Pacific Symposium on Biocomputing 4:505-516 (1999)

A COMBINED APPROACH FOR AB INITIO CONSTRUCTION OF LOW RESOLUTION PROTEIN TERTIARY STRUCTURES FROM SEQUENCE

RAM SAMUDRALA [a], YU XIA, MICHAEL LEVITT
Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305, USA

ENOCH S HUANG
Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Saint Louis, MO 63110, USA

An approach to construct low resolution models of protein structure from sequence information using a combination of different methodologies is described. All possible compact self-avoiding C_α conformations ($\approx$ 10 million) of a small protein chain were exhaustively enumerated on a tetrahedral lattice. The best scoring 10,000 conformations were selected using a lattice-based scoring function. All-atom structures were then generated by fitting an off-lattice four-state ϕ/ψ model to the lattice conformations, using idealised helix and sheet values based on predicted secondary structure. The all-atom conformations were minimised using ENCAD and scored using a second hybrid scoring function. The best scoring 50, 100, and 500 conformations were input to a consensus-based distance geometry routine that used constraints from each the conformation sets and produced a single structure for each set (total of three). Secondary structures were again fitted to the three structures, and the resulting structures were minimised and scored. The lowest scoring conformation was taken to be the "correct" answer. The results of application of this method to twelve proteins are presented.

1 Introduction

The prediction of protein three dimensional structure from sequence alone with accuracy rivalling that of experiment is an unsolved problem. However, for certain classes of small globular proteins, it is possible, in some cases, to computationally generate low resolution models of a sequence ($\approx$ 6 Å C_α root mean square deviation of the coordinates (cRMSD) from the experimental structure)[1,2]. As electron microscopists have demonstrated, even low resolution models can yield valuable insights about the function of a protein. Given the large number of sequences being determined and the relatively slow progress of protein structure prediction methods, low resolution models generated by current approaches can be used to elucidate details about structure and function for proteins whose atomic structure has not been determined experimentally.

[a] Corresponding author; E-mail: ram@zen.stanford.edu

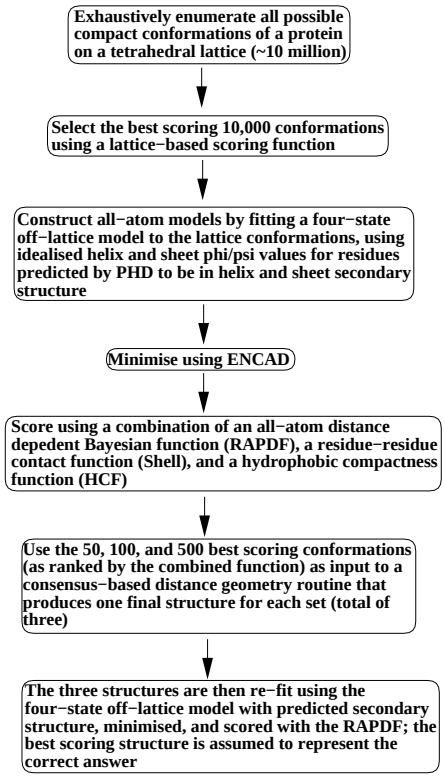

Figure 1: Flow chart describing the methodology used in this work. Many of these steps can be carried out in parallel via a pipeline.

In this work, we use a combination of approaches described in the literature, and primarily developed in-house, to construct tertiary models of protein sequences that have the correct topological arrangement of secondary structure elements (see Figure 1). A detailed description of the individual components of the combined approach follows.

2 Methods

2.1 Lattice enumeration and scoring on-lattice

Protein topology is captured by a self-avoiding tetrahedral lattice walk where each vertex represents 1-4 residues depending on the size of the protein, with a

maximum walk length of 38. (For a full description, see Hinds and Levitt, 1992 & 1994[3,4].) Lattice spacing between vertices is scaled based on the mean C_α-C_α distance obtained from a database of protein conformations. We enumerate all possible lattice walks that fit within a predefined elliptical bounding volume containing 20% to 50% more vertices than will be used by any particular structure. We pick out lattice walks that are reasonably compact to have radius of gyration of up to 1.14 times that of a sphere with the same volume.

