Delineating the mainchain topology of four-helix bundle proteins using the genetic algorithm and knowledge based on the amino acid sequence alone

Patrick Argos and Thomas Dandekar

European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10 22 09, 69012 Heidelberg, Germany

Abstract

The genetic algorithm was used to search mainchain conformational space in folding four-helix bundle proteins from knowledge derived from the amino acid sequence alone. The folding simulations were grid-free and relied on a selection of representative dihedral angular configurations for the backbone atoms. Successful fitness criteria and weights, optimized for an idealized helix bundle, included hydrophobic interactions, structural compactness without atomic clashes, and secondary structural nucleation regions predicted from the primary sequence. The proper mainchain fold for the all-helical proteins cytochrome b_{562}, hemerythrin, and cytochrome c' was achieved. The relative simplicity of the fitness function suggested that only a few basic energetic principles need be considered to achieve a molten-globule folding intermediate characterized by a lack of complete and specific sidechain interactions.

1 Introduction

Nature has produced (and continues to do so) surviving species in the Earth's planetary setting. Large phylogenetic spaces are scanned by introducing various combinational genetic species into the environment with the fittest, or at least those most facilely surviving, providing the chromosomal information for succeeding and more optimal generations. The natural techniques used to achieve a high quality solution from the near infinite genetic possibilities include selection of optimal species, mutation within a given individual, and recombination (cross-over) amongst chromosomes. John Holland [1] originally introduced these natural adaptation methods under the guise of the "genetic algorithm" for the solution of technical problems requiring searches over large combinatorial character spaces such as gas pipeline flow and artificial intelligence [2].

The prediction of a protein's fold and structure from only a knowledge of its amino acid sequence is also a prime target for the genetic algorithm, especially given the near limitless possibilities theoretically available for the atomic coordinates generally ranging in the thousands; for example, there are about 10 atoms on average composing each residue and

around 350 residues in an average protein. Nature has found and finds unique folds for unique and surviving amino acid sequences. In the work described here the genetic algorithm, based on a grid-free and three-dimensional system relying on standard and representative dihedral angles applied to the backbone atoms and taken from known tertiary protein structures, is applied to the amino acid sequence of three four-helix bundle proteins. Only simple and a few fitness criteria are utilized: hydrophobic interactions, nucleating secondary structural segments predicted by techniques reliant upon the primary structure alone, and structural compactness without atomic clashes. Selection of criteria and appropriate weights for the fitness terms were determined from an idealized four-helix bundle topology. Details of this effort can be found in Dandekar & Argos [3]. Previous efforts in this area include that of the present authors [4] who used the genetic algorithm for various sequence design tasks and grid-bound and idealized folding applications, Unger & Moult [5] who demonstrated that genetic algorithms are superior to Monte Carlo simulations in searching conformations for two-dimensional protein models, and Sun [6] who relies on full energy terms in the fitness function to elicit tertiary structure for very small proteins such as mellitin.

2 Materials and Methods

2.1 Models and structures

The chromosome was represented as a successive string of binary digits encoding for the various dihedral angles at the C_α atom of the mainchain from the N- to C-terminus. Several rotation angles (ϕ, Ψ) that characterize the conformational space utilized by known protein structures [7] were each assigned a specific binary code. The configurations included those of the α-helix (-65, -40); 3_{10} helix (-89, -1); β-strand (-117, 142); extended structure (-69, 140); glycine ((78, 20) and (103, -176)); and cis-proline often observed in turns (-82, 133). Standard peptide bond angles and distances were maintained amongst the mainchain atoms: C_α-C', 1.53Å; C'-O, 1.24Å; C'-N, 1.32Å; and N-C_α, 1.47Å. Sidechain atoms were not included though the entire backbone group was represented (C', O, C_α, N); nonetheless, sidechain characteristics such as hydrophobicity were assigned to each mainchain C_α atom for use in the fitness function.

