary structural elements. For the helix-turn-helix case, out of 256 structures generated, 83 had an RMSD <10 Å, with the lowest RMSD at 3.47 Å. The third protocol which we attempted takes recourse to the commonly observed and energetically preferred dihedrals in conformational analyses *viz*. *gauche*(+), *gauche*(−) and *trans*. Eclipsed conformation is also considered as a possibility. The middle two amino acids from each loop are selected and their (Φ,Ψ) dihedrals (four dihedrals per loop) are varied. Rest of the dihedrals of each loop region remain untouched during this structure generation process. Here also, a total of $(256)^{(n-1)}$ structures are generated for each system. For the helix-turn-helix case, out of the 256 structures generated, 96 had an RMSD <10 Å, with the lowest RMSD at 3.41 Å. From the number of structures generated and the RMSDs computed, it is evident that sampling in the dihedral space is comparatively more expensive, but more reasonable in terms of the quality of the structures generated. According to Go and Scheraga at least six degrees of freedom are required for the complete loop closure.[151] We have considered only four degrees of freedom by varying four dihedrals in the loop region, during trial structure generation for sampling the three dimensional space, but include the remaining unutilized loop dihedrals to complete the sampling *via* a Metropolis Monte Carlo (Boltzmann) in subsequent stages of structure refinement. This was done to ensure that the proposed methodology samples, at least in a coarse-grained manner, nearly all loop conformations.

(4) Biophysical filters

The trial structures generated are screened using biophysical filters, which have been developed on the basis of known physico-chemical properties of proteins. The first such property is named as *persistence length*, which is the maximum length of the uninterrupted polypeptide chain persisting in a particular direction. In literature, persistence length is defined as the distance over which the direction of the polymer segment persists and has been used extensively to describe the rigidity of synthetic polymers as well as DNA.[152,153] The algorithm for persistence length calculates the distance between the N-terminal of the ith residue (i = 1, 2, …, n − 1; n is the residue number), and the C-terminal of (i + j)th residue (j = 1, 2…) consecutively, as long as the distance is greater than the previous one. The above process is repeated starting from the (i + 1)th secondary structural element and the distance calculated is compared with the distance obtained from the previous

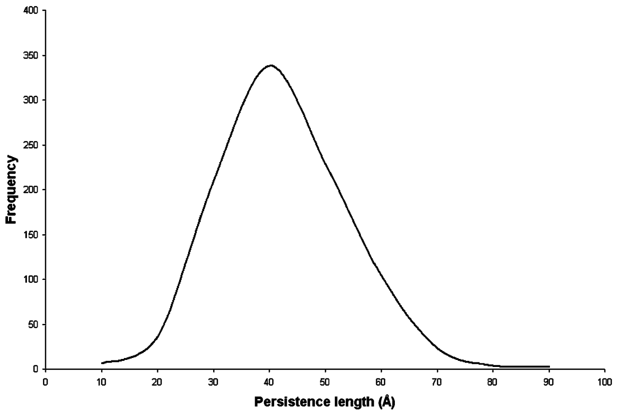

Fig. 2 Persistence length analysis on the dataset of ~1000 globular proteins.

calculation. This is continued till maximum distance of the polypeptide chain is obtained which is the persistence length for that protein in any given direction. Analysis of dataset of ~1000 globular proteins showed that the persistence length varies from 15 Å to 60 Å within the 95% confidence limits and averages around 40 Å (Fig. 2).

Another filter developed is the *radius of gyration*, which describes the overall spread of the molecule and is defined as the root mean square distance of the collection of atoms from their common center of gravity. Radius of gyration, used to describe polypeptide chains, was originally proposed 60 years ago by German chemist Kuhn,[154] where he emphasized that the shape of the random walk polymer is not spherically symmetric but is flexible having an isotropic end-to-end vector distribution. Flory[155] proposed that the growth of mean square end to end distance of the polymer is a function of the degree of polymerization. For an N amino acid long protein behaving as a random coil, radius of gyration, R_G, is directly proportional to the square root of N, suggesting that with the sequence information alone, for a given polymer, one can approximately predict the size. According to Flory, for large values of N, which is generally the case with proteins, a self avoiding random walk model rather than a simple random walk model is more appropriate. This is attributable to excluded volume effects[156] and results in a different scaling: R_G is directly proportional to $N^{3/5}$. The algorithm for radius of gyration first calculates the center of gravity of the whole protein molecule and then the root mean square distance of all the atoms from the center of gravity. Radius of gyration was calculated for the selected dataset of proteins and the values were plotted in Fig. 3 against $N^{3/5}$, where N is the number of amino acids. The R^2 value for the correlation is 0.86. Using least square fit method upper and lower limits for this filter were set up.

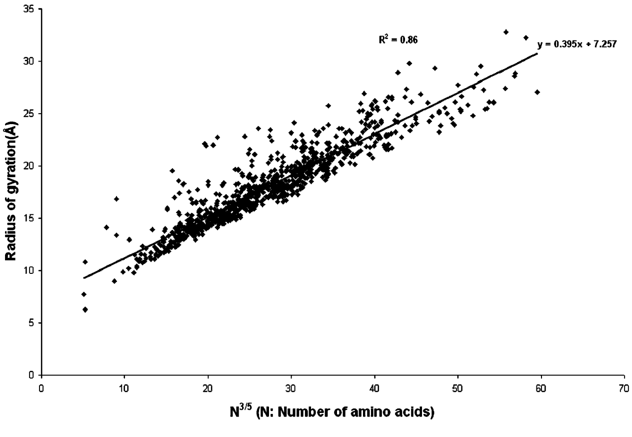

Fig. 3 Radius of gyration analysis on the dataset of ~1000 globular proteins.

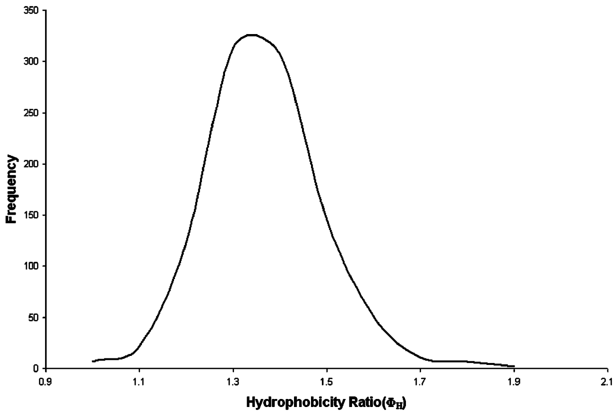

Fig. 4 Hydrophobicity ratio (Φ_H, eqn. (1)) analysis on the dataset of ~1000 globular proteins.

