METHOD FOR LOW RESOLUTION PREDICTION OF SMALL PROTEIN TERTIARY STRUCTURE

ANGEL R. ORTIZ, WEI PING HU, ANDRZEJ KOLINSKI, JEFFREY SKOLNICK*

Department of Molecular Biology, The Scripps Research Institute,
10666 N. Torrey Pines Rd., La Jolla, CA 92037

A new method for the *de novo* prediction of protein structures at low resolution has been developed. Starting from a multiple sequence alignment, protein secondary structure is predicted, and only those topological elements with high reliability are selected. Then, the multiple sequence alignment and the secondary structure prediction are combined to predict side chain contacts. Such contact map prediction is carried out in two stages. First, an analysis of correlated mutations is carried out to identify pairs of topological elements of secondary structure which are in contact. Then, inverse folding is used to select compatible fragments in contact, thereby enriching the number and identity of predicted side chain contacts. The final outcome of the procedure is a set of noisy secondary and tertiary restraints. These are used as a restrained potential in a Monte Carlo simulation of simplified protein models driven by statistical potentials. Low energy structures are then searched for by using simulated annealing techniques. Implementation of the restraints is carried out so as to take into account of their low resolution. Using this procedure, it has been possible to predict *de novo* the structure of three very diferent protein topologies: an α/β protein, the bovine pancreatic trypsin inhibitor (6pti), an α-helical protein, calbindin (3icb), and an all β- protein, the SH3 domain of spectrin (1shg). In all cases, low resolution folds have been obtained with a root mean square deviation (RMSD) of 4.5-5.5 Å with respect to the native structure. Some misfolded topologies appear in the simulations, but it is possible to select the native one on energetic grounds. Thus, it is demonstrated that the methodology is general for all protein motifs. Work is in progress in order to test the methodology on a larger set of protein structures.

1.- Introduction

Prediction of the three dimensional structure of a protein from its amino acid sequence is still one of the most important unsolved problems in contemporary molecular biology. Although protein structure determination by experimental methods has become more efficient, the ratio between the number of known sequences and the number of known structures is rapidly increasing. An advantage of this fast growth of protein sequence databases is that many sequences can be grouped into structurally conserved families (Sander & Schneider, 1991). This increasing number of protein families for which many homologous proteins are known affords a new opportunity to exploit evolutionary information. When aligned together, such families exhibit features of residue conservation that are directly related to their three dimensional structure. Using this principle, multiple aligned sequences in a family of homologous proteins have been used recently to improve the prediction of secondary structure in proteins (Rost & Sander, 1993). Prediction of contact maps has also been attempted by analyzing correlated mutations in multiple sequence analysis, showing interesting results (Göbel *et al.*, 1994). Combination of predictions based on multiple sequence alignments with structure calculations are starting to emerge aimed at predicting simple protein topologies (Hänggi & Braun, 1994; Numenthaler & Braun, 1995; Aszódi *et al.*, 1995). Along these lines, we investigate the possibilities of predicting low resolution structures of small proteins using restraints derived from multiple sequence alignments and Monte Carlo simulations. Two significant differences can be found with respect to previous studies: firstly, the signal-to-noise ratio of the predicted contacts is increased by means of a new two-step procedure (Ortiz *et al.*, 1996) that first detects pairs of topological elements in contact using correlated mutation analysis and then expands the number of contacts by using inverse folding. Secondly, a robust lattice Monte Carlo simulation, the MONSTER method (Skolnick *et al.*, 1996), is used to generate the protein model. The MONSTER method can efficiently find native-like structures with a RMSD of 3-4 Å using N/5 contact map restraints, where N is the number of residues in the chain. This approach has been applied to a small set of topologically different proteins, and preliminary results of its performance are described here.

2.- Methods

A flow chart of the procedure can be found in figure 1. It can be divided in two separate steps: restraint derivation and structure assembly, which are described in detail in the following paragraphs.

Restraint derivation method

Secondary structure prediction

Multiple sequence alignments for each of the proteins studied were obtained from the HSSP data base (Sander & Schneider, 1991). This alignment was used as input for the PHD method of secondary structure prediction (Rost & Sander, 1993). From the output, all elements predicted as strands were assumed to correspond to a strand in the real secondary structure. For the helices, only those elements with a reliability index higher than 3 were used. Chain reversals were then predicted by using the "U-turn" algorithm (Kolinski *et al.*, 1996). Elements predicted as turns override PHD predictions, as "U-turn" prediction has been proven to be highly reliable (Kolinski *et al.*, 1996).