Fitness criteria and weights were tested on an idealized four-helix bundle [8] with topology $\alpha_{10}L_5\alpha_9L_5\alpha_9L_5\alpha_{10}$ where α_N designates an α-helix of residue length N and similarly for L_N representing loop segments. Residues were assigned as hydrophobic (A) or non-hydrophobic (a) according to an amphipathic wheel [9] with pattern AaaAaaaAaa. Figure 1 illustrates the bundle fold and distribution helical C_α atom positions projected onto a circular wheel.

Simulation trials based on the use of the genetic algorithm and known primary structures were applied to three proteins with experimentally determined atomic coordinates; namely, hemerythrin from sipunculid worm [10]; cytochrome b_{562} from *E.coli* [11]; and cytochrome c' from *rhodospirillum molischianum* [12]. The respective codes given to the proteins in the Brookhaven data-bank of tertiary structures [13] are 1HMD, 256B, and 1CRN. The observed and predicted C_α positions were superposed by the method of McLachlan [14]. The overall root-mean-square distance deviation (RMSD) between structurally equivalent observed and simulated C_α atoms was used as a criterion for the accuracy of the predicted topology.

Predictions of secondary structural regions (α-helix, β-strand, coil) based on the amino acid sequence alone were taken from the method of Ptitsyn & Finkelstein [15]. This

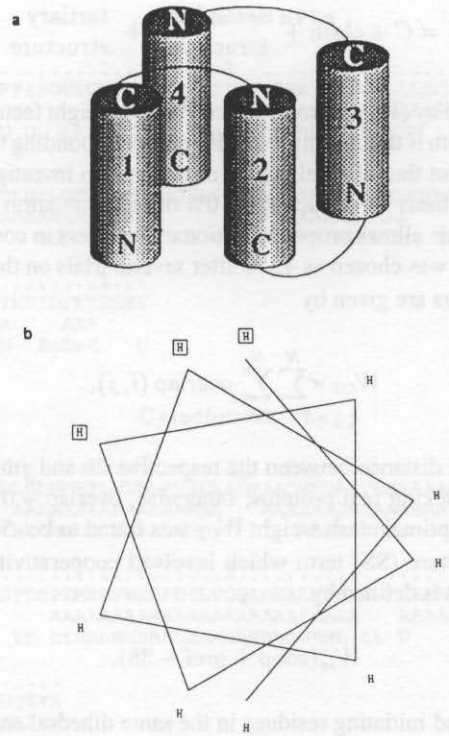

Figure 1: a) Schematic topology of the idealized four-helix bundle where helices are shown as cylinders with respective N- and C-termini along the primary sequence. Connecting loops are shown as thin, arched lines. b) Illustration of a helical 10-residue amphipathic wheel which is a projection of C_α atom positions at line intersections. The peak hydrophobic residues are boxed.

technique did not rely on sequence statistics that could have included the proteins tested here, thereby biasing the simulation results. Instead, their method was based upon stereochemical theory and hydrophobic templates.

2.2 Genetic algorithm

In the genetic algorithm procedures two parents were selected from the population with a probability of 0.2 and allowed one random crossover site. The number of generations allowed was set high to assure adequate convergence in the simulation runs (632 individuals for 632 generations); the mutation rate was set as one mutated bit per individual per generation.

Proper fitness criteria and associated weights are critical for successful optimization by the genetic algorithm. The fitness weights can be positive (reward) or negative (punishment) in sign. An idealized four-helix bundle as previously described was used to optimize

the criteria and weights. The fitness function (ff) summed four terms:

$$ff = C + \text{clash} + \begin{matrix}\text{secondary}\\ \text{structure}\end{matrix} + \begin{matrix}\text{tertiary}\\ \text{structure}\end{matrix}, \tag{1}$$

The term C is defined by (W_C*lchrom) where W_C is a weight factor, asterisk (*) refers to multiplication, and lchrom is the chromosomal length corresponding to the number of binary slots required to represent the dihedral angles of the protein investigated. The constant C normalizes the overall fitness value such that 10% of the population has no positive fitness in the first generation; this allows proper evolutionary progress in converging to an optimal (fittest) population. W_C was chosen as +350 after several trials on the idealized bundle.