The third filter developed is based on *hydrophobicity*[157] which is now well documented to be a major governing force in the folding of proteins. Proteins fold in a way, such that, the hydrophobic amino acids occupy the core region and the hydrophilic amino acids are relatively more exposed on the surface. Various hydrophobicity scales have been defined[158,159] which are based on the tendency of amino acids to be found inside the protein core or on the surface and on the basis of physico-chemical properties of side chains of amino acids.[160,161] We have first calculated the accessible surface areas[162] of side chains from the fully extended tripeptides (GLY–X–GLY). These values are in good agreement with the literature values.[163] The accessible surface area of amino acid residues changes upon folding of the polypeptide chain and is related to the hydrophobicity of the amino acid side chain, which governs the preference of the amino acid for the surface or the core of the protein. After considering several alternatives to convert hydrophobicity into a computational filter, we have decided on quantifying the "non-polar in and polar out" property as a ratio. The hydrophobicity ratio, Φ_H is defined here as the ratio of loss in accessible surface area (ASA) per atom of non-polar atoms to the loss in accessible surface area (ASA) per atom of the polar atoms,

$$\Phi_H = \frac{\text{Loss in ASA per atom of non-polar atoms}}{\text{Loss in ASA per atom of polar atoms}} \qquad (1)$$

For ~1000 protein dataset, the ratio Φ_H was calculated and a graph between the hydrophobicity ratio and frequency was plotted (Fig. 4). The hydrophobicity ratio varies from 1.0 to 1.9 in conformity with expectations. We have selected the range from 1.2 to 1.6, which lies within the 95% confidence limits as a filter.

Proteins are closely packed structures. The formation of secondary and tertiary structures drives them towards achieving high packing densities.[164,165] Studies in this area indicate correlation between packing density, sequence conservation, and folding nucleation.[166] Folded proteins are known to exhibit packing fractions around 0.7.[167] We have devised a grid method to compute the *packing fractions* and generated a distribution for the selected dataset of ~1000 proteins mentioned above. The values for packing fraction average around 0.70 and lies between 0.6 and 0.8 within 95% confidence limits (Fig. 5).

Essentially the aforementioned biophysical filters are designed to restrict the sample space of trial structures and to limit the number of candidates for further energy based processing in search of the native.

(5) Clash removal

Structures selected by the biophysical filters may have close van der Waals contacts due to overlaps of some parts of the

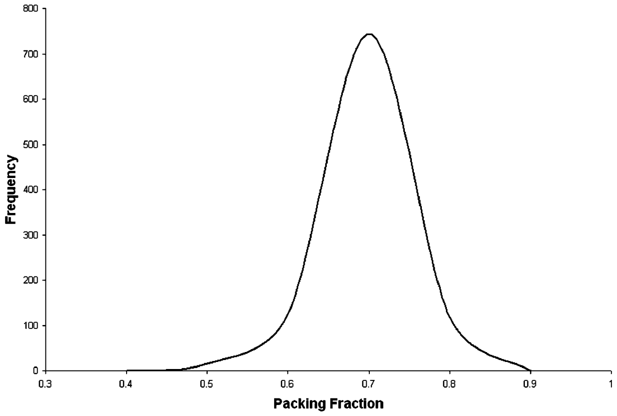

Fig. 5 Packing fraction analysis on the dataset of ~1000 globular proteins.

secondary structural elements or side chains during trial structure generation. For this, a distance based Monte Carlo sampling of the loop dihedrals is carried out. This is achieved by small optimizations of the main chain dihedrals in the loop regions and side chain dihedrals of the helix as well as loop regions during which small perturbations (2–10°) are made to ensure that the overall topology of the trial structure is not disturbed but the number of clashes is minimized.

(6) Energy minimization

At this stage, the trial structures are free from severe van der Waals contacts but local optimization of the side chains is required. A short minimization[139] of 150 steps (50 steepest descents + 100 conjugate gradients) in vacuum with distance dependent dielectric is performed on the trial structures.

(7) Energy ranking

According to the thermodynamic hypothesis, the native state of the protein corresponds to the global free energy minimum under normal physiological conditions.[3] A good scoring function, therefore, is the one that mimics a free energy function and discriminates between the native-like conformations and innumerable decoy conformations.

An all-atom based empirical scoring function developed in-house[168–171] and tested previously on publicly available decoys is employed to rank the optimized structures.[172] The empirical scoring function is expressed as a sum of three energy terms—electrostatic, van der Waals and hydrophobic.

$$E = \sum E_{\mathrm{el}} + E_{\mathrm{vdw}} + E_{\mathrm{hpb}} \qquad (2)$$

Here, E_{el} is the electrostatic contribution to the energy computed with a sigmoidal dielectric function, E_{vdw} is the van der Waals term, E_{hpb} is the hydrophobic contribution captured *via* Gurney approach[173,174] and the summation in eqn. (2) runs over all the atoms of the protein. Hydrogen bonding interactions are included in the electrostatics term as in the second-generation Amber force field. Based on energy ranking, 100 lowest energy structures were selected.

(8) Metropolis Monte Carlo simulations

The selected 100 lowest energy structures are further optimized by Metropolis Monte Carlo simulations (10 000 random moves for each structure at 300 K) in the space of side chain dihedrals of all the amino acids and main chain dihedrals of the amino acids in the loops.

(9) Characterization of candidate structures

The 100 candidate structures in each case are then characterized for their topological equivalence with the respective native structures using parameters such as percentage native contacts and RMSD in Cartesian space. The revised energy ranking of the structures is also obtained.

3 Calculations and results

We have chosen twelve small α-helical globular proteins, to test the performance and viability of the protocol outlined in Fig. 1, for generating native-like structures starting from sequence and secondary structural information. These proteins are free from metal ions, disulfide bridges as well as prosthetic groups and fold autonomously *in vitro*. The number of amino acids in these proteins ranges from 36 to 68 and the number of alpha helices ranges from 3 to 4. A master sheet of the results at the end of each step outlined in the Methods section is provided in Table 3.

Trial structure generation

Starting with the amino acid sequence and the secondary structural information from the native structure and dihedral angles obtained from the database analysis (Tables 1 and 2), initial structure containing the coordinates of all the atoms is built. Trial structures are then generated following sampling in the Φ, Ψ space of two residues in each loop, considering *gauche*(+), *gauche*(−) and *trans* as well as eclipsed conformations for each dihedral. For each of the helical proteins, the number of trial structures generated is equal to $256^{(n-1)}$, where n is the number of alpha helices. This is shown in Table 3, column (iv). Proteins with very short loops (<2 AAs) or very long loops (>6 AAs) and loops with proline require special treatment. This is taken up in the Discussion section.