Side-chain contact prediction

The prediction of residue contacts is performed in two stages (Ortiz *et al.*, 1996): first, a correlated mutation analysis of the multiple sequence alignment is carried out. Then, inverse folding is used to "enrich" the number of contacts. For details about the correlated mutation analysis, see Göbel *et al.* (1994). Briefly, the method is based on defining an exchange matrix or other similarity measure at each sequence position in a multiple alignment and then calculating a correlation coefficient between the exchange matrices at any two positions. In this work, the same multiple alignment used for secondary structure prediction was used, and only correlations between elements to be predicted in secondary structural regions (and not "U" turns) are considered. The rationale is that in this way it is possible to restrict the predictions to rigid elements of the core, for which the assumption of closeness in space for positions showing covariance in their mutational behavior is, in principle, more valid. This analysis delineates predicted secondary structure elements in contact. A relatively high cutoff for the correlation coefficient of 0.5 is used for contact prediction.

Correlated mutation analysis only provides a few points in the contact map, which we call "seeds", and usually do not provide enough restraint information. The set of restraints is then enriched by using inverse folding. This is based on the observation that the pairing of secondary structure elements in proteins is highly

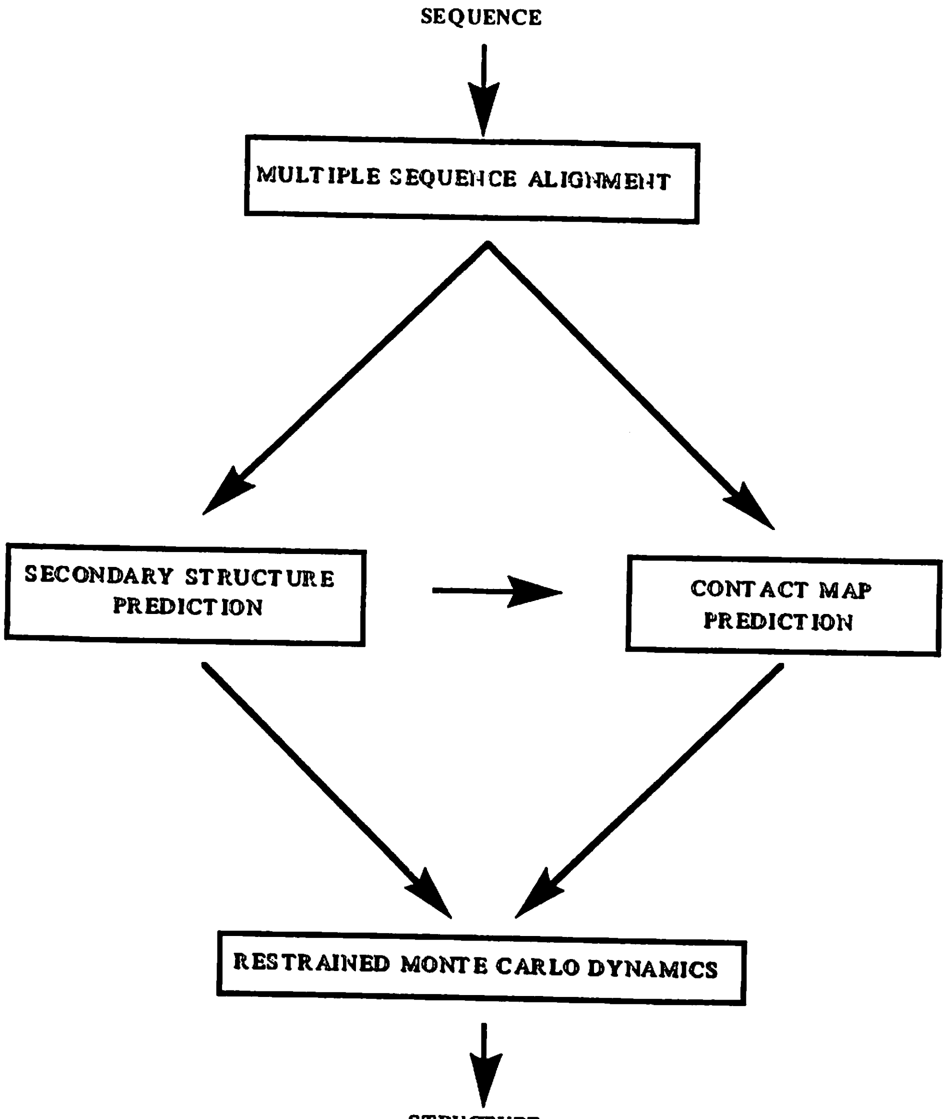

SEQUENCE
MULTIPLE SEQUENCE ALIGNMENT
SECONDARY STRUCTURE PREDICTION
CONTACT MAP PREDICTION
RESTRAINED MONTE CARLO DYNAMICS
STRUCTURE

Figure 1. Flowchart of the protein structure prediction method.

degenerate; that is, there is a limited number of ways of pairing secondary structure elements. Thus, pairs of secondary structure elements predicted to be in contact are scanned by using inverse folding, subjected to the constraint that the predicted contact should be present. A tolerance of ± 1 residue shift in each one of the members of the contact pair is allowed in order to take into account the inaccuracies in the correlated mutation analysis. Fragments are then clustered and scored. Fragments are superimposed in space by minimizing their corrdinate RMS deviation, and if they do not show a clear clustering, with an upper limit of 5 Å for the most divergent pair of fragments, the pair of fragments is discarded, and side chain restraints are not derived. On the other hand, if the fragments do cluster, they are scored by the inverse folding potential. The average fragment of the cluster is used as a representative, and a selection is done among the fragments that fulfill the constraints by means of the inverse folding score: the lowest scoring fragment is selected, and then the contact map for the selected fragment is projected onto the query sequence. With this procedure, the number of predicted contacts usually increases by about five times with respect to the number of contacts predicted from the correlated mutation analysis.