To obtain a model, a lattice walk is threaded with the target protein sequence such that no more than three residues are positioned between each pair of lattice points along the walk and that each lattice point is occupied by a specific residue. The score for this structure is evaluated using a residue-residue contact function derived from pairwise amino acid contact frequencies in a database of experimentally determined structures. We count residue-residue contacts in a lattice structure in such a way that total numbers of long-range contacts in lattice and actual structures are roughly similar.

For every compact lattice walk, an iterative dynamic programming procedure is used to identify one threading arrangement with residue assignment to lattice points such that a locally optimal pattern of tertiary interactions is formed. The score for this structure is calculated, and this procedure is repeated for all compact lattice walks. The 10,000 best scoring structures were selected as templates for building all-atom models.

2.2 Secondary structure prediction

Sequences to be modelled were submitted to the PHD PredictProtein Server (`predictprotein@embl-heidelberg.de`)[5] and the results returned from the server were used as-is, without further tuning of the multiple sequence alignments or the predictions.

2.3 Secondary structure fitting and all-atom model generation

The low-resolution tetrahedral lattice model only captures overall protein chain topology, and completely lacks fine detail of secondary structures. In order to build an all-atom model, we fit predicted secondary structures to the 10,000 best scoring lattice conformations using an off-lattice four-state ϕ/ψ model and a brute force build-up algorithm[6].

The conformation for a given residue is specified by a set of ϕ, ψ, and χ angles. The ϕ/ψ states used for the four-state model were taken from model B in Park & Levitt, 1995[6]: (-57/-47), (-129,124), (-36,108), (108,-36). The χ angles for side chains were fixed to those most frequently observed in a

database of protein structures[7]. All bond lengths and bond angles were set to idealised values.

Residues predicted to assume helix or sheet with high confidence (> 5) were assigned idealised helix and sheet ϕ/ψ values (the first two values in the four-state model). For all other residues, starting at the N terminus of the protein, we enumerated all possible conformations for the first ten non-fixed residues using the off-lattice model, saving 600 conformations which have lowest cRMSD relative to the corresponding C_α atoms of the lattice structure. Then we added an additional residue at the C terminus of each of the 600 saved conformations in all four possible states ($600 \times 4 = 2400$ conformations) and again saved the 600 conformations with the lowest cRMSD deviation from the corresponding C_α atoms of the lattice structure. This iterative procedure was repeated for all residues until the entire lattice model was fitted.

2.4 Off-lattice scoring functions

We used a scoring function that combined scores produced by three different functions: an all-atom distance-dependent conditional probability discriminatory function (RAPDF), a hydrophobic compactness function (HCF), and a residue-residue contact function (Shell). The scores were combined after being divided by the respective standard deviation calculated over 10,000 conformations.

Residue-specific all-atom probability discriminatory function (RAPDF)

The all-atom scoring function, RAPDF, was used to calculate the probability of a conformation being native-like given a set of inter-atomic distances[8]. The conditional probabilities were compiled by counting frequencies of distances between pairs of atom types in a database of protein structures[b]. All non-hydrogen atoms were considered, and a residue-specific description of the atoms was used, i.e., the C_α of an alanine is different from the C_α of a glycine. This resulted in a total of 167 atom types. The distances observed were divided into 1.0 Å bins ranging from 3.0 Å to 20.0 Å. Contacts between atom types in the 0-3 Å range were placed in a separate bin, resulting in a total of 18 distance bins. Distances within a single residue were not included in the counts.

We compiled tables of scores proportional to the negative log conditional probability that one is observing a native conformation given an an interatomic distance for all possible pairs of the 167 atom types for the 18 distance ranges.