Filtering the trial structures

The trial structures are passed through persistence length and radius of gyration filters to reduce the sample size of trial structures (Table 3, columns (v) and (vi)). We observe that the usage of persistence length filter followed by radius of gyration gives the maximum efficiency. In the case of small proteins where persistence length filter fails to reject any structure, a combination of the filters is efficient in screening trial structures. At this stage, the lowest RMSDs obtained with the end loops are in the range of 3.29–6.64 Å and without the end loops are in the range of 2.63–4.42 Å. These are given in columns (vii) and (viii), respectively of Table 3.

Clash removal and energy minimization

Generation of trial structures results in several close van der Waals contacts. A Monte Carlo procedure is used to remove the clashes. Close contacts occur among the main chain atoms, main chain and side chain atoms and among the side chain atoms. Small perturbations in the main chain dihedrals of the loop region and side chain dihedrals of the helix as well as loop regions, of the atoms involved in contact are attempted, until the structure is free from severe close contacts. Short energy minimization is performed thereafter on these structures with the protocol described in the Methodology section. The main purpose of these two steps is to relax the structure by removing any strain that may occur due to intramolecular clashes. The lowest RMSD ranges from 2.35 to 4.32 Å for all the twelve systems after this preliminary structure optimization and these are reported together with the corresponding energy rank in columns (ix) and (x) of Table 3.

Table 3 Results emerging from the proposed computational pathway for bracketing native-like structures for small alpha helical globular proteins

											Steps (7 and 8),[a] characterization of 100 lowest energy structures			
					Step (4),[a] no. of structures accepted				Steps (5 and 6),[a] MC optimization and energy minimization				Metropolis Monte Carlo simulations	
No.	PDB ID (i)	No. of residues (ii)	No. of helices (iii)	Step (3),[a] total no. of structures generated[b] (iv)	After persistence length (v)	After radius of gyration (vi)	Lowest RMSD (Å) (vii)	RMSD without end loops (Å) (viii)	Lowest RMSD (Å) (ix)	Rank (energy) (x)	Lowest RMSD (Å) (xi)	Rank (energy) (xii)	Lowest RMSD (Å) (xiii)	Rank (energy) (xiv)
1.	1VII	36	3	65 536	65 536	47 976	3.29	2.63	2.35	6958	2.85	3	2.88	1
2.	1DV0	45	3	65 536	65 536	28 606	4.23	3.72	3.78	7429	4.74	31	4.74	2
3.	1GVD	52	3	65 536	65 257	25 980	4.97	4.08	4.23	19 351	4.88	71	4.89	71
4.	1MBH	52	3	65 536	65 536	27 662	3.64	3.24	2.87	1774	4.66	72	4.63	24
5.	1GAB	53	3	65 536	65 483	18 941	3.89	3.37	3.16	838	4.01	50	4.08	25
6.	1IDY	54	3	65 536	65 536	18 953	4.85	2.97	2.38	2468	3.28	66	3.36	14
7.	1PRV	56	3	65 536	65 515	7545	5.56	3.40	2.7	727	4.23	52	3.87	2
8.	1HDD	57	3	65 536	61 427	16 523	4.08	3.29	2.46	1134	4.58	32	4.27	20
9.	1BDC	60	3	65 536	57 903	6800	6.64	4.42	4.12	5	4.12	5	4.21	2
10.	1HP8	68	3	65 536	48 171	5189	4.98	4.22	3.78	4610	3.89	90	4.20	41
11.	1BW6	56	4	262 144	254 975	44 872	5.99	4.13	4.32	6826	4.68	11	4.69	5
12.	2EZH	65	4	1048 576	1041 303	249 740	3.37	3.21	3.33	30 851	4.34	11	4.40	2

[a] Number of structures generated, accepted after biophysical filters and lowest RMSD together with energy ranking for each step of the protocol as shown in Fig. 1. [b] Total number of structures generated = $256^{(n-1)}$, where n is the number of helices. For systems 11 and 12, this number is less than expected because of the presence of PRO in one loop and a single amino acid (ASN) in another loop in the former, and presence of single amino acid (ASP) in one of the loops in the latter.

