# Ab Initio Protein Structure Prediction Using a Combined Hierarchical Approach

**Ram Samudrala,[1]* Yu Xia,[1] Enoch Huang,[2] and Michael Levitt[1]**
[1]*Department of Structural Biology, Stanford University School of Medicine, Stanford, California*
[2]*Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri*

*ABSTRACT* **As part of the third Critical Assessment of Structure Prediction meeting (CASP3), we predict the three-dimensional structures for 13 proteins using a hierarchical approach. First, all possible compact conformations of a protein sequence are enumerated using a highly simplified tetrahedral lattice model. We select a large subset of these conformations using a lattice-based scoring function and build detailed all-atom models incorporating predicted secondary structure. A combined all-atom knowledge-based scoring function is then used to select three smaller subsets from these all-atom models. Finally, a consensus-based distance geometry procedure is used to generate the best conformations from each of the all-atom subsets. With this method, we are able to predict the global topology/shape for all or a large part of the sequence for six out of the thirteen proteins. For two other proteins, the topology/shape for shorter fragments are predicted. This represents a marked improvement in ab initio prediction since CASP was first instigated in 1994. Proteins Suppl 1999;3:194–198.**

**Key words: lattice models; knowledge-based scoring functions; discrete-state models; distance geometry**

## INTRODUCTION

Ab initio prediction of protein structure from sequence can be divided into two major subproblems: (1) sampling the conformational space of the protein well so that a significant number of native-like conformations are generated, and (2) designing a discriminatory/scoring/energy function that will distinguish between native and non-native conformations in this sample. Exhaustive enumeration is a powerful means of sampling the global topology,[1–3] and all-atom scoring functions have been shown to be useful in identifying native-like folds from a range of conformations.[4,5] The approach we present, which uses a combination of hierarchical methods for generating and selecting structures, is partially successful in solving both of these problems.

## METHODS

### Combined Approach for Prediction

Table I gives a list of proteins predicted. For each protein, all possible self-avoiding compact conformations were exhaustively enumerated using a tetrahedral lattice model.[1,2] The computation is made tractable by reducing the chain length to no more than 50 lattice vertices (with two to three residues per vertex, depending on the size of the protein). This procedure yielded 10 million to 10 billion lattice conformations. Of these, up to 40,000 best-scoring conformations were selected using a simple lattice-based pairwise scoring function.[2]

All-atom models were constructed by “fitting” the predicted secondary structure to the best-scoring lattice models. The secondary structure prediction was accomplished by generating 20 multiple sequence alignments of a homologous set of sequences to the target protein (using a bootstrapping procedure) and using them as input for three previously published secondary structure prediction methods: PHD,[5] DSC,[6] and Predator.[7] The consensus of the 20 predictions for each method was used to assign helical and sheet residues where all three methods agreed. A greedy off-lattice build-up procedure with a four-state ($\phi,\psi$) representation (one state helix, one sheet, two other)[8] was used to minimize the root mean square deviation (RMSD) between the lattice model and the all-atom model, taking into account predicted helix and sheet assignments. The most frequently observed rotamer values in protein structures were used for constructing side chains. The all-atom models were refined by applying 200 steps of steepest descent minimization using ENCAD.[9–12]

Three subsets consisting of the best 50, best 100, and best 500 conformations from the set of all-atom models were selected by a combined scoring function. The combined function consisted of an all-atom distance-dependent conditional probability discriminatory function (RAPDF),[4] a simple residue-level pairwise contact function (Shell),[13] and a hydrophobic compactness function.[3] The most frequently observed $C_\alpha$-$C_\alpha$ distances in each of the three subsets were used as constraints to a distance geometry procedure (by the TINKER software suite)[14] to generate up to 36 models. Predicted secondary structures were once again fitted to the consensus distance geometry models, the models refined, and four best scoring models, as evaluated by the all-atom function (RAPDF), were submitted. The fifth model submitted was the best scoring

Grant sponsor: National Institutes of Health; Grant number: GM 41455.