The MONSTER Method

The protein model is based on a lattice representation of the Cα-backbone trace. The details of the model have been described previously (see Kolinski & Skolnick, 1996 for a review). Here, we give a brief summary for the reader's convenience. The procedure is divided in two steps: assembly and refinement. Slightly different protein representation is used in each one of the cases. Where appropriate, distinction will be made about the differences in implementation.

Lattice model of protein chain

The Cα backbone is a string of vectors of the type a.v with {v}={(3,1,1),... (3,1,0),... (3,0,0),... (2,2,1),... (2,2,0),....}, with a=1.22 Å. The lattice resolution is about 0.6-0.7 Å RMS (Godzik *et al.*, 1993). Virtual bond angles between successive Cαs are restricted to reproduce a protein-like distribution, and two successive backbone vectors provide the reference frame for the definition of a set of model (single interaction center for each side group) rotamers that cover the conformational space of the side chains with a 1 Å grid. A "rotamer" located at 0.3 Å from the Cα is assumed for glycine.

Force field of the protein model

The force field contains potentials of mean force (predominately of statistical origin) that account for the short range interactions, long range interactions, and hydrogen bond interactions (which could be short or long range). The total energy is given by:

$$E= 0.5E_{14} + 1.5E_{1} + 2.75E_{pair} + E_{H\text{-}bond} + E_{target,14} + V_{long} \quad (1)$$

All contributions to the potential are available via anonymous ftp (Skolnick, 1996), and detailed account of each one of the terms can be found in Skolnick *et al.* (1996a). Here only a brief description is given:

1.- Local conformational propensities (E_{14}) : This is a sequence dependent term divided into six conformational states that roughly correspond to extended right, and extended left handed states, wide right and left turns, and right and left handed helical states respectively. This component depends on the sequence through the pair of neighboring amino acids A_i and A_{i+1} , and controls the local chain geometry.

2.- One-body term (E_1): Centrosymmetric potential which reflects the tendency of some amino acids to be buried and some to be exposed. It is dependent of the expected radius of gyration of a single domain protein consisting of N amino acids in its native conformation ($S_0 = 2.2\ N^{0.38}$ in Å). This potential is applicable only to single domain proteins.