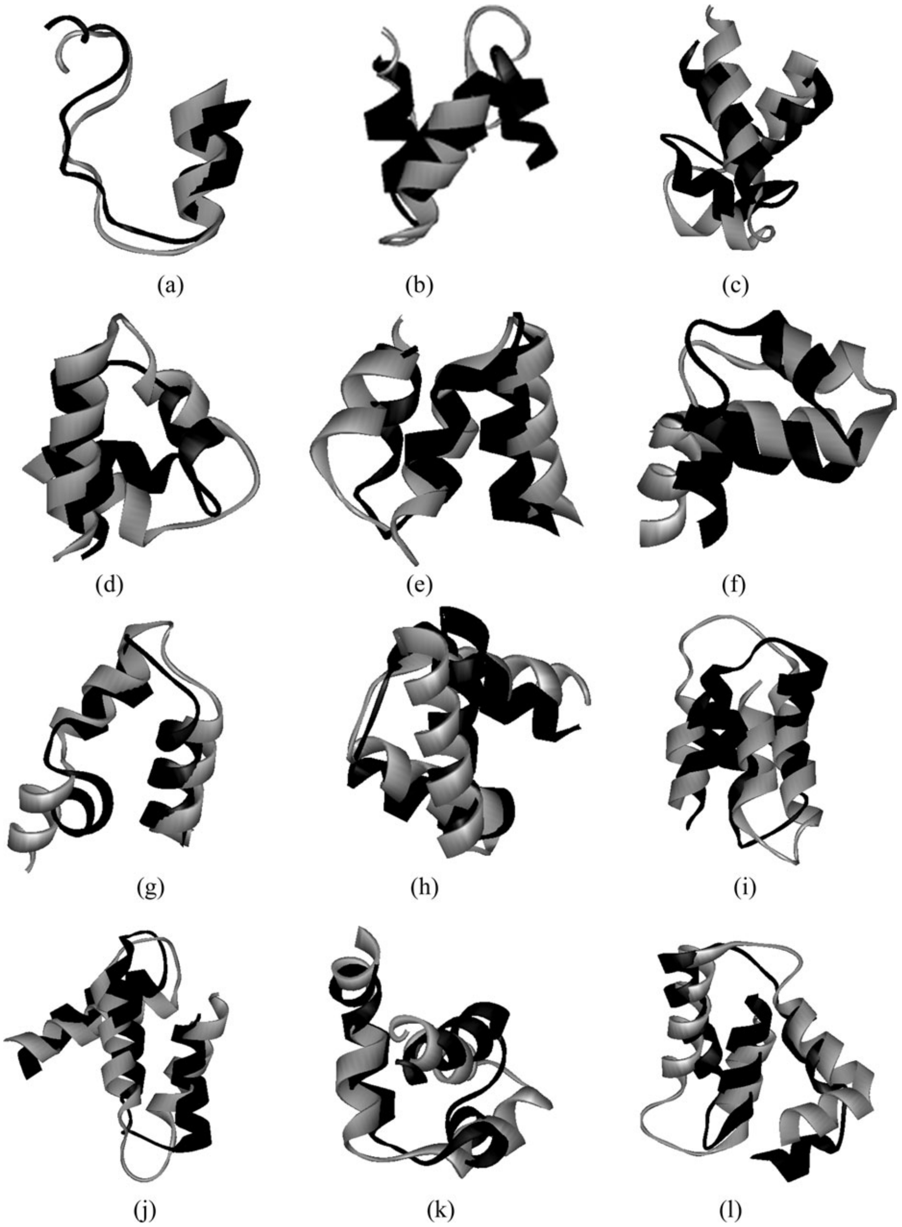

Fig. 6 The lowest RMSD structure emerging from the proposed pathway superimposed on the corresponding native structure for each of the twelve test proteins: (a) 1VII; (b) 1DV0; (c) 1GVD; (d) 1MBH; (e) 1GAB; (f) 1IDY; (g) 1PRV; (h) 1HDD; (i) 1BDC; (j) 1HP8; (k) 1BW6; (l) 2EZH. Native is represented in a darker shade and the predicted native-like structure in a lighter shade.

Energy scans using empirical scoring function

Energy analysis of these structures is carried out with the scoring function described in the Methodology section. The native structure was found to be the most stable energetically with this scoring function in all the systems. We selected 100 lowest energy structures for each system after energy analysis. The lowest RMSD together with the corresponding energy rank obtained amongst these 100 structures at this stage is given in Table 3 in columns (xi) and (xii). The selected 100 lowest energy structures have among them a few structures with native-like topology and RMSDs from the native in the range of 2.85 to 4.88 Å in all the twelve systems studied.

Monte Carlo optimization

The selected 100 structures are further optimized with a multiple move Metropolis Monte Carlo (Boltzmann) simulation in the dihedral space using the empirical scoring function. As the initial structures are generated by a change in only four dihedrals of the loop region, during the Metropolis Monte Carlo simulation we allow all the main chain dihedrals in the loops to vary, as also the side chain dihedrals to instigate better packing. In all the systems studied, even though we do not find a large variation in the RMSD of the structures after Monte Carlo optimization, we observe that the energies of the structures improve (columns (xiii) and (xiv) in Table 3). The superimposed lowest RMSD structures with their respective native structures are shown in Fig. 6.

Characterization of the selected structures

An assessment of the generated structures with RMSD in Cartesian space was carried out with respect to the native at each step of the pathway (Table 3, columns (vii), (viii), (ix), (xi), (xiii) and Table 4, column (ii)). In addition, percentage native contacts and RMSD in dihedral space were also calculated for the candidate structures after Metropolis Monte Carlo optimization as reported in Table 4, columns (iii) and (iv), respectively. Results in Table 4 indicate that side chain optimization has to be addressed in further studies towards the native.

Thus, starting from sequence and secondary structural information of the protein molecule, the methodology is able to

Table 4 Characterization of the predicted native-like structures from the pathway for each of the twelve proteins

No.	PDB ID (i)	Lowest RMSD after Metropolis Monte Carlo simulations (Å) (ii)	Native contacts (%) (iii)	Dihedral RMSD (°) (iv)
1	1VII	2.88	46.81	55.17
2	1DV0	4.74	26.56	75.58
3	1GVD	4.89	30.61	59.51
4	1MBH	4.63	36.84	49.47
5	1GAB	4.08	42.35	63.03
6	1IDY	3.36	48.10	51.63
7	1PRV	3.87	46.75	60.92
8	1HDD	4.27	56.38	37.02
9	1BDC	4.21	41.76	56.78
10	1HP8	4.20	38.10	61.60
11	1BW6	4.69	38.30	59.24
12	2EZH	4.40	49.52	44.35

generate an ensemble of structures by sampling the dihedral space around the loop regions. The biophysical filters reduce the sample size. After energy optimization, the scoring function selects 100 candidates with a few structures bracketing native to within 3–5 Å for each of the protein sequences investigated. The protocol is not a simulation of the folding pathway but an integrated computational suite to generate native-like decoys for small alpha helical globular proteins.

4 Discussion

We present here a computational protocol for bracketing native-like structures from sequence and secondary structural information and illustrate the methodology on twelve small alpha helical globular proteins. Native-like structures are seen to be bracketed to within 3–5 Å of the native in the 100 lowest energy structures in each case without exception. A critical assessment of each step is presented below. Sampling the dihedral space in a near complete manner to generate trial structures is not new but has become conceivable recently due to improved storage capacity and speed of current day computers. Biophysical filters utilizing the known physico-chemical properties of proteins have shown to reduce the number of candidates to a manageable number. When Persistence length is correlated with diameter of proteins for the dataset of proteins, the correlation is not high ($R^2 = 0.45$). On the other hand, when radius of gyration is correlated with the diameter of proteins, the correlation is good ($R^2 = 0.95$). Radius of gyration, therefore, is an indirect measure of the approximate radius of the protein. These two filters, which are based on different biophysical properties of protein, are nearly independent of each other and their usage as two different filters appears justified. The filters designed are in a sense complementary in eliminating wrong candidates which are either too extended or too compact. Due to the in-built characteristics of persistence length filter, it is able to reject extended structures very effectively. Radius of gyration on the other hand has proven to be more efficient in setting tight lower bounds as clear from Fig. 3. The empirical energy function utilized has been shown earlier[172] to separate the native from the decoys in a large number of decoy sets. Here the function is able to bracket native-like structures in the hundred lowest energy structures. The present protocol considers all atom representation of the protein molecule, from the structure generation step itself, unlike the reduced representations[175] and lattice models.[176] The results present a proof of concept of the protocol. It may also be mentioned that each of the steps is amenable to automation in an integrated pathway for native-like tertiary structure prediction at least for small alpha helical globular proteins.

We have further undertaken a comparison of the present methodology with homology modeling using four popular public domain softwares *viz* CPH models,[177] ESyPred3D,[178] Swiss-model[179] and 3D-PSSM.[180] CPH server builds the

Table 5 A performance appraisal of different modeling softwares for protein structure prediction (based on RMSD)

No	Protein PDB ID	CPH models[177] RMSD (Å)	ESyPred3D[178] RMSD (Å)	Swiss-model[179] RMSD (Å)	3D-PSSM[180] RMSD (Å)	Present work[b] RMSD (Å)
1	1IDY (1–54)[a]	3.96 (2–54)[a]	3.79 (2–51)[a]	5.73 (1–51)[a]	3.66 (1–51)[a]	3.36
2	1PRV (1–56)[a]	5.66 (2–56)[a]	5.56 (3–56)[a]	6.67 (3–56)[a]	5.94 (1–56)[a]	3.87

[a] Numbers in parenthesis represent the length (number of amino acids) of the protein model. [b] Structure with lowest RMSD bracketed in the 100 lowest energy structures.

