# Ab initio protein structure prediction results using a simple distance geometry-based method

Enoch Huang[1] Ram Samudrala[2] and Jay Ponder[1]

[1]Department of Biochemistry and Molecular Biophysics
Washington University School of Medicine
Saint Louis, MO 63110

[2]Department of Structural Biology
Stanford University School of Medicine
Stanford, CA 94305

## Abstract

An approach to *ab initio* protein tertiary structure prediction based on building and scoring libraries of folds was tested at the CASP3 experiment. We present blind predictions for five helical targets varying in length from 31 to 114 residues. Our method was able to predict most of each structure correctly to about 6 Å Cα root-mean-square deviation (RMSD), with the exception of the largest target. Because distance geometry is used to constrain every pair of predicted helices to a generic distance range, inadequate sampling of tertiary folds occurs for larger targets or targets with poorly assigned helical residues. However, for targets with 5 helices or fewer, the native inter-helical distances are more compatible with the generic bounds, and sampling improves to the extent that near-native fold selection is possible.

## Introduction

Methods for *ab initio* protein structure prediction are needed to model sequences for which there is no similar structure or sub-structure available in the database. One type of *ab initio* method calls for the construction of a library of candidate folds, from which a single structure is specified as most likely to be native-like (Cohen *et al.*, 1979; Covell & Jernigan, 1990; Hinds & Levitt, 1992; Hinds & Levitt, 1994; Huang *et al.*, 1998; Huang *et al.*, 1999; Park & Levitt, 1996; Park *et al.*, 1997; Shortle *et al.*, 1998). By

decoupling the search from the scoring, this approach shifts the burden towards reasonable heuristics for fold construction and away from the reliance on a particular scoring function to guide a folding simulation.

Our method, the details of which are published elsewhere (Huang et al., 1999), can be applied to small (< 100 residues) targets that are known or predicted to be helical. Encouraged by our results from controlled testing, we tested the method at the third Critical Assessment of techniques for protein Structure Prediction (CASP3: `http://predictioncenter.llnl.gov/`). For this international community-wide experiment, participants submitted models of proteins for which experimentally-determined three-dimensional structures were imminent. Here we present our results for five small helical targets. Three of the targets were assessed in the "*ab initio*" category, and two in the "fold recognition" category, as structurally-similar folds were found in the Brookhaven protein databank (PDB) (Bernstein *et al.*, 1977). For all but the largest target (114 residues) the method was able to predict most of the structure correctly to about 6 Å Cα root-mean-square deviation (RMSD).

**Methods**

A library of structures for each target is generated using metric matrix distance geometry and a list of predicted and generic restraints. First, we used the PHD server (`http://www.embl-heidelberg.de/predictprotein/predictprotein.html`) to guide our prediction of helical boundaries, which we kept as fixed units throughout the procedure (Rost & Sander, 1993; Rost *et al.*, 1994). Idealized intra-helical distances were derived from these predicted helical segments. Next, we used the predicted solvent accessibility profiles, also taken from PHD, to choose a contact residue near the center of each predicted helix. We assigned a generic inter-Cα distance range between every pair of these designated residues. A range of 5 – 11 Å was chosen for smaller targets; for larger targets the upper bound was raised to 15 Å. After metrization and embedding, each

structure was subjected to refinement against the input distances and torsion restraints that enforce the correct handedness of the helices.

After the addition of side-chains (Levitt, 1992) and 200 steps of steepest-descent energy minimization against the ENCAD force field (Levitt *et al*., 1995), each structure was then scored with a normalized linear combination of three scoring functions: an all-atom distance-dependent probability function (Samudrala & Moult, 1998), a simple inter-residue contact function (Park et al., 1997), and a hydrophobic compactness function (Samudrala *et al*., 1999). The top-scoring folds in each library were then collected for a consensus distance geometry procedure, which yielded up to 18 different structures (Huang et al., 1998). These various consensus models were a product of two parameters: the number of top-scoring structures in the input set {50, 100, 500} and the weighting scheme {unweighted, Boltzmann-weighted, linearly weighted}. Taking into account both mirror images resulted in total of 3 x 3 x 2 = 18 structures. Consensus structures tended to be characterized by malformed helices caused by mutually incompatible restraints. Using the Cα trace of the consensus structure as a template, we fit the predicted secondary structure using a off-lattice build-up algorithm with a 4-state Ramachandran representation (Park & Levitt, 1995).