3.- Pairwise tertiary interactions (E_{pair}): The pair interactions beyond the fourth neighbor are derived from the statistics of the database, and neglected between nearest neighbors in sequence, assuming that these interactions are already accounted for by the hydrogen bond potential (see below) and secondary structure preferences.. For residues i and j,

$$E_{ij} = \begin{array}{ll} E^{rep}, & \text{for } r_{ij} < R_{ij}^{rep} \\ e_{ij}, & \text{for } R_{ij}^{rep} < r_{ij} < R^{con}_{ij}, \text{ and } e_{ij} > 0 \\ fe_{ij}, & \text{for } R_{ij}^{rep} < r_{ij} < R^{con}_{ij}, \text{ and } e_{ij} < 0 \\ 0 & \text{Otherwise.} \end{array} \qquad (2)$$

where: r_{ij} is the distance between the side chain centers of mass, R_{ij}^{rep} and are R^{con}_{ij} the cut-off values for hard core excluded volume interactions and for square-well, soft pairwise interactions, respectively. The amino acid pairwise specific parameters, e_{ij} are described elsewhere (Skolnick *et al.*, 1996). In order to facilitate assembly, the magnitude of this interaction for a repulsive pair of residues is decreased by a factor of two. E^{rep} is the repulsive energy associated with the overlap of alpha carbons and side chains, and it is on the order of $5k_BT$. f is an angular factor which weakly favors almost parallel or almost antiparallel orientations of the secondary structure elements such as occur in all beta and mixed α/β proteins (Skolnick *et al.*, 1996).

3.- Hydrogen bonds ($E_{H\text{-}bond}$): It operates only between α-carbons, following an scheme very much in the spirit of Levitt and Greer (Levitt & Greer, 1977). Each α-carbon can participate in at most two hydrogen bonds (the α-carbon of proline is an exception and can participate in only one hydrogen bond), and there is no directionality (donor-acceptor) in the scheme. This scheme reproduces 90 % of the main chain hydrogen bonds assigned to the structure by DSSP Kabsch-Sander algorithm (Kabsch & Sander, 1983). Explicit cooperativity is also introduced. When two neighboring pairs of residues are hydrogen bonded, the system gets an additional favorable energy (cooperativity). Note that we ignore all side chain - side chain and side chain - backbone hydrogen bonding. Hydrogen bond energy is modulated depending upon the hydrophobicity of the residues under consideration. For hydrophobic residues, it also depends on whether or not the residue is buried or exposed. In refinement, a different model for the short range interactions and a different hydrogen bond scheme are used. Here, explicit coordinates for the protein backbone are used, and thus orientational correlations of the peptide bond plates are simulated in the short range term of the potential. The short range interactions are of two types: generic (sequence independent) and sequence specific. The role of sequence independent potentials is to provide a strong bias towards short and intermediate range correlations that mimic conformational regularities in proteins. These include a bias towards the right handed geometry of helix-like fragments, a slight bias toward left handed expanded states, and terms reflecting the conformational stiffness of more regular secondary structure fragments, regarless of amino acid composition. The hydrogen bond scheme takes the form of a simplified Coulombic potential that is moderated by an angular factor. Details of the implementation and relative scaling with respect to the long range term of the potential can be found in Kolinski *et al.* (1995a, 1995b).

Restraint Contributions

Secondary structure restraints and a limited number of tertiary restraints are used. Furthermore, a set of knowledge-based restraints is used. The implementation of each kind of restraint is discussed in turn.

Short range restraints

1. For those residues assigned to be helical, hydrogen bonds beyond the fifth residue along the chain are not allowed. Similarly, a β-assigned residue cannot have a helix hydrogen bond pattern nor can it hydrogen bond to a residue that has been assigned to be in a helical region of the molecule.

2. A given residue can be in one of five conformational states, assigned on the basis of the local chain geometry. For those residues which have an assigned secondary structural type, energetic biases for the various allowed conformational states are assigned. Left handed helices experience a repulsion of 1., in k_BT units. Turns are encoded on a generic basis, i.e., their chirality is not specified. Rather they behave as flexible joints between regular secondary structural elements. See Skolnick *et al.* (1996) for details. The resulting background target local conformational energy is

$$E_{target,14} = \sum_{k=2}^{N-2} 2.5\mu(k) \sum_{j=1}^{5} \sec(k,j) + (1-\mu(k)) \sum_{j=1}^{6} ag(k,j) \quad (3)$$

where μ(k)=1 if the secondary structure is assigned *a priori*, and it is equal to zero otherwise. ag(k,j) is an amino acid pair specific matrix describing local interactions, that depends on the identity of A_k and A_{k+1}. It is defined with respect to the six conformational bins, left and right hand extend/beta states, left and right hand wide turns, and left and right handed helices/tight turns, and acts propagating secondary structure elements. sec(k,j) contains the local energetic bias for a particular conformational bin, as predicted from the secondary structure prediction procedure.