[b] A set of 312 unique folds from the SCOP database[9] was used.

Given a set of distances in a conformation, the probability that the conformation represents a "correct" fold was evaluated by summing the scores for all distances and the corresponding atom pairs. A complete description of this formalism has been published elsewhere [8].

Hydrophobic compactness function (HCF)

The Hydrophobic compactness function (HCF) score for a given conformation is calculated using the formula:

$$HCF = \frac{\sum_i^N (\bar{x} - x_i)^2 + (\bar{y} - y_i)^2 + (\bar{z} - z_i)^2}{N} \quad (1)$$

where N is the number of carbon atoms in the protein, and x, y, and z are the three-dimensional coordinates of those atoms. This measure is the square of the radius of gyration of the carbon atoms.

Residue-residue contact function (Shell)

The Shell scoring function is described in detail elsewhere [10]. Briefly, it is a simple pairwise contact function with the form:

$$E = \sum_{i=1} \sum_{j>i+1} e_{ij}^{ab} \quad (2)$$

where e is the contact score for residues i and j of types a and b, respectively. $e_{ij}^{ab} = e^{ab}$ if $d_{ij} < 7.0$ Å and zero otherwise. All inter-residue distances d_{ij} were measured from an interaction center located 3.0 Å from the C_α atom along the C_α-C_β vector.

$$e^{ab} = -\ln n_{obs}^{ab} / n_{exp}^{ab} \quad (3)$$

where n_{obs}^{ab} is the number of residue types a and b within 7.0 Å in a database of proteins. n_{exp}^{ab} is the number of contacts expected in a random mixture of residue types in the database:

$$n_{exp}^{ab} = \sum_p \frac{C_p \times 2R_p^{ab}}{(N_p - 2) \times (N_p - 1)} \quad (4)$$

For each protein p, C_p is the total number of contacts, R_p^{ab} is the number of residue pairs of type a and b separated by at least two residues in the sequence, and N_p is the number of residues.

2.5 Consensus-based distance geometry

Restraints for metric matrix distance geometry were taken directly from the best scoring conformation sets. Each inter-C_α distance was measured and stored in 1 Å bins. The upper and lower bounds for a given C_α-C_α distance were determined by a jury process. Each distance received a weight equal to the Boltzmann weight of the structure from which it was measured, i.e.

$$W_i = \frac{exp(-E_i/kT)}{Q} \tag{5}$$

where E is the score of fold i, and Q is the partition function:

$$Q = \sum_i exp(-E_i/kT) \tag{6}$$

Here, kT was set to 10. In the jury process, the distance bin that received the most Boltzmann-weighted votes was used to set the upper and lower bound for a given C_α-C_α distance.

Distance geometry calculations were performed with the program distgeom from the TINKER suite. Structures were generated using 10% random pairwise metrization. Efficient metrization was achieved via a fast shortest path update algorithm used to re-smooth the lower and upper bounds matrices every time a trial inter-atomic distance is chosen. Trial distances were selected from approximately Gaussian distributions between the lower and upper bounds. The center of the distribution between the upper and lower bounds is a function of the number and type of input restraints and is consistent with the expected radius of gyration of the structure. Following metrization, embedding and majorization, the generated structure is refined via 10,000 steps of simulated annealing against a set of penalty functions which enforce local geometry, chirality, excluded volume, and the input distance restraints. A full description of this method is given in Huang, *et al* [11].

2.6 Minimisation procedures and generation of final models

All-atom models generated after the fitting procedure were minimised for 200 steps using ENCAD [12,13,14,15]. For each protein, the three structures from the consensus-based distance geometry were minimised for 2000 steps after fitting using both high confidence and complete secondary structure assignments by PHD, as described previously. The conformation with the lowest score, as evaluated by the all-atom scoring function RAPDF, was taken to represent the final selection.