For each target, 5 models were submitted, though for some targets we used a threading-style approach in addition to this distance geometry method. We discuss only the *ab initio* distance geometry results here. Human intervention played a role in the selection of the final models. For instance, consistency with putative biochemical function and structural motifs (e.g. helix-turn-helix), was sought whenever applicable. Often, high-scoring models that were very similar to other high-scoring models were disregarded. Typically the submitted set comprised a diverse mix of top-scoring models in the library and consensus structures. The targets selected were T0065 (31 residues, SinI protein), T0056 (114 residues, DnaB helicase), T0061 (76 residues, protein HDEA), T0079 (96 residues, MarA protein), and the first 75 residues of T0083 (cyanase).

These sequences did not have any detectable sequence similarity with proteins of known structure. The number of residues predicted for each target was based strictly on the sequence provided by the CASP3 organizers (with the exception of T0083). For T0065, T0061, and T0079, the sequences of the respective experimental structures were shorter in length than the sequence given. These RMSD values thus reflect truncation of the models for consistency with the structures determined by experiment. Computation of side-chain relative solvent accessibility was performed with the NACCESS software (Hubbard & Thornton, 1993). Software for distance geometry calculations is available at `http://dasher.wustl.edu/tinker/`.

## Results and Discussion

*Quality of fold libraries*

In Table I, we list for each target the identifier, length of sequence, length of experimental structure in common with the model, the library size, the RMSD range in the library, and the secondary structure prediction accuracy. With the exception of the small helical target T0065, near-native folds were only sparsely generated. The RMSD ranges reflect the raw distance geometry structures, i.e. before side-chain construction and energy minimization.

T0056 (DNAB): Despite knowing the secondary structure assignment (provided by the CASP3 organizers), sampling of near-native folds proved difficult. Good sampling was hampered by the generic specification of restraints between all pairs of helices. Although the side-chains of all the predicted contact residues are buried in the NMR structure (PDB: 1jwe), they were not all in the generic 5 to 15 Å range. For a protein with six helices, there are 15 unique inter-helical distances, 5 of which were greater than 15 Å in DNAB (the greatest of which was 25 Å). One structure was only 6.7 Å RMSD

from the native; however, this was the only structure within 8 Å out of 2616 folds. The structure of DNAB is depicted in Figure 1.

T0061 (HDEA): We collected two different libraries for HDEA, totaling 2428 folds between them (Table I). Both assumed that the correct structure had 5 helices, but the placement of helices 4 and 5 was slightly different between the two libraries. The reported three-state accuracy (Q3) in Table I reflects the average of the two values (66% and 71%). The library that used the better secondary structure prediction exhibited lower average RMSD and minimum RMSD (data not shown). The quality of the sampling was compromised by the presence of the first predicted helix (residues 3 to 8 in the entire 89 residue sequence). The first 9 residues are missing in the crystal structure (PDB: 1bg8; Figure 2). If these missing residues do not actually form a helix that abuts the rest of the native structure, it is likely that considering the N-terminal helix would disrupt the building of a native-like model. To test this hypothesis, we removed the first 9 residues from our model and regenerated the library, keeping the positions of the helices and contact residues as before. After 1200 trials, the average RMSD of the library with better secondary structure prediction dropped from 11.3 Å to 10.7Å, and the best fold in the library improved from 6.2 Å to 5.3 Å.

T0065 (SINI): Only 31 out of the original 57 residues were visible in the crystal structure of SINI (PDB: 1b0n:B). This is a helix-hairpin that docks with its binding partner (SINR) to form a repressor / anti-repressor complex (Lewis *et al*., 1998). The mode of oligomerization is very distinctive, as two helices from each molecule interdigitate to bury an unusually extensive core (Lewis et al., 1998). The effect of this interaction is to splay the two helices of SINI apart (Figure 3), resulting in a distance of 14.8 Å between predicted contact residues. Nevertheless, since the location of the helical boundaries

was reasonably predicted by PHD, many structures in the library resembled the SINI molecule in isolation; roughly 10% (183 / 1880) of the folds were within 4 Å RMSD.

T0079 (MARA): As in the case of DNAB, above, model construction of MARA was frustrated by the designation of more inter-helical contacts that might be reasonably expected. The crystal structure of MARA (PDB: 1bl0; Figure 4) is composed of two subdomains, each of which contains a DNA-binding helix-turn-helix motif (Rhee *et al.*, 1998). Thus, 10 out of the 15 predicted inter-helical distances were greater than the upper-bound of 15 Å. The increased distance error is a result of the nature of the bipartite fold and the misprediction of Val33 as a buried residue (side-chain relative solvent accessibility: 39%). Four of the five inter-helical distances that involve this residue are greater than 15 Å. No fold in the library was within 8 Å RMSD of the native structure. This target was assessed in the fold recognition category.