*Correspondence to: Dr. Ram Samudrala, Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305. E-mail: ram@zen.stanford.edu

Received 8 February 1999; Accepted 14 May 1999

**TABLE I. List of Targets[†] Predicted for CASP3**

| | | | | | Number | Best all | | Best fragment | | | | SS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Walk | selected[b] | $C_\alpha$ RMSD[c] | | $C_\alpha$ RMSD[d] | | | $C_\alpha$ RMSD | Q3[f] |
| | Target | Size | Class | length[a] | folds | model | (Å) | model | (Å) | (size) | range[e] (Å) | (%) |
| (−) | T43/hppk | 158 | $\alpha/\beta$ | 50 | 40,000 | 2 | 14.5 | 2 | 6.3 | (48) | 10.0–19.5 | 70 |
| (−) | T46/adg | 119 | $\beta$ | 50 | 23,120 | 1 | 13.9 | 4 | 6.6 | (39) | 10.1–19.2 | 67 |
| (−) | T52/cvn | 98 | $\beta$ | 50 | 278 | 2 | 13.6 | 4 | 6.6 | (33) | 10.6–16.3 | 50 |
| (−) | T54/vanx[g] | 202 | $\alpha + \beta$ | 51 | 19,816 | 3 | 15.5 | 3 | 15.5 | (202) | — | — |
| (*) | T56/dnab | 114 | $\alpha$ | 50 | 40,000 | 5 | 13.0 | 1 | 6.8 | (60) | 6.2–17.8 | 100 |
| (**) | T59/smd3 | 71 | $\beta$ | 38 | 20,000 | 2 | 11.6 | 2 | 6.7 | (46) | 7.4–15.7 | 80 |
| (**) | T61/hdea | 76 | $\alpha$ | 45 | 22,794 | 1 | 10.1 | 4 | 7.4 | (66) | 6.0–14.0 | 62 |
| (−) | T63/if5a | 135 | $\beta$ | 50 | 40,000 | 2 | 15.1 | 1 | 6.4 | (35) | 10.8–22.0 | 60 |
| (**) | T64/sinr | 103 | $\alpha$ | 50 | 40,000 | 5 | 11.2 | 5 | 4.8 | (68) | 8.0–18.8 | 90 |
| (*) | T65/sini | 31 | $\alpha$ | 29 | 20,422 | 2 | 4.1 | 2 | 4.1 | (31) | 2.4–7.6 | 90 |
| (*) | T74/eps15 | 98 | $\alpha$ | 49 | 37,296 | 3 | 11.3 | 1 | 7.0 | (60) | 6.3–16.5 | 88 |
| (**) | T75/ets1 | 88 | $\alpha$ | 50 | 40,000 | 1 | 9.8 | 1 | 7.7 | (77) | 6.0–17.0 | 78 |
| (*) | T84/rlz[h] | 30 | $\alpha$ | — | 5 | 5 | 1.0 | 5 | 1.0 | (30) | — | 82 |
| | Average | 102 | — | 47 | 26,440 | — | 11.1 | — | 6.7 | (61) | 7.6–16.8 | 77 |

[†] For 4 of 13 cases, marked with (**), we correctly predict the topology/shape for all or a large portion of the sequence; for another 4 of 13 cases, marked with (*), we correctly predict the topology/shape for small fragments of the sequence (or these were "easy" predictions), and there were five cases, marked with (−), where the method did not yield a good prediction.

[a] In some cases, sets of tetrahedral lattice conformations were generated with two different walk lengths. The table lists one value, and the other value, which is sometimes used, is 40. The number of residues per lattice point is simply the protein size divided by the walk length and varies from 1 (for T65/sini) to 4 (for T54/vanx).

[b] The number of conformations for which all-atom models were built and evaluated using the combined scoring function.

[c] The $C_\alpha$ root mean square deviation (RMSD) between the best model (of five) and the experimental structure for all the residues in the protein. The model number is indicated.

[d] The $C_\alpha$ RMSDs of the fragments predicted best among the five models (see Fig. 1). The model number is indicated.

[e] The range of RMSDs for all of the all-atom conformations sampled.