Table 6 CPU time required for each step of the pathway for two representative proteins

Protein	Step (in Fig. 1)	Stage	Time required for single structure (s)	Total structures	Time required for all the structures per processor (s)	Total time required on 50 processors
1GVD[a]	Step (3) + Step (4)	Trial structure generation and application of filters	~0.48	25 980	~12 600	
	Step (5)	Removal of clashes	~78.37	25 980	~2036 052	~40 721 s (~11 h)
	Step (6)	Minimization	~5.87	25 980	~152 502	~3050 s (~51 min)
	Step (7)	Energy calculation	~0.1	25 980	~2598	~52 s (~1 min)
	Step (8)	Metropolis MC	~4067	100	~406 700	~8134 s (~2.3 h)
2EZH[b]	Step (3) + Step (4)	Trial structure generation and application of filters	~0.346	249 740	~86 400	
	Step (5)	Removal of clashes	~124.78	249 740	~31 162 557	~623 251 s (~7 days)
	Step (6)	Minimization	~7.46	249 740	~1 863 060	~37261 s (~10 h)
	Step (7)	Energy calculation	~1	249 740	~249 740	~4995 s (~1.4 h)
	Step (8)	Metropolis MC	~10 228	100	~1 022 800	~20 456 s (~5.7 h)

[a] Contains 3 helices and 52 residues. [b] Contains 4 helices and 65 residues.

models according to non-redundant and PDB databases and 3D-PSSM provides a 3D model according to the fold library. On the other hand both EsyPred3D and Swiss-model provide the option for template selection for generation of homology models. Sequence alignments for all of the twelve systems were performed using the PHI BLAST from NCBI (http://www.ncbi.nlm.nih.gov/) and all the templates having more than 30% sequence similarity were selected leaving out native and sequence homologs belonging to the same family. The selected templates were used to generate the 3D model of the protein using SWISS-MODEL and ESyPred3D. We found the present methodology to result in comparable accuracies, if not better, for two out of twelve proteins than the homology models (Table 5). In these two proteins (PDB codes 1IDY and 1PRV) tertiary structural models were built according to templates having maximum sequence similarity of 38% and 48%, respectively. For the rest of the proteins, because of the presence of closely related structures in the database, homology models lie within an RMSD of less than 1 Å of the native structure. This study provides an example where *ab initio* techniques can provide candidates for native-like structure, when comparative modeling techniques fail due to database inadequacies. In our calculations, almost at every stage (except for trial structure generation and application of filters), the calculations are performed on 50 UltraSparc III 900 MHz processors. CPU times required for each step, as shown in Fig. 1, are presented in Table 6, for two proteins namely 1GVD and 2EZH, having 3 and 4 helices, respectively. Jobs are data parallel and hence the ease and advantage of implementing the pathway on large clusters towards a resolution of the problem in realistic time scales. However, there is scope for further speed up.

The length of intervening loops (*i.e.* loops flanked by the secondary structural elements) was found to affect the quality of the tertiary structures predicted. A suggestion for improvement in systems with smaller or larger loops is the generation of more conformations per dihedral of the loop or at least six dihedrals per loop instead of the present four or the selection of residues other than the middle two for trial structure generation. We have used the last strategy by selecting residues other than the middle two for trial structure generation for one protein (2EZH, Table 3). For this case an RMSD of less than 5 Å from the native structure was obtained. When proline residue occurs in loop, all four conformations are not possible in view of the restrictions on phi dihedral. This fact is evident in the case of protein 1BW6 (Table 3) which has proline in one of the loops. Occurrence of proline in fact presents an opportunity to restrict the sample size. In all cases irrespective of the length of the loops, the protocol yields structures to within 5 Å RMSD of the native.

The proposed methodology has shown to bracket reliable structures for all alpha helical proteins, given the second generation force field parameters for amino-acids, secondary structural information and large rotamer libraries. It may be important to classify proteins into different structural classes for better predictions as indicated in literature.[181,182] Further improvements to the methodology besides improvements to scoring function include introduction of a new filter based on loop dihedrals which can further reduce the size of probable candidates to less than 100. It is also anticipated that more efficient Monte Carlo strategies or explicit solvent molecular dynamics simulations on the selected structures can aid in optimizing side chain orientations and facilitate favorable packing interactions. Hydrophobicity and packing fraction filters could be utilized at this stage for selecting some representatives for the native and thus help in reducing the number of candidate structures further. *Post facto* free energy analyses of MD trajectories, on the candidates to narrow down the choice of the native-like structures are contemplated as the last step. Work on these lines is in progress.

5 Conclusions

The proposed computational pathway arrives at the tertiary fold of proteins starting from the secondary structure, for small alpha helical globular proteins comprising less than 100 amino acids. Studies on twelve proteins consisting of three to four helices demonstrated that structure to within 3–5 Å of the native are bracketed in 100 lowest energy structures in all cases without exception.

Acknowledgements

Funding from the Department of Biotechnology is gratefully acknowledged. Ms Kumkum Bhushan is a recipient of the Senior Research Fellow award from the Council of Scientific and Industrial Research (CSIR), India. The authors are thankful to Dr N. Pant for a critical reading of the manuscript.

References

1 D. D. Shoemaker, E. E. Schadt, C. D. Armour and Y. D. He *et al.*, *Nature*, 2001, **6822**, 922–927.

2 J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li and R. J. Mural *et al.*, *Science*, 2001, **291**, 1304–1351.