The clash term factors are given by

$$W_{CI} * \sum_{i=1}^{N} \sum_{j=i}^{N} \text{overlap}\,(i,j), \tag{2}$$

where overlap = 1 if the distance between the respective ith and jth C_α atoms is less than or equal to 3.8 Å, mimicking real proteins; otherwise, overlap = 0. N is the number of mainchain atoms. The optimal clash weight W_{CI} was found to be -500.

The secondary structure (SS) term which involved cooperativity (coop) and dihedral angular preference (pref) is defined by

$$W_{ss}(\text{coop} + \text{pref} - 38), \tag{3}$$

where two successive and initiating residues in the same dihedral state were each awarded +10 under coop, followed by successive additions of +1 for each adjacent extension of the same conformation in either sequential direction. If residues were in the same secondary structural state as that predicted, their pref value was assigned a value of +20; extensions of this state under the aforementioned coop conditions was also allowed beyond the predicted regions up to 3 residues within another nucleation region or extension site. The subtraction of 38 normalized according to the number of helical residues in the idealized bundle. The preferred weight W_{ss} for the secondary structure component was +14.

The tertiary structure term consisted of the addition of two compactness measures:

$$\left(W_{gs} * \begin{matrix}\text{global}\\ \text{scatter}\end{matrix}\right) + \left(W_{hs} * \begin{matrix}\text{hydrophobic}\\ \text{scatter}\end{matrix}\right), \tag{4}$$

where scatter is simply the sum of the distances of each C_α atom from the center-of-mass of the current fold. The global scatter took into account all C_α atoms while the hydrophobic term involved only those associated with strongly hydrophobic residues (Met, Ile, Leu, Val, Tyr, Cys, Phe) (16). Under optimal conditions, $W_{gs} = -24$ and $W_{hs} = -19$.

Fitness terms were also used to mimic the formation of hydrogen bonds between a mainchain carbonyl oxygen and peptide nitrogen amongst any two residues in the fold or between residues following a helical pattern (residue i to residue $i + 4$). The H-bond effects are not discussed as such fitness terms did not foster the fold of the idealized bundle, either by their exclusion from the fitness function built with the other criteria or from their inclusion with one or more of the remaining criteria deleted. Dandekar & Argos [3] provide details.

Hemerythrin

```
  1-50     ....,....1....,....2....,....3....,....4....,....5
SEQ        GFPIPDPYCWDISFRTFYTIVDDEHKTLFNGILLLSQADNADHLNELRRC
                     AAA     AAAAAAAAAAAAAAAAAAA   AAAAAAAAAA
            U U  tEEEeEeE  EEeEEEthhhhhHHhhHHHH ttttttHhhhHhhH

 51-100    ....,....1....,....2....,....3....,....4....,....5
SEQ        TGKHFLNEQQLMQASQYAGYAEHKKAHDDFIHKLDTWDGDVTYAKNWLVN
           AAAAAAAAAAAAAA      AAAAAAAAAAAAAAA       AAAAAAAAAA
               UU    UU    U  U         hHHhhH  Ttt HhHhhhHHHh

101-113    ....,....1...
SEQ        HIKTIDFKYRGKI
           AA    AAA
           hH  EeEeE   U
```

Cytochrome b_{562}

```
  1-50     ....,....1....,....2....,....3....,....4....,....5
SEQ        ADLEDNMETLNDNLKVIEKADNAAQVKDALTKMRAAALDAQKATPPKLED
             AAAAAAAAAAAAAAAAA   AAAAAAAAAAAAAAAAA     AAA
             HhhhHhhHhhhHhHHh       HhhhHhhHh   U      tttU

 51-100    ....,....1....,....2....,....3....,....4....,....5
SEQ        KSPDSPEMKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNA
                AAAAAAAAAAAAAAAAAAAAAAAAA   AAAAAAAAAAAAAAAAA
             tt ttHhhHhhhHhHHHhhHhhhHhHh tt U     hhhhHhh

101-106    .....,.
SEQ        YHQKYR
           AAAAA
           U   U
```

Figure 2: Secondary structure predictions for hemerythrin and cytochrome b_{562} according to the method of Ptitsyn and Finkelstein [15] are either β-strand (e or E), α-helix (h or H), turn (t or T) or coil (blank or U). It must be emphasized that only helical and strand predictions were used as nucleation regions with fixed dihedral angles; the turn predictions shown are only informative. Observed secondary helix structures are also indicated by an A. Peak hydrophobic residues according to the scale of Manavalan and Ponnuswamy [16] are marked by capital letters in the predicted regions (H, E, or T) or U if no prediction was given by Ptitsyn and Finkelstein (1983). SEQ refers to the amino acid sequence given in single letter code.