T0083 (CYNS): The structure of CYNS is divided into two domains. We submitted models for the 5-helical N-terminal domain (Figure 5), which was evaluated as a fold recognition target. The domain prediction was based on visual inspection of the secondary structure prediction, which was quite accurate (Q3 = 83%). Even though 2 out of the 5 designated contact residues were partially exposed to solvent, only 2 out of the 10 inter-helical distances exceeded the upper bound set at 15 Å. Distance geometry generated 48 folds out of 1472 within 8 Å RMSD of the native, the best of which was 6.2 Å RMSD.

*Quality of predicted models*

For each target, up to 5 models were selected for submission. Table I lists the accuracy of the best submitted model (in its entirety). Also shown is the contiguous fragment with lowest RMSD and its corresponding length in residues. In a recent study

(Huang et al., 1999), we found that modeling a structure with the consensus inter-Cα distances from a Boltzmann-weighted subset of 50 folds (the "bw50" model) was more effective overall than simply saving the best-scoring fold, though there were exceptions. For purposes of comparison, we also list the best-scoring fold in the library and the consensus model built from the aforementioned parameters (Table II). In some cases, these models may have actually been members of the submitted set.

T0056 (DNAB): None of the submitted models were native-like, as the best complete model was over 12 Å RMSD. Neither the best-scoring nor the bw50 model could have improved upon this model. The best possible consensus model, which coincidentally is the best-scoring of the 18 consensus models, was 9.05 Å RMSD from the native. One of the 3 submitted models had a 70 residue fragment, spanning the last four helices, correctly predicted within 7.65 Å RMSD. Despite knowing the exact placement of the six helices, the poor sampling of tertiary arrangements precluded the possibility of a good prediction.

T0061 (HDEA): We consider this target to be a partial success. The best submitted model, which was 89 residues in length, matched the corresponding 76 residues of the four-helical native structure to 9.8 Å RMSD. A contiguous 60 residue segment matched the native to an accuracy of 6.7 Å (PDB residues 16 to 75); this corresponded to the first three helices of 1bg8 (Figure 2). Given the difficulty posed by marginal secondary structure prediction and the problematic effect of the non-existent N-terminal helix (see above), we were satisfied with how the method fared with this target. The CASP3 assessors mentioned our model of HDEA in their analysis of the targets of "medium" difficulty (Orengo et al., in preparation).

T0065 (SINI): This small motif was considered by the CASP3 assessor to be an "easy" target (Orengo et al., in preparation). Assuming one used a reasonable secondary structure prediction, most models for this 31-residue inhibitor would not stray too far from the correct fold. The difficulty of this target lay in predicting the rather unusual mode of binding with its molecular partner (Lewis et al., 1998) which in turn reveals the relative disposition of the two helices. Our best submitted model was 3.8 Å RMSD, and using the best-scoring fold in the library would have been a slight improvement at 3.4 Å (Figure 3).

T0079 (MARA): Although the final models were rather unimpressive for this target (best submitted model: 11.4 Å RMSD), we note that the central core of four helices (70 residues) is correctly predicted to 5.7 Å in one of our 4 models. The bipartite shape of 1bl0 separates the first and sixth helices by 24 Å at our designed residues; clearly this is a source of error. Indeed, Figure 4 shows that these terminal helices have been tethered closely together in our best model, which otherwise matches the native structure quite well.

T0083 (CYNS): One of our 5 models for the first domain of CYNS was correct to 8.7 Å. Most of the coordinate error can be attributed to the unstructured tail at the N-terminus (Figure 5) and the incorrect placement of the fifth predicted helix. Neglecting these segments, our best model (4 helices, 60 residues) was accurate to 5.4 Å RMSD. Had we submitted the bw50 model for this target, our entire 75-residue prediction would have been within 7 Å RMSD from the native (Table II).

**Conclusions**

Based on our experiences at CASP3 and with our test cases (Huang et al., 1999), it appears that our method is most effective with targets of 5 helices and fewer. The

recurring theme throughout this study is that successful prediction begins with a good library, and a good library depends critically on the choice of input restraints. For larger structures, our scheme of enforcing generic inter-helical distances is inadequate. We see this shortcoming in the cases of T0056 and T0079, wherein low near-native concentration in the libraries doomed the prediction of the entire structure to failure. For smaller targets like T0083, if the number and location of the helices are accurately predicted, the fraction of near-native structures present in the library rises to more acceptable levels. When secondary structure prediction is less accurate, or when a predicted helix is absent in the native structure, the quality of the library will also suffer, leading to partially correct structures (e.g. the case of T0061).

The scoring function has performed up to expectation in blind prediction testing. After disregarding T0065 as a stringent test case, the function was unable to select the best structures in the respective libraries for three of the four remaining targets (T0056, T0061, and T0079). However, as discussed earlier, this failure is primarily a consequence of the poor quality and scarcity of the best folds. In the single case (T0083) for which a substantial population of native-like folds was present, the scoring function performed quite well, selecting a native-like fold as the best scoring in the library. The bw50 fold (not submitted) would have been even a better choice, indicating that there were other native-like folds present in the low-energy subset.