2.7 Handling mirror images

The lattice enumeration procedure only generates low resolution C_α structures with no secondary structure information, and as a result, for a given lattice walk, a structure and its mirror image cannot be distinguished. Likewise, the embedding of the distance matrix in three-dimensions has two possible solutions: a structure and its mirror image. Since mirror images cannot be distinguished by the lattice-based residue contact potential or by the distance geometry procedure, for this particular work we chose the conformation that has lower cRMSD compared with native structure (note that this procedure cannot be used for "blind" prediction). However, further analysis showed that in almost all cases, this structure was readily discernible by the handedness of the α-helices or by the all-atom scoring function (RAPDF), since the local environment is different between right and left handed helices. This supports the view that all-atom models of mirror image structures will have different local environments that can be distinguished by all-atom potentials, due to handedness of amino acids and secondary structures.

2.8 Selection of a test set of proteins

A set of twelve small proteins (<= 110 residues) representing different fold classes was chosen as a test set. Half these proteins were targets for the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2), but the model building described in this work is not blind prediction. The reason we used CASP2 proteins is because they are more realistic test cases (for example, the secondary structure prediction accuracy for this set of six proteins is generally lower compared to the other six and the sizes of the proteins are generally larger). Table 1 lists the proteins that were used to generate test sets, along with the results. All proteins involved in the test sets were not used in compilation of the scoring functions, i.e., the procedure was properly jack-knifed.

3 Results and discussion

3.1 Accuracy of model construction for twelve proteins

Table 1 gives the cRMSDs for the structure with the lowest score after passing it through all the filters. For five out of twelve proteins, we are able to identify the correct topology of the protein and produce conformations that are ≈ 6.0 Å to the experimental structure (see example in Figure 2). For nine out of twelve proteins, we sample the conformational space adequately to ensure that

Table 1: Results of application of the combined approach for *ab initio* structure prediction to a set of twelve proteins. For each protein, the Protein Data Bank (PDB) [16] identifier, the length, the approximate class, and the three-state (helix, sheet, other) secondary structure prediction accuracy (Q3) of the PHD prediction, relative to the DSSP [17] assignments is given. Also shown are the range of cRMSDs for the 10,000 conformations after secondary structure fitting, and the cRMSD for the final selection. Proteins that were targets for the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2) are indicated, but we emphasise that the model building described in this work is not blind prediction. In general, the method fails on large mostly β proteins and works best on small α-helical proteins.

Protein (PDB code)	Size	Class	CASP2	Secondary structure prediction accuracy (Q3/%)[a]	cRMSD range (Å)	**cRMSD (Å)**
1fca	55	β		78.1	5.09 - 12.06	**5.90**
1pgb	56	$\alpha+\beta$		57.1	5.60 - 13.30	**8.41**
1trl-A	62	α		96.8	5.30 - 13.16	**6.35**
1fgp	67	β	Y	65.7	7.80 - 14.40	**10.93**
1ctf	68	$\alpha+\beta$		72.0	5.45 - 13.54	**5.75**
1dkt-A	72	β		72.2	6.68 - 14.79	**7.80**
1sro	76	β	Y	64.5	7.30 - 15.42	**9.68**
4icb	76	α		85.5	4.74 - 13.28	**4.95**
1nkl	78	α	Y	78.2	5.26 - 14.23	**5.70**
1beo	98	α	Y	54.0	6.96 - 15.94	**11.13**
1aa2	108	α	Y	75.9	6.18 - 15.28	**11.08**
1jer	110	β	Y	69.0	9.55 - 17.53	**13.60**
average	77	-	-	72.4	6.32 - 14.41	**8.44**

a conformation representing the correct topology is available in the sample space. The correct topologies are sampled and identified even in cases where the secondary structure assignments were not necessarily very accurate (Figure 3). There is no clear dependence of success on protein size, but it is notable that the three failres (PDB codes 1fgp, 1sro, and 1jer) are all β class proteins.

