doi:10.1006/jmbi.2000.3835 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **300**, 171–185

JMB

Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach

Yu Xia[1], Enoch S. Huang[2], Michael Levitt[1]* and Ram Samudrala[1]*

[1]*Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305, USA*

[2]*Cereon Genomics, 45 Sidney Street, Cambridge MA 02139, USA*

**Corresponding authors*

We present a hierarchical method to predict protein tertiary structure models from sequence. We start with complete enumeration of conformations using a simple tetrahedral lattice model. We then build conformations with increasing detail, and at each step select a subset of conformations using empirical energy functions with increasing complexity. After enumeration on lattice, we select a subset of low energy conformations using a statistical residue-residue contact energy function, and generate all-atom models using predicted secondary structure. A combined knowledge-based atomic level energy function is then used to select subsets of the all-atom models. The final predictions are generated using a consensus distance geometry procedure. We test the feasibility of the procedure on a set of 12 small proteins covering a wide range of protein topologies. A rigorous double-blind test of our method was made under the auspices of the CASP3 experiment, where we did *ab initio* structure predictions for 12 proteins using this approach. The performance of our methodology at CASP3 is reasonably good and completely consistent with our initial tests.

Keywords: protein structure prediction; lattice model; knowledge based; discriminatory function; decoy approach

Introduction

Ab initio protein structure prediction remains one of the most important unsolved problems in molecular biophysics after 30 years of intensive research. This problem is in principle solvable: if we know the exact formulation of the physical micro-environment within a cell where proteins fold, we will be able to mimic the folding process in nature by computing the molecular dynamics based on our knowledge of the physical laws (McCarmmon & Harvey, 1987; van Gunsteren, 1998; Duan & Kollman, 1998). Complementarily, we can rely on the much-debated thermodynamic hypothesis, i.e. that the native protein structure is thermodynamically stable and is located at the global free energy minimum (Anfinsen, 1973). However, we do not yet have a complete understanding of the driving forces behind protein folding. Perturbations introduced by errors in the potential energy landscape may possibly result in a different folding pathway and a different folded structure. Even if we have an accurate enough potential energy function, we are still hampered by the huge search space (Levinthal, 1968).

Novel methods have recently been proposed for *ab initio* protein structure prediction with impressive results (Simons *et al.*, 1999; Ortiz *et al.*, 1999; Osguthorpe, 1999; Lomize *et al.*, 1999; Lee *et al.*, 1999; Huang *et al.*, 1999; Eyrich *et al.*, 1999). Current methods for structure prediction can be roughly grouped into two categories. The first set of methods include Monte Carlo and deterministic energy minimization (Hansmann & Okamoto, 1999; Scheraga, 1996; Levitt & Lifson, 1969) and genetic algorithms (Pedersen & Moult, 1996), which generally start from either one or a small set of random starting points and attempt to drive the conformation to a low energy in an iterative manner. The major advantage of these methods is that they more or less mimic the physical process of protein folding. Besides the folded structure, the pathway that leads to the folded structure may also be obtained. However, the success of these methods depends crucially on the very high

Abbreviations used: cRMS, C^{α} root mean square deviation; dRMS, distance root mean square deviation; CASP, critical assessment of protein structure prediction methods; RAPDF, residue-specific all-atom conditional probability disriminatory function; SCOP, structural classification of proteins.

E-mail addresses of the corresponding authors: levitt@zen.stanford.edu; ram@csb.stanford.edu

0022-2836/00/010171–15 $35.00/0

quality of energy function: it not only has to discriminate the native structure from all the other possible structures along any possible simulation run, it also has to lead any random starting configuration toward the native structure. It is not clear whether current energy functions can satisfy the two requirements simultaneously (Moult, 1997).

The second set of methods use a sampling procedure to produce trial structures, known as decoys (Chelvanayagam *et al*., 1998; Huang *et al*., 1999; Park & Levitt, 1996), that are subsequently evaluated by an energy function. The structure with the lowest energy is assumed to be the native structure.

We chose to work within the second paradigm for the following reasons. First, the prediction procedure is separated into sampling and selection processes. Each process is modular and can be developed and calibrated separately. Ineffectiveness of the whole procedure can be attributed to one or both parts that can be corrected as needed. Second, we ignore pathway prediction and focus our attention on structure prediction. Protein folding is a process that involves hundreds of degrees of freedom. Any single simulation can easily be trapped in one of many local minima along the folding pathway, and the chances of overcoming a local energy minimum decrease exponentially with the height of the free energy barrier. With the decoy approach, it is possible to explore millions of local energy minima of protein conformations in parallel, thereby sampling the protein conformational space effectively without the need to overcome high energy barriers. Third, the requirements demanded of the energy function have been significantly reduced: the only requirement is discrimination between near-native and non-native structures. This allows for the use of powerful statistical energy functions as discriminatory functions, which may not perform as well as folding potentials.