3 C. B. Anfinson, *Science*, 1973, **181**, 223–230.

4 K. A. Dill, *Biochemistry*, 1990, **29**, 7133–7155.

5 H. S. Chan and K. A. Dill, *Phys. Today*, 1993, **46**, 4–32.

6 K. A. Dill, *Curr. Opin. Struct. Biol.*, 1993, **3**, 99–103.

7 K. A. Dill, H. S. Chan and K. Yue, *Macromol. Symp.*, 1995, **98**, 615–617.

8 K. Yue and K. A. Dill, *Protein Sci.*, 1996, **5**, 254–261.

9 R. Doyle, K. Simons, H. Qian and D. Baker, *Proteins*, 1997, **29**, 282–291.

10 K. W. Plaxco, D. S. Riddle, V. Grantcharova and D. Baker, *Curr. Opin. Struct. Biol.*, 1998, **8**, 80–85.

11 D. Baker, *Nature*, 2000, **405**, 39–42.

12 R. L. Baldwin, *Nat. Struct. Biol.*, 2001, **8**, 92–93.

13 C. L. Brooks IIIrd, M. Gruebele, J. N. Onuchic and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 11037–11038.

14 J. E. Shea and C. L. Brooks IIIrd, *Annu. Rev. Phys. Chem.*, 2001, **52**, 499–535.

15 C. L. Brooks IIIrd, *Nature*, 2002, **420**, 33–34.

16 C. L. Brooks IIIrd, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 1099–1100.

17 G. D. Rose and T. P. Creamer, *Proteins*, 1994, **19**, 1–3.

18 V. Daggett and A. R. Fersht, *Trends Biochem. Sci.*, 2003, **28**, 18–25.

19 T. Lazaridis and M. Karplus, *Biophys. Chem.*, 2003, **100**, 367–395.

20 K. W. Plaxco and M. Gross, *Nat. Struct. Biol.*, 2001, **8**, 659–670.

21 S. E. Radford, *Trends Biochem. Sci.*, 2000, **25**, 611–618.

22 S. E. Radford, *Fold. Des.*, 1998, **3**, R59–R63.

23 S. Gianni, U. Mayor and A. R. Fersht, *Ital. J. Biochem.*, 2003, **52**, 154–161.

24 N. Ferguson and A. R. Fersht, *Curr. Opin. Struct. Biol.*, 2003, **13**, 75–81.

25 J. S. Fetrow, A. Giammona, A. Kolinski and J. Skolnick, *Curr. Pharm. Biotechnol.*, 2002, **3**, 329–347.

26 M. R. Betancourt and J. Skolnick, *J. Comput. Chem.*, 2000, **22**, 339–353.

27 J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes, *Annu. Rev. Phys. Chem.*, 1997, **48**, 545–600.

28 W. A. Eaton and P. G. Wolynes, *Phys. World*, 1999, **12**, 39–44.

29 J. N. Onuchic and P. G. Wolynes, *Curr. Opin. Struct. Biol.*, 2004, **14**, 70–75.

30 C. M. Dobson, A. Sali and M. Karplus, *Angew. Chem., Int. Ed.*, 1998, **37**, 868–893.

31 A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson and M. Karplus, *Trends Biochem. Sci.*, 2000, **25**, 331–339.

32 K. A. Dill and H. S. Chan, *Nature Struct. Biol.*, 1997, **4**, 10–19.

33 M. Mayor, N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sato, G. S. Jas, S. M. V. Freund, D. O. V. Alonso, V. Daggett and A. R. Fersht, *Nature*, 2003, **421**, 863–867.