3.2 Computation times

For small proteins (less than 80 residues), the computation time for each protein is approximately three CPU days on a 533 MHz alpha processor for the entire process of building a model from sequence. The method is highly parallelisable (via a pipeline) and a large number of proteins (for example, complete small genomes) can be modelled using a farm of independent processors.

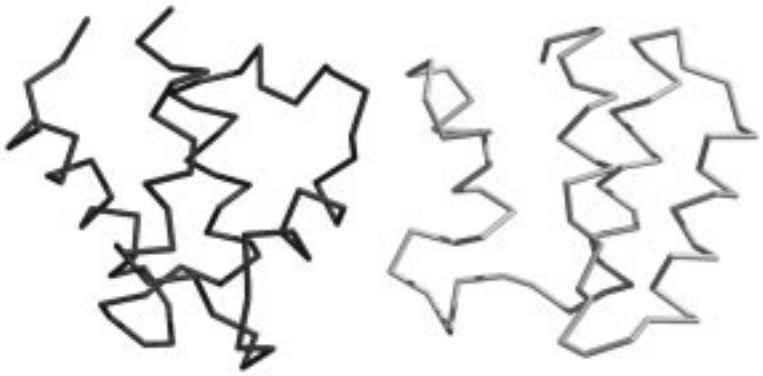

Figure 2: Comparison of the experimental conformation (left) and the final model (right) for 4icb/Calbindin. The model on the right has a cRMSD of 4.95 Å relative to the experimental structure.

```
1nkl ---HHHHHHHHHHHHHHH-----HHHHHHHHHHHHHHH----HHHHHHHHHH-----HHHHH----HHHHHHH------
PHD  -----HHHHHHHHHHH----------HHHHHHH------------HHHHHHHHHHHHHHH--------EEE-------
```

Figure 3: Comparison of the DSSP assignment for NK-Lysin (1nkl) and the high confidence PHD secondary structure prediction (H - α-helix; E - β-strand). The PHD assignments above were used as is for the fitting of the 10,000 structures, resulting in sampled conformations as low as 5.26 Å cRMSD, with a final selection of 5.70 Å. Conformations with these relatively low cRMSD values are observed even though the residues in final helix in NK-Lysin are fixed in β-stand conformation in the generated models.

3.3 Comparison to other methods

The results presented here are fairly promising for *ab initio* structure prediction, given that low resolution models can provide important biological insights into function and/or suggest further means of probing function. Recently, Ortiz *et al* [2] have reported interesting results for a set of twenty small proteins. Based on a qualitative comparison, it would appear that the results presented here are not as impressive as the results by Ortiz *et al* [2]. However, the method applied here is completely automated, is fairly tolerant of secondary structure prediction accuracy (see Figure 3), does not require multiple sequence alignment information, and can be applied to be proteins as large as a hundred residues or more. Further, it is difficult to make a direct comparison as the proteins for which final models were constructed by the two methods are different. In cases where both methods build a model for the same protein (1nkl/NK-Lysin and 4icb/Calbindin), the C_α cRMSDs are similar: 5.6 Å vs.

5.7 Å and 4.5 Å vs. 4.9 Å respectively for the approach of Ortiz *et al* [2] and the one described here.

Baker and colleagues [1] have reported on the *ab initio* generation of low RMSD conformations for a set of six small proteins, but their scoring function was not able to distinguish the conformation with the correct topology. For the two proteins common to the studies (4icb/Calbindin and 1pgb/Protein G), the best C_α cRMSD values in the sample space are similar: 4.7 Å vs. 4.9 Å and 6.3 Å vs. 5.6 Å respectively for the approach of Baker and colleagues [1] and the one described here.

It must be stressed that all the models constructed here are not "blind" prediction. We are testing this method at the third meeting on the Critical Assessment of Protein Structure Prediction (CASP3) which will enable us to make a definitive statement on the utility of this approach, particularly in comparison to other *ab initio* structure prediction methods.