3 Results

3.1 Idealized bundle

The accomplishment of an idealized four-helix fold without clashes as depicted in Figure 1 was used to judge the significance of the several folding principles represented in the fitness function. Tests from various combinations of criteria and weights, including exclusion of particular terms and variable helical (8 to 16) and loop residue (3 to 6) lengths, showed that compactness of all residues with emphasis on the strongly hydrophobic ones was essential

Table 1: Comparison of C_α RMSD (Å) fits amongst observed and optimal simulated folds including loop atoms.

	Hemerythrin		Cytochrome b_{562}
	simulated	observed	observed
Cytochrome b_{562} simulated	11.9	11.2	6.1
Cytochrome b_{562} observed	12.9	12.7	
Hemerythrin observed	6.7		

along with secondary structural nucleation sites (as small as three residues) for each helix of the bundle and subsequent cooperative extension. Nucleation sites at the helix termini were particularly effective.

3.2 Real all-α structures

The criteria and weights determined effective in the idealized case were then applied in genetic algorithm simulations involving folding of sequences from four-helix proteins with experimentally determined tertiary structures; namely, hemerythrin, cytochrome c' and cytochrome b_{562}. The Ptitsyn/ Finkelstein secondary structure prediction method was applied to each of the amino acid sequences; those segments predicted as helix and occasionally strand were assigned as preferred regions (Eq. (3)). Typical predictions are shown for hemerythrin and cytochrome b_{562} in Figure 2. In ten simulation trials for each of the proteins with different randomly chosen starting conformations, one-half generally achieved the proper four-helix topology, handedness and orientation; i.e., each helical axis was 20^o within that observed, antiparallelity in successive helices along the sequence was maintained, and secondary structural length was within 1 or 2 turns of that observed for individual helices. In each case, the fold selected was that with the largest fitness value in the various populations. Folds with lower values gave higher C_α RMSD values relative to the observed structure and did not display proper topology.

Table 1 lists the RMSD values for all structurally equiva- lent simulated and observed C_α atom positions for each of the three protein structures tested and over all atoms excluding loops and only over each helix observed. It is clear that the residue phasing of all helices is proper except for helix I in hemerythrin which is rotated by one-residue. These results are achieved despite cases where only one-third of an observed helix sequence span was predicted, single helical predictions were interrupted by prediction of a turn span, or N- or C-terminal regions (or both) were not predicted or mispredicted. The hemerythrin four-helix fold was also accomplished despite a long symmetry breaking N-terminal extended region. It is clear that the genetic algorithm achieved near proper orientation of the helices despite many other possible outcomes. For example, largely skewed helices could well minimize scatter about the center-of-mass. The simultaneous action of all the criteria in the fitness function is critical.

Table 2: Comparison of structurally equivalent simulated and observed C_α atom positions using RMSD values given in Å.

Protein	with loops	without loops	individual helices (from N–terminus)			
(fittest fold)			I	II	III	IV
Cytochrome b_{562}	6.1	5.8	0.4	0.6	0.8	1.2
Hemerythrin	6.7	5.6	2.7	1.7	0.4	0.7
Cytochrome c'	6.1	3.9	1.4	1.2	0.4	0.9

Table 3: Comparison of absolute and component values in the fitness functions. Notes: The absolute values for the fitness function components as well as the total values obtained in the simulation for cytochrome c' are given. Values are also given for the observed folds where predicted and observed secondary structures are taken as the same. Further, fitness values averaged over the failed simulation trials are listed.