[f] The three-state (helix, sheet, loop) secondary structure accuracy for the prediction that was used to build the all-atom models (except for T56/dnab, where the assignments were made available to the predictors).

[g] Some data for T54/vanx are missing because the experimental coordinates were not made available to the predictors; the data shown were provided by the CASP organizers.

[h] This model was so simple to construct that no searching of the conformational space was required.

model (evaluated by RAPDF) from the set of all-atom models without the distance geometry step.

### Postprocessing of Models and Experimental Structures for RMSD Calculation

In some cases, the number of residues predicted was based on the sequence given and is larger than the number of residues in the corresponding experimental structures, which sometimes have missing residues. In these cases, the models, the conformations sampled, and the experimental structures were postprocessed for consistency in calculating the RMSDs, and the numbers given are those calculated after any postprocessing. The RMSDs also are generally for the best models among the five that were submitted, are sequence dependent, and are based on a global superposition of the coordinates. Residue numbering is based on the experimental structures provided to us by the prediction center.

## RESULTS

Table 1 gives the numerical results for the 13 predictions, and Figure 1 illustrates the performance of this method at CASP3 for the more successful predictions. We describe the results for all targets briefly and comment on what went right, what went wrong, and why.

### T43/hppk (−)

T43/hppk is a mixed $\alpha/\beta$ protein. There are a few fragments between 40 and 50 residues that are predicted to within 6.0 Å, but the poor RMSDs for the best conformations that were sampled (10.0 Å), the large size of the protein (158 residues), and the relatively low accuracy of the secondary structure prediction (70% Q3) all contribute to a mostly incorrect prediction.

### T46/adg (n-)

We could not adequately sample the conformational space for T46/adg (we generated 23,120 instead of the usual 40,000 conformations because of time limitations). The fact that our lattice representation has difficulty representing all-$\beta$ proteins and the weak secondary structure prediction (67% Q3) leads us to a situation where the RMSD of the best conformation sampled was 10.1 Å. Given the sampling, our prediction of 13.9 Å (model 1) is not too surprising, but in our fourth model, the topology for three of eight strands is approximately captured (Fig. 1).

### T52/cvn (−)

T52/cvn was our first prediction and our sampling was woefully inadequate given the time constraints for this prediction. We used the published disulfide information to