3.4 Predictive power of this approach

The prediction quality of our method appears dependent on the secondary structure content of the protein to be modeled. This method does worse on β proteins, particularly if they are relatively large (over 100 residues). One reason for this results from simplicity of the lattice representation used in this work. Table 1 shows that the sampling range for the 10,000 structures for mostly β proteins is not adequate for the scoring functions used to be sufficiently discriminative. $\alpha+\beta$ proteins appear to have have mixed performance based on the limited data in Table 1. On a positive note, the approach works fairly well for helical proteins, both in terms of sampling and in terms of final selection.

With the current lattice scheme, we are limited by the degree of exhaustive enumeration that is done. Further, it is not possible to justify modification or tuning of secondary structure predictions when the correct answer is known. For the predictions at CASP3, we are exploring the conformational space to the extent that computational limits will permit (using longer lattice walk lengths and larger boundaries). We also use a consensus-based secondary structure prediction approach, which should lead to improved accuracy. The flip side is that most of the CASP3 targets are generally larger than 100 residues.

Even though the average model building accuracy is 8.44 Å cRMSD, the average cRMSD for the best conformation in the 10,000 structures is 6.32 Å for the twelve proteins. Thus better discrimination on the part of the scoring functions could potentially lead to more folds being correctly identified *ab initio*.

3.5 Availability of test sets and software

The best scoring sets of structures for the twelve proteins and the software for fitting and scoring of these conformations is available via the Decoys 'R Us database at <http://dd.stanford.edu/dd/>. The TINKER suite of programs is available at <http://dasher.wustl.edu/tinker/>.

Acknowledgments

We are extremely grateful to Patrice Koehl for providing us with efficient FORTRAN source code to construct protein models given a set of $\phi/\psi/\chi$ angles and to calculate the best-fit RMSD between conformations, and to Jay Ponder for TINKER and helpful advice on its application. This work was supported in part by a Burroughs Wellcome Fund Postdoctoral Fellowship awarded by the NSF Program in Mathematics and Molecular Biology to Ram Samudrala, a Howard Hughes Medical Institute Predoctoral Fellowship to Yu Xia, a Jane Coffin Childs Memorial Fund Fellowship to Enoch Huang, and NIH Grant GM 41455 to Michael Levitt.

References

1. K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268:209–225, 1997.
2. A. Ortiz, A. Kolinski, and J. Skolnick. Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.*, 277:419–448, 1998.
3. D.A. Hinds and M. Levitt. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA*, 89:2536–2540, 1992.
4. D.A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668–682, 1994.
5. B. Rost, C. Sander, and R. Scheider. Phd - an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.*, 10:53–60, 1993.
6. B. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, 249:493–507, 1995.
7. R. Samudrala, E.S. Huang, and M. Levitt. Side chain construction on non-native main chains using an all-atom discrimatory function. *In*

preparation, 1998.

8. R. Samudrala and J. Moult. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275:895–916, 1997.
9. T.J.P. Hubbard, A.G. Murzin, S.E. Brenner, and C. Chothia. Scop: a structural classification of proteins database. *Nucleic Acids Res.*, 25:236–239, 1997.
10. B. Park, E.S. Huang, and M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, 266:831–846, 1997.
11. E.S. Huang, R. Samudrala, and J. Ponder. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. *Protein Sci. (in press)*, 1998.
12. M. Levitt and S. Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, 46:269–279, 1969.
13. M. Levitt. Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.*, 82:393–420, 1974.
14. M. Levitt. Molecular dynamics of native protein. *J. Mol. Biol.*, 168:595–620, 1983.
15. M. Levitt, M. Hirshberg, R. Sharon, and V. Daggett. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Comm.*, 91:215–231, 1995.
16. F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.E.J. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tsumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.
17. W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.