	secondary structure	tertiary structure		total fitness
	(minus clashes	gs	hs	(+ positive constant C)
Cytochrome c'	52604	-25782	-6353	62819
observed fold	53480	-26958	-5371	63501
failed simulations	42010	-24544	-5706	54110

The genetic algorithm responds to the specific sequence under test. Table 2 lists the RMSD values for equivalent and superposed C_α atoms in cytochrome b_{562} and hemerythrin over all possible cross comparisons of simulated and observed topologies. The two proteins as simulated possessed the same number of residues allowing direct equivalencing. The observed and simulated folds for a given sequence are much closer (~6Å RMSD) than are those, simulated or observed, for different sequences (~12Å). Exemplary values for the various criteria in an optimized cytochrome c' fitness function resulting from a successful simulation are listed in Table III. These values are compared to those obtained from observed and average failed structures. The observed fold produces the best (maximum) fitness value with the simulated close by while the failed structures are clearly lower. Hydrophobic interactions and compactness contribute about 40% of the optimal value while secondary structure nucleation and subsequent cooperativity are responsible for the remaining 60%.

Figure 3 shows a stereo mainchain trace of cytochrome b_{562} where the observed structure and optimal simulation can be compared. Figure 4 illustrates a failed hemerythrin simulation

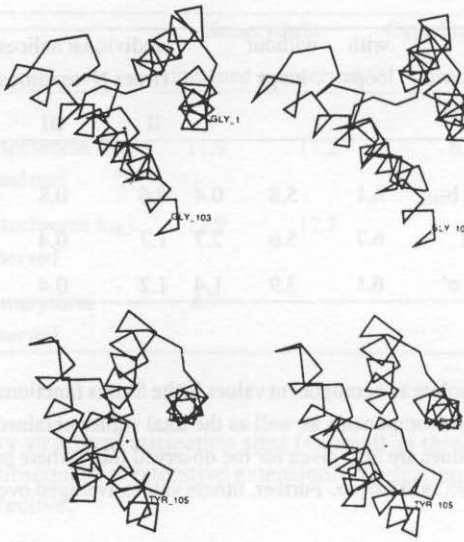

Figure 3: Fittest and observed folds (stereo view) of cytochrome b_{562}. The secondary structure predictions of Ptitsyn and Finkelstein [15] were used as nucleation sites (Figure 2). The top view illustrates the final fold in the simulation while the bottom is that of the experimentally determined fold (corresponding amino acids given). Virtual bonds connecting successive C_α atoms are shown as connecting lines.

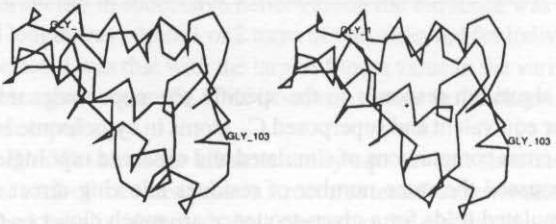

Figure 4: Example of a failed simulation trial for hemerythrin. Virtual bonds connecting sequential C_α atoms are shown as connecting lines in the stereo view.

where a bent region in one helix disturbs proper helix aggregation and topology for the bundle. Figure 5 shows exemplary and successive folding steps from early to late generations in the simulation of a small idealized helix bundle.