Long range restraints

Long range restraints operate on the level of distances between the centers of mass of the side groups. Due to the fuzzy representation of the side chains in our model and the ambiguity in the restraint derivation, long range restraints are implemented as follows: If residues i and j are predicted to be in contact, then the residue based pair potential of eq.2 is modified so that e_{ij}=-1.25. The long range tertiary restraints are as follows. Let $d_{ij}=r_{ij}-R^{con}_{ij}$

$$V_{long}(r_{ij}) \begin{cases} =0 & \text{if } r_{ij} < R^{con}_{ij} \\ =g\, d_{ij}^{\;2} & \text{If } R^{con}_{ij} < d_{ij} < 34.5\ \text{Å} \\ =g(34.5)^2 & \text{otherwise.} \end{cases} \qquad (4)$$

Typically, the value of g ranges from 0.5 to 2. In folding from random compact states, the restraints are not implemented simultaneously. Rather, the "sequential growing strategy" is used. This appears to increase the folding efficiency by decreasing the extent of kinetic trapping.

Knowledge-based restraints

Finally, for one of the proteins tested (6pti, see below), knowledge based information about protein topology is used (Skolnick *et al.*, 1996). This knowledge based information acts to reduce the number of misfolded structures. In particular, it helps to eliminate the problem of topological mirror image states, and thereby enhances the folding efficiency. By a topological mirror image, we refer to structures where the chirality of the secondary structure connections is reversed (Pastore *et al.*, 1991).

Conformational sampling

The sampling of conformational space occurs via a standard asymmetric Monte Carlo Metropolis scheme (Metropolis *et al.*, 1953). The conformational updates are composed of several types of local conformational micromodifications of the chain backbone and their associated side groups, side group equilibration cycles, and rare (small distance) motions of larger chain fragments.

Fully extended chains are selected as initial structures in the assembly stage. Each simulation started at a reduced temperature in the range of 2.0, and then the temperature is slowly lowered to 1.0. The final structure obtained from the assemply step is then subjected to refinement, again using simulated annealing with a temperature range of 2.0-1.0. Typically, three kinds of structures result: There are misfolded states of higher energy that can be trivially dismissed; the correct native like folds with various root mean square deviations of the α-carbons from native; and the topological mirror image folds. These pseudo mirror image folds could be identified in two ways. First, they may exhibit an *a priori* violation of the known connectivity handedness rules for some supersecondary connections. Second, the average energies of the mirror image structures are higher than that for the correct folds. The predicted structure is the one which exhibits the lowest average and minimum energies.

Test cases studied

Three different test cases have been studied (Table 1), chosen as small protein representatives of different structural classes, in order to explore the methodology with different protein motifs: an α/β protein, the bovine pancreatic trypsin inhibitor (6pti); an α-helical protein, calbindin (3icb); and an all β-protein, the SH3 domain of spectrin (1shg). In all cases the starting conformation was a fully extended chain, and ten simulations for each protein were carried out.

Table 1. Summary of results of the folding simulations (for the lowest energy structures).

Prot[a]	Nres[b]	Nsq[c]	Npc[d]	Ntc[e]	rms_r[f]	rms_n[g]	E_n[h]	rs_n[i]	rms_w[j]	E_w[k]	rs_w[l]
6pti	56	45	19	92	4.7	6.0	-410	19	9.7	-397	18
3icb	75	67	25	154	4.5	6.8	-406	21	12.6	-342	11
1shg	57	20	39	109	4.5	4.5	-198	19	8.22	-125	17

[a] *Prot* refers to the pdb access number of the protein studied.

[b] *Nres* is the number of residues of the protein in the pdb file.

[c] *Nsq* stands for the number of sequences aligned in the HSSP (Sander & Schneider, 1991) multiple sequence alignment file.

[d] *Npc* corresponds to the number of predicted contacts used in the simulation.

[e] *Ntc* is the number of side chain contacts in the experimental structure.