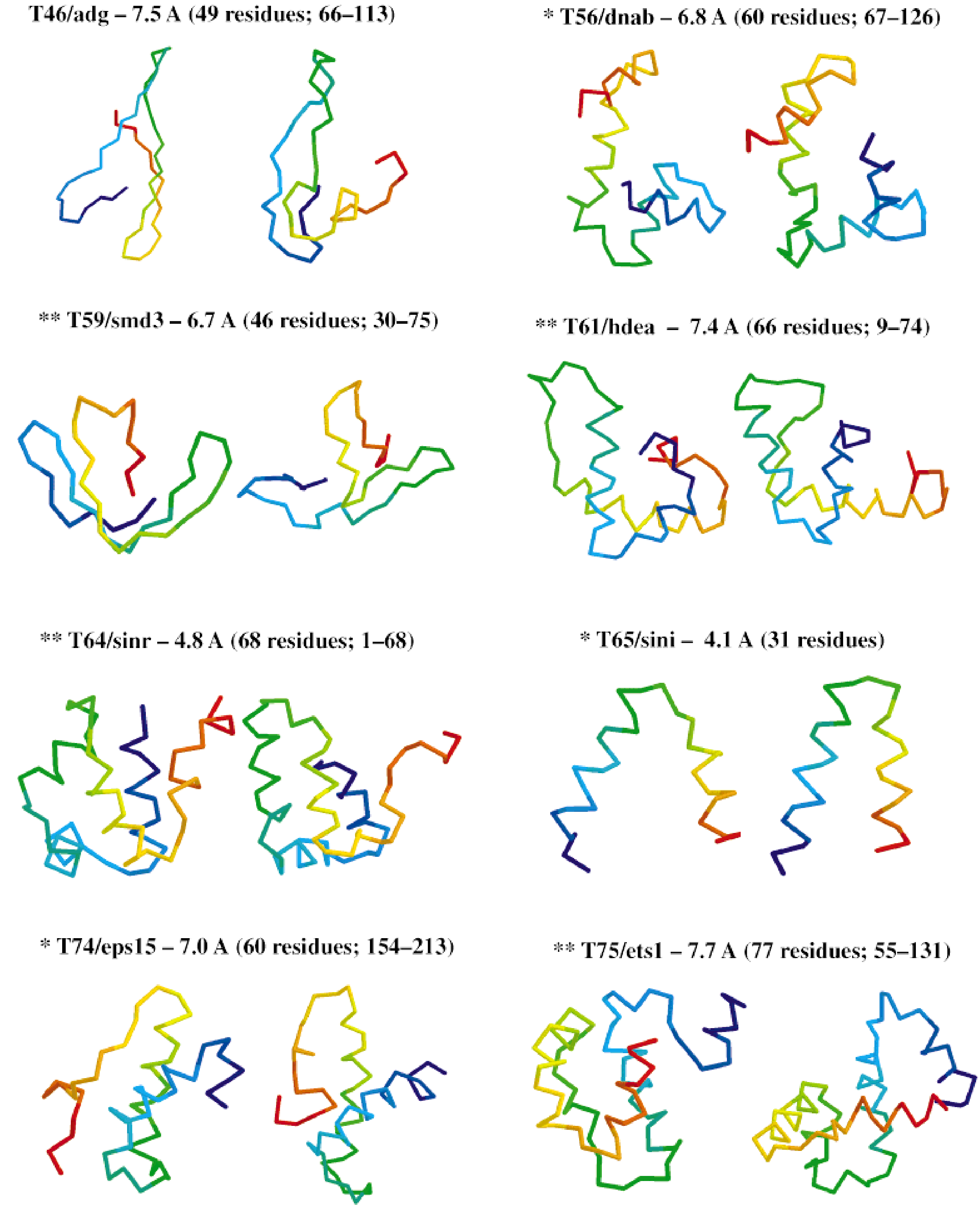
T46/adg – 7.5 A (49 residues; 66–113)
* T56/dnab – 6.8 A (60 residues; 67–126)
** T59/smd3 – 6.7 A (46 residues; 30–75)
** T61/hdea – 7.4 A (66 residues; 9–74)
** T64/sinr – 4.8 A (68 residues; 1–68)
* T65/sini – 4.1 A (31 residues)
* T74/eps15 – 7.0 A (60 residues; 154–213)
** T75/ets1 – 7.7 A (77 residues; 55–131)

Fig. 1. Illustrations of some of the more successful predictions obtained by our combined approach. In all cases, the experimental structure is on the left and the predicted model is on the right. The chains are colored according to sequence order (from the N-terminus in blue to the C-terminus in red). For six cases (four of these are shown in the figure and marked [**]), we were able to predict the topology/shape for all or a large portion of the sequence; for three other cases (two are shown and marked [*]), we were able to predict the topology/shape for small fragments of the sequence. The particular model shown is the one of the five generated that has the Best Fragment $C_\alpha$ RMSD; please refer to Table 1 for the model number in each case.

limit the number of conformations scored to only 278. The inadequate sampling (best RMSD was 10.6 Å) and the poor secondary structure prediction (50% Q3) contribute to a wrong prediction.

### T54/vanx (−)

The experimental coordinates for T54/vanx had not been released to the predictors at the time of writing, and therefore it is difficult to ascertain whether there are regions in the protein that were predicted well or whether native-like topologies were sampled. This is the largest protein predicted, and the RMSD (as provided to us by the CASP organizers) is a high 15.5 Å between our model (3) and the experimental structure for all the 202 residues.

### T56/dnab (*)

T56/dnab is a particularly disappointing misprediction for two reasons:

1. The secondary structure for this protein was completely specified.
2. The conformations sampled included four structures between 6.0 and 7.0 Å RMSD for the 114 residues.

Our scoring function was unable to select any of these native-like conformations, perhaps because of their scarcity (1 in 10,000). However, there is some partial success because fragments of size 60 are predicted to within 7.0 Å (Fig. 1).