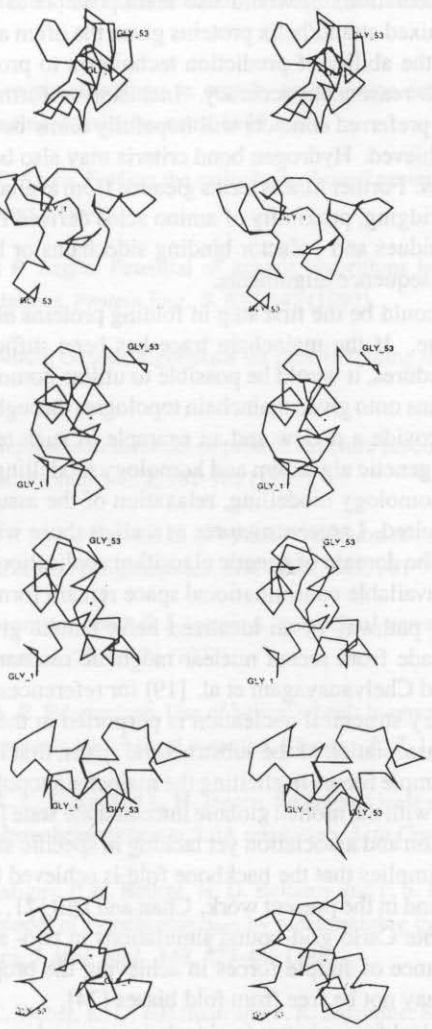

Figure 5: *Ab initio* folding simulation of a small idealized four-helix bundle as described in Material and Methods. A representative individual (a) is shown from the random start population, as are the fittest individuals after (b) 1, (c) 10, (d) 30, (e) 100 generations, and (f) the optimal fold at the simulation end. Virtual bonds connecting successive C_α atoms are given as connecting lines. Double images of each case are shown for a stereo perspective. Terminal residues are indicated as Gly.

4 Discussion

The results presented here show that the mainchain topology of small all-helical proteins can be identified by the genetic algorithm when used as a conformational search tool and coupled

with surprisingly simple folding (fitness) rules and with only a knowledge of the protein's amino acid sequence and information derivable from it (hydrophobic residue distribution and secondary structural prediction). It would also seem possible to achieve such optimal folds for all β-strand and mixed strand/helix proteins given the often amphipathic character of extended structure and the ability of prediction techniques to provide nucleation sites of such conformations with reasonable accuracy. Inclusion of further fitness terms such as residue size, shape, and preferred contacts will hopefully allow better orientation of the substructural units to be achieved. Hydrogen bond criteria may also be necessary to induce formation of sheet structures. Further fitness terms gleaned from available experimental data could involve disulphide bridging, proximity of amino acids derived from crosslinking and spatially close catalytic residues and cofactor binding sidechains or hydrophobic residues well conserved in multiple sequence alignments.

The genetic algorithm could be the first step in folding proteins *ab initio* from primary structural information alone. If the mainchain trace has been sufficiently delineated by the genetic algorithm procedures, it would be possible to utilize homology modelling techniques which place sidechains onto given mainchain topologies through energy calculations. Eisenmenger et al. [17] provide a review and an example of such techniques. No doubt further development of the genetic algorithm and homology modelling methods is required for such applications. In homology modelling, relaxation of the assumed but partly erroneous backbone fold is required. Larger structures as well as those with various secondary structural types must enter the domain of genetic algorithm application. The calculationally problemsome increases in available conformational space remain formidable.

The illustrative folding pathway of an idealized helix bundle given in Figure 5 corresponds to suggestions made from recent nuclear magnetic resonance experiments (see Baldwin and Roder [18] and Chelvanayagam et al. [19] for references) on folding intermediates where some secondary structural nucleation is purported in the early folding phase, followed by extension and association of the substructural spans, first locally and then globally. The effectiveness of simple forces in eliciting the mainchain topology from the genetic algorithm is also consistent with the molten globule intermediate state [20], characterized by secondary structure formation and association yet lacking in specific sidechain interactions. Observation of such states implies that the backbone fold is achieved by relatively nonspecific interactions as also found in the present work. Chan and Dill [21, 22] and Skolnick and Kolinski [23] who use Monte Carlo grid-bound simulations in two- and three-dimensions also emphasize the importance of simple forces in achieving the proper fold though their particular grid topologies may not be free from fold biases [24].

Complex physico-chemical forces between sidechains, secondary structural propensity and the overall cooperativity of protein folding have been translated in this work into abstract and simple rules which rely only on knowledge of the protein amino acid sequence. For the foreseeable future, until the exact physicochemical forces are known and can be modelled and calculated with sufficient detail and speed, the genetic algorithm approach provides a way to bridge the gap between secondary structure predictions and tertiary fold.

Acknowledgement

The authors are grateful for a postdoctoral fellowship to Thomas Dandekar from the Deutsche Forschungsgemeinschaft without whose financial support this work could never have been accomplished.