[f] rms_r is the rms deviation with respect to the experimental structure of the final predicted structure after refinement (measured in Å).

[g] rms_n corresponds to the rms deviation of the computed structure to the experimental structure after the assembly runs for the native-like topology (measured in Å).

[h] E_n is the total energy of the protein after the assembly run for the native-like topology (measured in k_BT units).

[i] rs_n refers to the number of restraints satisfied in the assembled native-like structure.

[j] rms_w corresponds to the rms deviation of the alternative topology of lowest energy with respect to the experimental structure, after the assembly run (measured in Å).

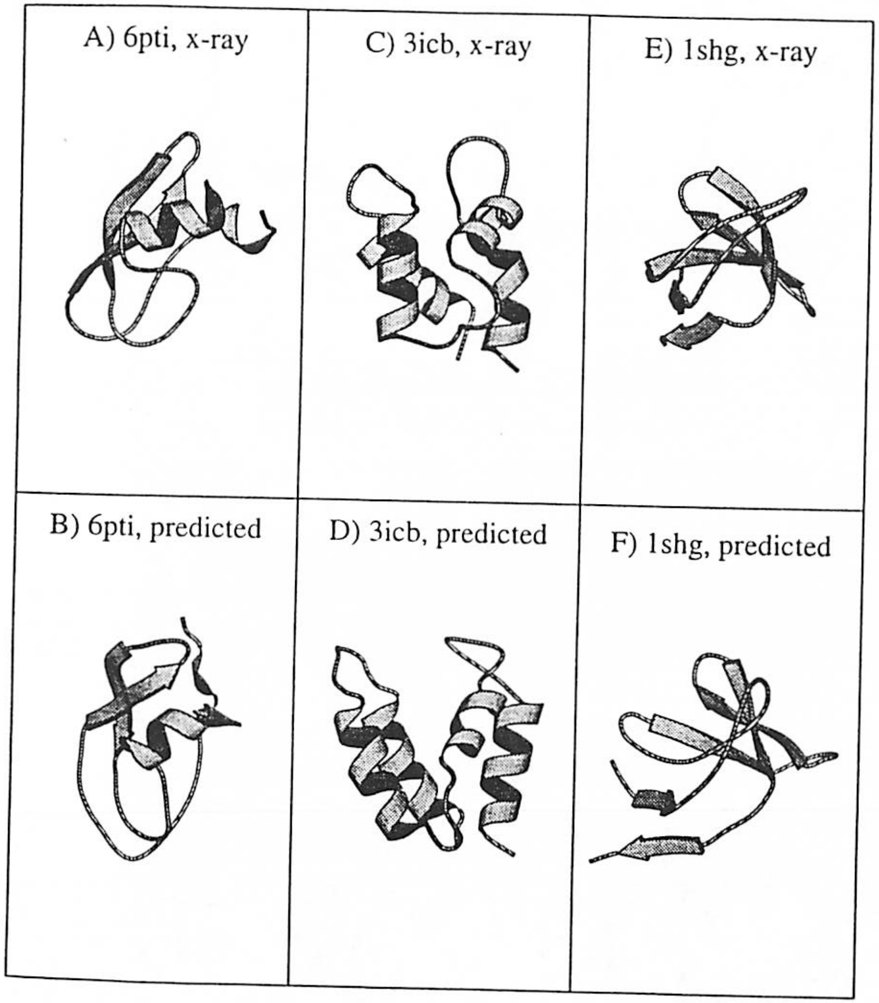
[k] E_w is the total energy of the protein in the lowest energy alternative topology found, after assembly runs (measured in k_BT units).

[l] rs_w stands for the number of restraints satisfied in the alternative topology after assembly runs.

3.- Results and Discussion.

6pti

6pti is a small α/β protein of 57 residues (Table 1). The structure of 6pti (figure 2a) consists of a small α-helix, involving residues 48 to 55, packed against a twisted β-hairpin, associated with residues 18-35. These elements are connected by long loops. There are 3 disulfide bridges in the structure. The secondary structure prediction correctly locates the elements of secondary structure, including the turn in the β-hairpin, altough there is a shift and slight overprediction in the first strand of the β-hairpin, and it misses the C-terminal part of the helix (figure 3a). The prediction accuracy for this protein is 80.4 %. The contact map prediction predicts 19 contacts (Table 1). They locate the strand pairing and the packing of the α-helix against the β-hairpin, although there is a shift in the residues involved in the contact of the secondary structure elements.