34 V. S. Pande, A. Yu Grosberg, T. Tanaka and D. S. Rokhsar, *Curr. Opin. Struct. Biol.*, 1998, **8**, 68–79.

35 D. J. Brockwell, D. A. Smith and S. E. Radford, *Curr. Opin. Struct. Biol.*, 2000, **10**, 16–25.

36 S. Nauli, B. Kuhlman and D. Baker, *Nat. Struct. Biol.*, 2001, **8**, 602–605.

37 J. Venkatraman, S. C. Shankaramma and P. Balaram, *Chem. Rev.*, 2001, **101**, 3131–3152.
38 I. L. Karle, C. Das and P. Balaram, *Proc. Natl. Acad. Sci. USA*, 2000, **28**, 3034–3037.
39 A. Smith, *Nature*, 2003, **6968**, 883.
40 C. M. Dobson, *Nature*, 2003, **6968**, 884–890.
41 C. M. Dobson, *Trends Biochem. Sci.*, 1999, **24**, 329–332.
42 C. M. Dobson, *Nature*, 2002, **418**, 729–730.
43 R. Sitia and I. Braakmam, *Nature*, 2003, **6968**, 891–894.
44 A. L. Goldberg, *Nature*, 2003, **6968**, 895–899.
45 D. J. Selkoe, *Nature*, 2003, **6968**, 900–904.
46 F. E. Cohen and J. W. Kelly, *Nature*, 2003, **6968**, 905–910.
47 M. C. Peitsch, *Biochem. Soc. Trans.*, 1996, **24**, 274–279.
48 S. H. Bryant and C. E. Lawrence, *Proteins*, 1993, **16**, 92–112.
49 D. T. Jones, *Curr. Opin. Struct. Biol.*, 1997, **7**, 377–387.
50 C. Venclovas, *Proteins*, 2001, **5**, 47–54.
51 B. Al-Lazikani *et al.*, *Curr. Opin. Struct. Biol.*, 2001, **5**, 51–56.
52 J. Moult, *Curr. Opin. Biotechnol.*, 1999, **10**, 583–588.
53 B. Rost and C. Sander, *Annu. Rev. Biophys. Biomol. Struct.*, 1996, **25**, 113–136.
54 A. Tramontanoa and V. Morea, *Proteins*, 2003, **53**, 352–368.
55 N. Guex, A. Diemand and M. C. Peitsch, *Trends Biochem. Sci.*, 1999, **24**, 364–367.
56 A. Sali, E. I. Shakhnovich and M. Karplus, *Nature*, 1994, **477**, 248–251.
57 A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson and M. Karplus, *Trends Biochem. Sci.*, 2000, **25**, 331–339.
58 B. Honig, *J. Mol. Biol.*, 1999, **293**, 283–293.
59 K. A. Dill, S. Bromberg, K. Z. Yue and K. M. Fiebig, *Protein Sci.*, 1995, **4**, 561–602.
60 C. Levinthal, *J. Chim. Phys.*, 1968, **65**, 44–45.
61 D. J. Osguthorpe, *Curr. Opin. Struct. Biol.*, 2000, **10**, 146–152.
62 C. Hardin, T. V. Pogorelov and Z. Luthey-Schulten, *Curr.Opin. Struct. Biol.*, 2002, **12**, 176–181.
63 R. Srinivasan, P. J. Fleming and G. D. Rose, *Methods Enzymol.*, 2004, **383**, 48–66.
64 C. B. Anfinsen and H. A. Scheraga, *Adv. Protein Chem.*, 1975, **29**, 205–300.
65 J. M. Yon, *Biochemie*, 1978, **60**, 581–591.
66 M. G. Rossmann and P. Argus, *Annu. Rev. Biochem.*, 1981, **50**, 497–532.
67 J. Garnier, *Biochemie*, 1990, **72**, 513–524.
68 D. J. Thomas, *FEBS Lett.*, 1992, **307**, 10–13.
69 R. A. Friesner and J. R. Gunn, *Annu. Rev. Biophys. Biomol. Struct.*, 1996, **25**, 315–342.
70 H. A. Scheraga, *Biophys. Chem.*, 1996, **59**, 329–339.
71 M. Levitt, M. Gerstein, E. Huang, S. Subbiah and J. Tsai, *Annu. Rev. Biochem.*, 1997, **66**, 549–579.
72 M. Karplus, *Fold. Des.*, 1997, **2**, S69–75.
73 R. J. Ellis, *Curr. Biol.*, 2003, **13**, R881–883.
74 T. E. Crieghton, *Adv. Biophys.*, 1984, **18**, 1–20.
75 D. P. Goldenberg, *Annu. Rev. Biophys. Biophys. Chem.*, 1988, **17**, 481–507.
76 A. R. Fersht, A. Matouschek, J. Sancho, L. Serrano and S. Vuilleumier, *Faraday Discuss.*, 1992, **93**, 183–193.
77 J. Skolnick, A. Kolinski and A. Godzik, *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 2099–2100.
78 R. L. J. Baldwin, *J. Biomol. NMR*, 1995, **5**, 103–109.
79 J. S. Weissman, *Chem. Biol.*, 1995, **2**, 255–260.
80 D. Thirumalai and D. K. Klimov, *Curr. Opin. Struct. Biol.*, 1999, **9**, 197–207.
81 R. Zwanzig, A. Szabo and B. Bagchi, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 20.
82 A. Mukherjee and B. Bagchi, *Proc. Indian Acad. Sci.*, 2003, **115**, 621.
83 G. Srinivas and B. Bagchi, *Curr. Sci.*, 2002, **82**, 179.
84 A. Mukherjee and B. Bagchi, *J. Chem. Phys.*, 2003, **118**, 4733.
85 M. Lipton and W. C. Still, *J. Comput. Chem.*, 1988, **9**, 343–355.
86 M. Saunders, *J. Am. Chem. Soc.*, 1987, **109**, 3150–3152.
87 D. M. Ferguson and D. J. Raber, *J. Am. Chem. Soc.*, 1989, **111**, 4371–4378.
88 Z. Q. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 6611–6615.
89 G. Chang, W. C. Guida and W. C. Still, *J. Am. Chem. Soc.*, 1989, **111**, 4379–4386.
90 D. E. Goldberg, in *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.
91 N. Gautham and Z. A. Rafi, *Curr. Sci.*, 1992, **63**, 560–564.
92 K. Vengadesan and N. Gautham, *Biophys. J.*, 2003, **84**, 2897–2906.
93 H. A. Scheraga, *Biophys. Chem.*, 1996, **59**, 329–339.
94 L. Piela, J. Kostrowicki and H. A. Scheraga, *J. Phys. Chem.*, 1989, **93**, 3339–3346.
95 J. Kostrowicki and H. A. Scheraga, *J. Phys. Chem.*, 1992, **96**, 7442–7449.
96 K. D. Gibson and H. A. Scheraga, *J. Comput. Chem.*, 1990, **11**, 468–486.
97 M. Saunders, K. N. Houk, Y. D. Wu, W. C. Still, M. Lipton, G. Chang and W. C. Guida, *J. Am. Chem. Soc.*, 1990, **112**, 1419–1427.
98 A. R. Leach, in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, VCH Publishers, New York, 1991, pp. 1–55.
99 H. A. Scheraga, in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, VCH Publishers, New York, 1992, pp. 73–142.
100 H. A. Scheraga, J. Lee, J. Pillardy, Y. J. Ye, A. Liwo and D. Ripoll, *J. Global Optim.*, 1999, **15**, 235–260.
101 M. Vásquez, G. Nemethy and H. A. Scheraga, *Chem. Rev.*, 1994, **94**, 2183–2239.
102 A. Neumaier, *SIAM Rev.*, 1997, **39**, 407–460.
103 C. A. Floudas, J. L. Klepeis and P. M. Pardalos, in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, ed. M. Farach-Colton, F. S. Roberts, M. Vingron and M. Waterman, American Mathematical Society, 1999, pp. 141–171.
104 J. K. Shin and M. S. John, *Biopolymers*, 1991, **31**, 177–185.
105 J. Lee, H. A. Scheraga and S. Rackovsky, *J. Comput. Chem.*, 1997, **18**, 1222–1232.
106 J. L. Klepeis, M. J. Pieja and C. A. Floudas, *Biophys. J.*, 2003, **84**, 869–882.
107 B. Fain and M. Levitt, *J. Mol. Biol.*, 2001, **305**, 191–201.
108 G. Hanggi and W. Braun, *FEBS Lett.*, 1994, **344**, 147–153.
109 I. D. Kuntz, G. M. Krippen and P. A. Kollman, *Biopolymers*, 1979, **18**, 939–957.
110 J. Skolnick, A. Kolinski and A. R. Ortiz, *J. Mol. Biol.*, 1997, **65**, 217–241.
111 E. S. Huang, R. Samudrala and J. W. Ponder, *Protein Sci.*, 1998, **7**, 1998–2003.
112 E. S. Huang, R. Samudrala and J. W. Ponder, *J. Mol. Biol.*, 1999, **290**, 267–281.
113 R. Samudrala, Y. Xia, E. Huang and M. Levitt, *Proteins*, 1999Suppl. 3), 194–198.
114 Y. Xia, E. S. Huang, M. Levitt and R. Samudrala, *J. Mol. Biol.*, 2000, **300**, 171–185.
115 J. Lee, A. Liwo, D. R. Ripoli, J. Pillardy and H. A. Scheraga, *Proteins*, 1999Suppl. 3), 204–208.
116 B. Özkan and I. Bahar, *Proteins*, 1998, **32**, 211–222.
117 M. Vásquez and H. A. Scheraga, *Biopolymers*, 1985, **24**, 1437–1447.
118 K. D. Gibson and H. A. Scheraga, *J. Comput. Chem.*, 1987, **8**, 826–834.
119 S. Vajda and C. DeLisi, *Biopolymers*, 1990, **29**, 1755–1772.
120 S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, *Science*, 1983, **220**, 671–680.
121 S. R. Wilson, W. Cui, J. W. Moskowitz and K. E. Schmidt, *Tetrahedron Lett.*, 1988, **29**, 4373–4376.
122 L. B. Morales, R. Garduno-Juarez and D. Romero, *J. Biomol. Struct. Dyn.*, 1991, **8**, 721–735.
123 Y. Okamoto, T. Kikuchi and H. Kawai, *Chem. Lett.*, 1992, **1992**, 1275–1278.
124 A. Kolinski and J. Skolnick, *Proteins*, 1994, **18**, 353–356.
125 A. R. Dinner, A. Sali and M. Karplus, *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 8361–8365.
126 K. Ishikawa, K. Yue and K. A. Dill, *Protein Sci.*, 1999, **8**, 716–721.
127 A. Monge, E. J. P. Lathrop, J. R. Gunn, P. S. Shenkin and R. A. Friesner, *J. Mol. Biol.*, 1995, **247**, 995–1012.
128 M. Levitt and A. Warshel, *Nature*, 1975, **253**, 694–698.
129 J. A. McCammon, B. R. Gelin and M. Karplus, *Nature*, 1977, **267**, 585–590.
130 J. Skolnick and A. Kolinski, *Annu. Rev. Phys. Chem.*, 1989, **40**, 207–235.
131 R. Day and V. Daggett, *Adv. Protein Chem.*, 2003, **66**, 373–403.
132 R. M. Levy and E. Gallicchio, *Annu. Rev. Phys. Chem.*, 1998, **49**, 531–567.
133 R. M. Levy, D. Perahia and M. Karplus, *Proc. Natl. Acad. Sci. USA*, 1982, **79**, 1346–1350.
134 M. Schaefer and M. Karplus, *J. Phys. Chem.*, 1996, **100**, 1578–1599.
135 Y. Duan and P. A. Kollman, *Science*, 1998, **282**, 740–744.
136 S. Chowdhury, M. C. Lee, G. Xiong and Y. Duan, *J. Mol. Biol.*, 2003, **327**, 711–717.