References

[1] J. Holland, Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, 1975.

[2] D.E Goldberg, Genetic algorithms in search, optimization and machine learning. Addison Wesley Publ., Reading, Massachu- setts, 1989.

[3] T. Dandekar and P. Argos, Folding the mainchain of small proteins with the genetic algorithm, *J. Mol. Biol.*, in press, (1994).

[4] T. Dandekar and P. Argos, Potential of genetic algorithms in protein folding and protein engineering simulations, *Protein Eng.*, **5**, 637-645 (1992).

[5] R. Unger and J. Moult, Genetic algorithms for protein folding simulations, *J. Mol. Biol.*, **231**, 75-81 (1993).

[6] S. Sun, Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms, *Protein Sci.*, **2**, 762-785 (1993).

[7] M. J. Rooman, J.-P. A. Kocher and S. J. Wodak, Prediction of protein backbone conformation based on seven structural assignments, *J. Mol. Biol.*, **221**, 961-979 (1991).

[8] P. Argos, M. G. Rossmann and J. E. Johnson, A four-helical supersecondary structure, *Biochem. Biophys. Res. Commun.*, **75**, 83-86 (1977).

[9] M. Schiffer and A. B. Edmundson, Use of helical wheels to represent the structures of proteins and to identify segments with helical potential, *Biophys. J.*, **7**, 121-135 (1967).

[10] R. E. Stenkamp, L. C. Sieker and L. H. Jensen, Restrained least-squares refinement of *themiste dyscritum* methydroxohemerythrin at 2.0Å resolution, *Acta Cryst.*, **B38**, 784-792 (1982).

[11] F. Lederer, A. Glatigny, P. H. Bethge, H. D. Bellamy and F. S. Mathews, Improvement of the 2.5Å resolution model of cytochrome b_{562} by redetermining the primary structure and using molecular graphics, *J. Mol. Biol.*, **148**, 427-448 (1981).

[12] B. C. Finzel, P. C. Weber, K. D. Hardman and F. R. Salemme, Structure of ferricytochrome c' from rhodospirillum molischianum at 1.67Å resolution, *J. Mol. Biol.*, **186**, 627-643 (1985).

[13] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. jr. Meyer, M. C. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The protein data bank: A computer-based archival file for macromolecular structures, *J. Mol. Biol.*, **112**, 535-542 (1977).

[14] A. D. McLachlan, Gene duplications in the structural evolution of chymotrypsin, *J. Mol. Biol.*, **128**, 49-79 (1979).

[15] O. B. Ptitsyn and A. V. Finkelstein, Theory of protein secondary structure and algorithm of its prediction, *Biopolymers*, **22**, 15-25 (1983).

[16] P. Manavalan and P. K. Ponnuswamy, Hydrophobic character of amino acid residues in globular proteins, *Nature*, **275**, 673-674 (1978).

[17] F. Eisenmenger, P. Argos, and R. Abagyan, A method to configure protein sidechains from the mainchain trace in homology modelling, *J. Mol. Biol.*, **231**, 849-860 (1993).

[18] R. L. Baldwin and H. Roder, Characterizing protein folding intermediates, *Current Biology*, **1**, 218-220 (1991).

[19] G. Chelvanayagam, Z. Reich, R. Bringas, and P. Argos, Prediction of protein folding pathways, *J. Mol. Biol.*, **227**, 901-916 (1992).

[20] K. Kuwajima, The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure, *Proteins 6*, **87-103** (1989).

[21] H. S. Chan and K. A. Dill, Origins of structure in globular proteins, *Proc. Natl. Acad. Sci. USA*, **87**, 6388-6392 (1990).

[22] H. S. Chan and K. A. Dill, The protein folding problem, *Physics Today*, **46**, 24-32 (1993).

[23] J. Skolnik and A. Kolinski, Simulations of the folding of a globular protein, *Science*, **250**, 1121-1125 (1990).

[24] L. M. Gregoret and F. E. Cohen, Protein folding. Effect of packaging density on chain conformation, *J. Mol. Biol.*, **219**, 109-122 (1991).