### T59/smd3 (**)

T59/smd3 is a 71-residue, eight-stranded β-barrel. The topology of the last six strands is predicted to 6.7 Å (Fig. 1), making this one of our successful β-protein predictions. Part of this can be attributed to the reasonable sampling (RMSD range of 7.4–15.7 Å for 22,794 conformations) because of the simple topology and the relatively high secondary structure prediction accuracy (80% Q3) for a β-protein.

### T61/hdea (**)

T61/hdea is a 76-residue protein with a four-helix core and a long loop. The topology of the first three helices (66 residues) is predicted accurately, to an RMSD of 7.4 Å relative to the experimental structure (Fig. 1). However, the last helix is misplaced, leading to an overall large RMSD.

### T63/if5a (−)

T63/if5a is a complicated two-domain β-barrel. The lattice models are unable to capture the complicated topology, leading to inadequate sampling (10.8–22.0 Å RMSD). Combined with the poor secondary structure prediction accuracy (60% Q3), this results in an incorrect prediction.

### T64/sinr (**)

T64/sinr is a protein whose sequence is 30% identical to that of a protein with known structure, but this knowledge was not used in our ab initio prediction. For the first 68 residues, which form a four-helix bundle, we predict a conformation of 4.8 Å RMSD to the experimental structure (Fig. 1), and the remaining 35 residues, which form a two-helix bundle, we predict to 5.8 Å RMSD. The packing between the two domains is predicted incorrectly, leading to high RMSDs both in the sampling and in the final model.

### T65/sini (*)

T65/sini is a two-helix bundle. The helices in the native structure are noncompact compared to our model. For the 31 residues, the RMSD between the experimental and predicted structures is 4.1 Å (Fig. 1), which is probably not significant given the small size and simple packing observed in the protein.

### T74/eps15 (*)

T74/eps15 is sampled adequately (30 conformations within 7.0 Å for 98 residues), and the secondary structure prediction accuracy is high (88% Q3). Like dnab/t56, our scoring function is unable to select a conformation with native-like topology, and the best predictions are 60-residue fragments to 7.0 Å RMSD, which capture the topology for three helices (Fig. 1).

### T75/ets1 (**)

T75/ets1 is one of our most successful predictions. The RMSD between our model (1) and the experimental structure is 9.8 Å for the entire 88 residues and 77 of the residues (27–103) are predicted to 7.7 Å. Even though the RMSD is fairly large, the global topology is captured well (Fig. 1).

### T84/rlz (*)

T84/rlz (30 residues) was correctly predicted to be a single long helix. The model was constructed manually (by setting the torsion angles to idealized helix values), leading to a RMSD of 1.0 Å between our fifth model and the experimental structure. This is not considered to be very significant.

### Computational Issues

The time required for a single prediction is about 1 week on a single 533 MHz DEC alpha processor. However, the procedure can be run in a massively parallel manner, and 10 processors in parallel will predict a protein structure of less than 200 residues in 24 hr.

## DISCUSSION

Using the combined hierarchical approach, we see consistent fragmentary and/or topological prediction for most of the targets we predicted at CASP3. This represents significant progress in ab initio prediction relative to CASP1 and CASP2.

For six proteins (T59/smd3, T61/hdea, T64/sinr, T65/sini, T74/eps15, and T84/rlz), we predict models that capture the global topology for all or large portions of the sequence (Fig. 1). For two others (T56/dnab and T74/eps15), we predict the correct topology for relatively short fragments of the sequence. There are four failures (T43/

hppk, T46/adg, T52/cvn, and T63/if5a), three of which are all or mostly β proteins and one of which is a mixed α/β protein. For eight proteins (T59/smd3, T61/hdea, T56/dnab, T64/sinr, T65/sini, T74/eps15, T75/ets1, and T84/rlz), native-like topologies (RMSD below 7.6 Å for the whole sequence) are sampled, but not necessarily predicted, for all or large portions of the protein (Table 1).

### Problems With This Approach

Although the results presented here are encouraging, the conformations predicted are not really useful for functional studies.