137 V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. Snow, E. Sorin and B. Zagrovic, *Biopolymers*, 2003, **68**, 91–109.
138 M. R. Lee, J. Tsai, D. Baker and P. A. Kollman, *J. Mol. Biol.*, 2001, **313**, 417–430.
139 D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Chetham III, S. DeBolt, D. Ferugson, G. Seibel and P. Kollman, *Comput. Phys. Commun.*, 1995, **91**, 1–41.
140 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acid Res.*, 2000, **28**, 235–242.
141 G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.*, 1968, **23**, 283–437.
142 J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, 1987, **193**, 775–791.
143 P. Tuffery, C. Etchebest, S. Hazout and R. Lavery, *J. Biomol. Struct. Dyn.*, 1991, **8**, 1267–1289.
144 R. L. Dunbrack Jr. and M. Karplus, *J. Mol. Biol.*, 1993, **230**, 543–574.
145 P. Tuffery, C. Etchebest and S. Hazout, *Protein Eng.*, 1997, **10**, 361–372.
146 R. L. Dunbrack, Jr. and F. E. Cohen, *Protein Sci.*, 1997, **6**, 1661–1681.
147 M. De Maeyer, J. Desmet and I. Lasters, *Fold. Des.*, 1997, **2**, 53–66.
148 S. C. Lovell, J. M. Word, J. S. Richardson and D. C. Richardson, *Proteins*, 2000, **40**, 398–408.
149 Z. Xiang and B. Honig, *J. Mol. Biol.*, 2001, **311**, 421–430.
150 R. L. Dunbrack Jr., *Curr. Opin. Struct. Biol.*, 2002, **12**, 431–440.
151 N. Go and H. A. Scheraga, *Macromolecules*, 1970, **3**, 178–187.
152 C. Calladine and H. Drew, in *Understanding DNA*, Academic Press, London, 1992.
153 S. B. Smith, L. Finzi and C. Bustamante, *Science*, 1992, **258**, 1122–1126.
154 W. Kühn, *Kolloid Z.*, 1934, **68**, 2–11.
155 P. J. Flory, in *Principles of Polymer Chemistry*, Cornell Univ. Press, Ithaca, New York, 1953.
156 P. G. de Gennes, in *Scaling Concepts in Polymer Physics*, Cornell Univ. Press, Ithaca, New York, 1979.
157 W. Kauzmann, *Adv. Protein Chem.*, 1959, **14**, 1–63.
158 J. Janin, *Nature*, 1979, **277**, 491–492.
159 C. C. Palliser and D. A. D. Parry, *Proteins*, 2001, **42**, 243–255.
160 R. Wolfenden, L. Anderson, P. M. Cullis and C. C. Southgate, *Biochemistry*, 1981, **20**, 849–855.
161 J. Kyte and R. F. Doolittle, *J. Mol. Biol.*, 1982, **157**, 105–132.
162 B. Lee and F. M. Richards, *J. Mol. Biol.*, 1971, **55**, 379–400.
163 S. Miller, J. Janin, A. M. Lesk and C. Chothia, *J. Mol. Biol.*, 1987, **196**, 641–656.
164 C. Maritan, C. Michelletti, A. Trovato and J. R. Banavar, *Nature*, 2000, **406**, 287–290.
165 A. Stasial and J. H. Maddocks, *Nature*, 2000, **406**, 251–253.
166 O. B. Ptitsyn, *J. Mol. Biol.*, 1998, **278**, 655–666.
167 P. G. Squire and M. E. Himmel, *Arch. Biochem. Biophys.*, 1979, **196**, 165–177.
168 N. Arora and B. Jayaram, *J. Phys. Chem.*, 1998, **102**, 6139–6144.
169 N. Arora and B. Jayaram, *J. Comput. Chem.*, 1997, **18**, 1245–1252.
170 (*a*) B. Jayaram, in *Proceedings of the Ninth Conversation in Biomolecular Stereodynamics*, ed. R. H. Sarma, Adenine Press, New York; (*b*) B. Jayaram, *J. Biol. Struct. Dyn.*, 1996, **1**, 109.
171 M. A. Young, B. Jayaram and D. L. Beveridge, *J. Phys. Chem. B*, 1998, **102**, 7666–7669.
172 P. Narang, K. Bhushan, S. Bose and B. Jayaram, *J. Phys. Chem.*, submitted.
173 B. Jayaram and D. L. Beveridge, *J. Phys. Chem.*, 1991, **95**, 2506–2516.
174 P. S. Ramanathan and H. L. Friedman, *J. Chem. Phys.*, 1971, **54**, 1086.
175 A. Monge, R. A. Friesner and B. Honig, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 5027–5029.
176 D. A. Hinds and M. Levitt, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 2536–2540.
177 O. Lund, M. Nielsen, C. Lundegaard and P. Worning, *Abstract at the CASP5 Conference*, 2002, A102.
178 C. Lambert, N. Leonard, X. De Bolle and E. Depiereux, *Bioinformatics*, 2002, **18**, 1250–1256.
179 T. Schwede, J. Kopp, N. Guex and M. C. Peitsch, *Nucleic Acid Res.*, 2003, **31**, 3381–3385.
180 L. A. Kelley, R. M. MacCallum and M. J. E. Sternberg, *J. Mol. Biol.*, 2000, **299**, 499–520.
181 B. Rost and C. Sander, *Protein Eng.*, 1993, **6**, 831–836.
182 M. M. Gromiha and S. Selvaraj, *Protein Eng.*, 1998, **11**, 249–251.