The method as it currently stands is easily improved due to its extreme simplicity. One way to circumvent the problems of over-specifying distance restraints between helices is to be more selective in predicting distances. In this regard, one promising technique is contact prediction by correlated mutation analysis (Gobel *et al.*, 1994; Ortiz *et al.*, 1998). We have begun to assess the suitability of this approach for our prediction protocol.

**Acknowledgements**

This work was supported by Department of Energy grant DE-FG07-96ER14693 to J.W.P. and a grant from The Jane Coffin Childs Memorial Fund for Medical Research. E.S.H. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research. The authors wish to thank Dr. Patrice Koehl for supplying code used for constructing protein structures.

**References**

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol 112*: 535-42.

Cohen, F. E., Richmond, T. J. & Richards, F. M. 1979. Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J Mol Biol 132*: 275-88.

Covell, D. G. & Jernigan, R. L. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry 29*: 3287-94.

Gobel, U., Sander, C., Schneider, R. & Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins 18*: 309-17.

Hinds, D. A. & Levitt, M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci U S A 89*: 2536-40.

Hinds, D. A. & Levitt, M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol 243*: 668-82.

Huang, E. S., Samudrala, R. & Ponder, J. W. 1998. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. *Protein Sci 7*: 1998-2003.

Huang, E. S., Samudrala, R. & Ponder, J. W. 1999. Folding proteins with distance geometry and predicted inter-residue distances. *J Mol Biol in press.*

Hubbard, S. J. & Thornton, J. M. (1993). 'NACCESS' Computer Program. Department of Biochemistry and Molecular Biology, University College London.

Levitt, M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol 226*: 507-33.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. 1995. Potential-Energy Function and Parameters For Simulations of the Molecular-Dynamics of Proteins and Nucleic-Acids in Solution. *Computer Physics Communications 91*: 215-231.

Lewis, R. J., Brannigan, J. A., Offen, W. A., Smith, I. & Wilkinson, A. J. 1998. An evolutionary link between sporulation and prophage induction in the structure of a repressor:anti-repressor complex. *J Mol Biol 283*: 907-12.

Ortiz, A. R., Kolinski, A. & Skolnick, J. 1998. Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J Mol Biol 277*: 419-48.

Park, B. & Levitt, M. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol 258*: 367-92.

Park, B. H., Huang, E. S. & Levitt, M. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol 266*: 831-46.

Park, B. H. & Levitt, M. 1995. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol 249*: 493-507.

Rhee, S., Martin, R. G., Rosner, J. L. & Davies, D. R. 1998. A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc Natl Acad Sci U S A 95*: 10413-8.

Rost, B. & Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*: 584-99.

Rost, B., Sander, C. & Schneider, R. 1994. PHD--an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci 10*: 53-60.

Samudrala, R. & Moult, J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol 275*: 895-916.

Samudrala, R., Xia, Y., Levitt, M. & Huang, E. S. 1999. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac Symp Biocomput*: 505-516.

Shortle, D., Simons, K. T. & Baker, D. 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci U S A 95*: 11158-62.

**Table I**

| Target | len (seq) | len (str) | nFolds | RMSD Range (Å) | <RMS> (Å) | nMod | Best Model RMSD (Å) | Frag RMSD (len) |
|---|---|---|---|---|---|---|---|---|
| T0056/dnab | 114 | 114 | 2616 | 6.73 - 18.04 | 13.99 | 3 | 12.07 | 7.65 (70) |
| T0061/hdea | 89 | 76 | 2428 | 6.18 - 14.83 | 11.54 | 5 | 9.83 | 6.67 (60) |
| T0065/sini | 57 | 31 | 1180 | 3.01 - 7.42 | 4.93 | 4 | 3.82 | n/a |
| T0079/mara | 104 | 96 | 3018 | 8.01 – 17.29 | 13.20 | 4 | 11.39 | 5.69 (70) |
| T0083/cyns | 75 | 75 | 1472 | 6.15 – 14.22 | 12.47 | 5 | 8.70 | 5.43 (60) |

*Table II*

| Target | len | Consen | Select |
|---|---|---|---|
| T0056/dnab | 114 | 12.11 | 13.12 |
| T0061/hdea | 76 | 10.25 | 11.78 |
| T0065/sini | 31 | 4.61 | 3.39 |
| T0079/mara | 96 | 12.00 | 15.38 |
| T0083/cyns | 75 | 6.97 | 8.70 |