Assembly Monte Carlo simulations locate native-like topologies with an RMSD of about 6 Å with respect to the experimental structure (Table 1). However, another alternative topology appears, corresponding to the topological mirror image of the native structure (Pastore *et al.*, 1991). Distinction between both topologies can be made in this case on the basis of the force field energy and the number of restraints fulfilled (Table 1). Refinement runs allow to obtain structures with an RMSD of 4.7 Å with respect to the experimental structure. Figure 2b shows the predicted structure. The general fold topology is well reproduced. However, there are still wrong details in the predicted structure. The most salient ones are the shift in the helix to hairpin packing and the wrong twist of the β-hairpin. On the other hand, it is worth noting that the simulations are able to correct the wrong prediction of the C-terminal helix as a coil, and thus in the predicted structure, it forms part of the α-helix, as in experiment.

3icb

3icb is a four helix bundle (figure 2c) of 75 residues (Table 1). The secondary structure prediction successfully locates the four helices. However, the PHD method overpredicts the number of residues in helical conformation for the second helix, but the "U-turn" prediction algorithm detects a clear "U-turn" in that region and hence successfully corrects the PHD prediction (figure 3b). The contact map prediction detects 25 side chain contacts (Table 1). They correspond to contacts between helices II and III; III and IV; and I and IV. Assembly runs typically produce structures in the range of 6 Å with respect the experimental structure. The predicted structure of 3icb, after refinement, can be found in figure 2d. The RMSD of the final refined conformation is 4.5 Å with respect to the experimental structure (table I). The main differences with respect to the native structure are the different conformations of the loops, the more open structure in the calcium binding loops face, the more parallel arrangement of the third helix, and the lack of bend in the first helix. The alternative mirror image topology appears frequently in this case, but again can be dismissed on the basis of the number of restraints fulfilled and the force energy (Table 1).

1shg

1shg is a small all β-protein (figure 2e) of 57 residues (Table 1), consisting of 5 strands arranged so as to form a barrel structure. Because of its all beta topology and the presence of bends in the strands in order to configurate the barrel, together with the long loops connecting secondary structure elements, this protein represents a considerable challenge for the methodology. The secondary structure prediction is shown in figure 3c. It is worth noting that in this case the prediction accuracy is only 65 %. There are several differences between the predicted secondary structure and the observed one: the number of residues in the first strand is overpredicted. Furthermore, there is an overprediction of a strand between residues 18 and 20. Aditionally, the small helix between 50 and 52 is not predicted, and the two last strands are considerably shortened. The "U-turn" prediction also shifts the prediction of some turns by about 2-3 residues. The number of predicted contacts in this case is 17 (Table 1). Succesful assembly simulations for this structure provide folds in the order of 6-7 Å RMSD with respect the experimental structure. However, a considerable number of misfolded alternative topologies appear, most cases are trapped intermediates with wrong pairings of strands, or structures with almost parallel angles between between the two β-sheets. In this case the force field energy averaged over short runs does not discrimate between the alternative anwers and the correct topology, perhaps because the structures obtained after assembly are still relatively away from the experimental structure in conformational space, and because of the shifts in the "U-turns" and predicted contacts introduce a considerable distortion in theresidue environment, which reflects in the computed energy. The slightly higher number of restraints satisfied by the correct topology does not provide in this case clear criteria for selecting the right topology unambiguously. However, it was possible to select it on the basis of long isothermal runs. The results are reported in Table I, where can be observed that the correct topology can be identified. The refined structures provide folds with a RMSD of 4.5 Å with respect to experiment. Figure 2f shows the predicted structure. The main differences with the experimental structure come from the positioning of the loops and the angles between the strands. On the other hand, strands are highly "idealized" and they do not show the bend observed in the experimental structure.

```
A)  protein:      6pti              length       56    Q'=80.4

   ....,....1....,....2....,....3....,....4....,....5....,....6
AA|RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCG|
OB|  HHHH           EEEEEEE    EEEEEEE           HHHHHHHH |
PR|55555555555533444444444443334444444555555555555222222555|

B)  protein:      3icb              length       75    Q'=89.3

   ....,....1....,....2....,....3....,....4....,....5....,....6
AA|KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKGPSTLDELFEELDKNGDGE|
OB|  HHHHHHHHHHHH          HHHHHHHHHHH        HHHHHHHH       |
PR|552222222222555111111111222222222222333335552222222225533336|

   ....,....7....,....8....,....9....,....10...,....11...,....12
AA|VSFEEFQVLVKKISQ|
OB|  HHHHHHHHHHHH |
PR|552222222222255|

C)  protein:      1shg              length       57    Q'=64.9

   ....,....1....,....2....,....3....,....4....,....5....,....6
AA|KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD|
OB|  EEEE                 EEEEEE     EEEEEE   EEEEEEHHHEEEE|
PR|554444444555333354445533444445533335544555535444555544555|
```