In most cases where the method fails (T43/hppk, T46/adg, T52/cvn, and T63/if5a), inadequate sampling appears to be the major hurdle preventing selection of good ab initio models. In particular, sampling for all or mostly-β proteins is worse than that for mixed α-β proteins, which is worse than for all or mostly-α proteins. This trend is reflected in the quality of the final model.

Even when native-like topologies are sampled, the large number of incorrect conformations can overwhelm the scoring function, resulting in an incorrect final model with perhaps some correct fragments (T56/dnab and T74/eps15). In these cases, it is difficult to ascertain whether it is a failure of the selection procedure or whether the sample does not include sufficient native-like topologies. The low resolution of the models also makes it hard to assess how much of the failure can be attributed to limitations of the discriminatory function.

### Advantages of This Approach

The procedure is mostly automatic and very little human intervention was used in making the predictions, including the selection of the final five models submitted to CASP.

The method is robust with respect to secondary structure prediction. Even though some of the Q3 percentages are reasonable (60%–70%), secondary structure prediction is poor in many of these targets (T43/hppk, T46/adg, T61/hdea, and T63/if5a) where entire α-helices are predicted as β-strands and vice versa.

The method is also robust with respect to the details of the knowledge-based databases used. Database information from known protein structures is only used for the simple and all-atom scoring function and for the secondary structure prediction. For the rest of the method, no database information is used.

### The Road Ahead

In comparison to CASP1 and CASP2, there has been progress in terms of RMSDs for all the residues and across several targets. However, a general solution to the structure prediction problem has not been found. Even in “good” cases, i.e., where native-like topologies are observed in our sample space, the best conformations are between 6.0 and 8.0 Å RMSD relative to the experimental structure for 60-residue fragments. Approaches that sample closer to the native structure are necessary for detailed all-atom scoring functions to discriminate effectively. Combining exhaustive search methods with predicted constraints,[15] nonexhaustive knowledge-based sampling methods,[16] and better secondary structure prediction (e.g., results of the Jones group at CASP3) is a possible path for future work.

### Availability of Software and Decoys

The ensembles of structures that were generated and much of the software used to generate them are available at <http://dd.stanford.edu> and <http://www.ram.org/computing/ramp/ramp.html>, respectively. The TINKER suite of programs, used for the consensus distance geometry, is available at <http://dasher.wustl.edu/tinker/>.

## ACKNOWLEDGMENTS

The authors thank everyone in the Levitt laboratory for stimulating discussions and their flexibility in giving us priority for computing resources. We also thank Patrice Koehl for his efficient FORTRAN code used for constructing protein structures and calculating RMSDs, and Jay Ponder for the TINKER package. This work was supported in part by a Burroughs Wellcome Fund/Program in Mathematics and Molecular Biology fellowship to Ram Samudrala, a Howard Hughes predoctoral fellowship to Yu Xia, and a Jane-Coffin Childs Fellowship to Enoch Huang.

## REFERENCES

1. Hinds D, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
2. Hinds D, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
3. Samudrala R, Xia , Levitt M, Huang E. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. In: Altman R, Dunker A, Hunter L, Klein T, Lauderdale, K, editors. Proceedings of the Pacific Symposium on Biocomputing. World Scientific Publishing Co: Singapore; 1999. p 505–516.
4. Samudrala R, Moult J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1997;275:895–916.
5. Rost B, Sander C. Prediction of protein structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
6. Ross D, Sternberg M. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 1996;5:2298–2310.
7. Frishman D, Argos P. Knowledge-based secondary structure assignment. Proteins 1995;23:566–579.
8. Park B, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
9. Levitt M, Lifson S. Refinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol 1969; 46:269–279.
10. Levitt M. Energy refinement of hen egg-white lysozyme. J Mol Biol 1974;82:393–420.
11. Levitt M. Molecular dynamics of native protein. J Mol Biol 1983;168:595–620.
12. Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp Phys Comm 1995;91:215–231.
13. Park B, Huang E, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
14. Huang E, Samudrala R, Ponder J. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. Protein Sci 1998;7:1998–2003.
15. Ortiz A, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.
16. Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.