Figure 3. A) Secondary structure assignment for 6pti; B) Secondary structure assignment for 3icb; C) Secondary structure assignment for 1shg. "AA" refers to the aminoacid sequence of the protein; "OB" is the observed secondary structure in the experimental conformation, according to the DSSP assignment; "PR" is the assigned secondary structure state in the folding simulations, according with the prediction results (see Methods). In the numeration scheme adopted, "1" stands for coil assignment; "2" for helix assignment; "3" for turn assignment"; "4" for strand assignment and "5" correspond to no assignment of secondary

4.- Conclusions.

Low resolution folds of small proteins have been obtained using a hierarchical approach that starts from multiple sequence alignments. For the test cases presented, the method is able to predict reliably the correct fold with an accuracy of 4.5-5.5 Å. Certainly, a larger test set of proteins is needed in order to correctly assess the methodology. Such work is now in progress. Still, even demonstrating the generality of the method, considerable improvements are required. It is important is to achieve a higher yield of successful simulations, which currently is in the order of 30 % for the cases studied. It is also crucial to select the correct topology from the misfolded alternatives. The results with 1shg suggest that long isothermal runs may be required to obtain an adequate energy spectrum that allows reliable identification. In any event, the approach presented here seems to be a promising one for predicting protein structure from sequence.

References

Aszódi, A., Gradwell, M. J. & Taylor, W. R. (1995). *J. Mol. Biol.* **248**, 308-326.
Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). *Proteins*, **18**, 309-317.
Godzik, A., Kolinski, A. & Skolnick, J. (1993) *J. Comp. Chem.* **14**, 1194-1202.
Hänggi, G. & Braun, W. (1994). *FEBS Letters* 344, 147-153.
Kabsch, W. & Sander, C. (1983). *Biopolymers* **22**, 2577-2637.
Kolinski, A. *et al.* (1995). *J.Chem.Phys.* **103**, 4312-4323.
Kolinski, A., Galazka, W., Skolnick, J. (1995). *J.Chem.Phys.* **103**, 10286-10297.
Kolinski, A. & Skolnick, J. (1996). *Lattice Models of Protein Folding, Dynamics and Thermodynamics*, R. G. Landes Company, Austin, TX.
Kolinski, A., Skolnick, J. & Godzik, A. (1996). *Proteins* (submitted).
Kraulis, P. (1991) *J. Appl. Cryst.* **24**, 946-950.
Levitt, M. & Greer, J. (1977). *J. Mol. Biol.* **114**, 181-293.
Metropolis, N. *et al.* (1953). *J.Chem.Phys.* **51**, 1087-92.
Numenthaler, C. & Braun, W. (1995). *Prot. Science* 4, 863-871.
Rost, B. & Sander, C. (1993). *J. Mol. Biol.* **232**, 584-599.
Sander, C. & Schneider, R. (1991). *Proteins* 9, 56-68.
Skolnick, J., Kolinski, A. & Ortiz, A. R. (1996). *J. Mol. Biol.*. (submitted).
Skolnick, J. (1996). ftp.scripps.edu in directory /pub/skolnick/nmr.
Skolnick, J. *et al.* (1996). *Prot. Engn.*, (submitted).
Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). *Prot. Engn.* **6**, 605-614.
Ortiz, A. R., Hu, W. P. & Skolnick, J. (1996). *Unpublished results.*
Pastore, A. *et al.* (1991). *Proteins*, 10, 22-32.
Wlodawer, A. *et al.* (1987). *J. Mol. Biol.* **198**, 469-480.