Structure sampling and evaluation have conflicting needs. We need simplified models to reduce the dimensionality of the sampling space to make the computations tractable. At the same time, to make the best selections, we need structures with full atomic detail to represent potential energy surface with enough accuracy to allow discrimination by energy functions. Unfortunately, generating and evaluating all-atom structures is a time-consuming process and cannot be done for huge numbers of conformations.

We tackle this problem by sampling low resolution structures exhaustively, and performing the final selection with a limited, yet promising, set of all-atom structures. Our approach starts with an exhaustive enumeration of all possible folds using a highly simplified tetrahedral lattice model. A set of filters are then applied to these folds, primarily in the form of discriminatory functions. As the filters are applied, we add more detail to the models, until one final all-atom model remains. Using this purification scheme, many non-native structures are pruned out due to high energy even before all-atom structures are built. Here, we describe our methodology in detail and provide a comprehensive analysis of its performance.

Results and Discussion

Simplified lattice representation is able to represent protein conformations well

Table 1 shows the parameters that we used in the lattice prediction procedure for the 12 test proteins. All the parameters are predetermined and are only dependent on protein size. Our simple lattice model will only be useful if it can represent native protein features to a good approximation. How well can the tetrahedral lattice model represent native protein structures? To answer this question, we compute for each test protein the distance root mean square deviation (dRMS) of the lattice structure with the highest number of native contacts (Table 2). This structure will be picked out if we have a perfect energy function, and is a measure of how well the lattice can represent protein structures.

In general, larger proteins are represented less accurately. However, there is a tremendous degree of variation: the best dRMS fit for 1aa2 with 108 residues is less than 3 Å, whereas the best dRMS fit for 1fgp with only 67 residues is slightly larger at 3.15 Å. In most cases, the best dRMS fit ranges from 2.4 Å to 3.3 Å. This shows that our simplified lattice model is able to represent the full variety of supersecondary structure topologies that occur in native proteins. The selection of these best structures depends entirely on the energy function.

Simplified energy function is able to select good subsets of lattice models

We use a simple statistical contact energy function for both threading optimization and selection of low energy structures. Performance of different energy functions is characterized by how far the dRMS distribution of the low energy population is pushed away from that of all lattice structures towards native structure. Using energy criteria for threading and selection pushes the structure population towards lower dRMS in all 12 test cases (Table 2). In the case of protein 1dkt-A, energy selection improves the mean of the dRMS distribution by 2 Å. On average, the dRMS distribution shifts 0.86 Å towards the lower end by applying energy criteria. A likely reason for the selection power of our energy function is that the function we use is complementary to the geometry of the lattice scheme. Even though our simple lattice models ignore local geometrical information such as side-chain orientation and secondary structure, they have well-formed interiors that can represent the hydrophobic core of native structures. Since our energy function captures the dominant hydro-

Table 1. Proteins and parameters used in lattice structure enumeration and selection

Protein[a]	Size[b]	Class	Walk length[c]	Edge size (Å)[d]	Bounding box vertex count[e]	R_g cutoff[f]
A. *Test set*						
1aa2	108	α	38	5.71	51	1.08
1beo	98	α	38	5.53	51	1.08
1ctf	68	$\alpha + \beta$	34	5.08	50	1.10
1dkt-A	72	β	36	5.08	51	1.10
1fca	55	β	28	5.08	50	1.12
1fgp	67	β	34	5.08	50	1.10
1jer	110	β	38	5.75	51	1.08
1nkl	78	α	38	5.12	51	1.08
1pgb	56	$\alpha + \beta$	28	5.08	50	1.12
1sro	76	β	38	5.08	51	1.08
1trl-A	62	α	31	5.08	50	1.10
4icb	76	α	38	5.08	51	1.08
B. *CASP3 predictions*[g]						
T0043	158	α/β	50/40	5.92/6.37	60/56	1.08/1.14
T0046	119	β	50/40	5.38/5.80	60/56	1.08/1.14
T0052	101	β	50	5.08	60	1.08
T0054	202	$\alpha + \beta$	51	6.38	60	1.08
T0056	114	α	50/40	5.31/5.72	60/56	1.08/1.14
T0059	75	β	38	5.08	56	1.14
T0061	89	α	45	5.09	60	1.08
T0063	138	β	50/40	5.66/6.09	60/56	1.08/1.14
T0064	111	α	50/40	5.26/5.67	60/56	1.08/1.14
T0065	57	α	29	5.08	60	1.80
T0074	98	α	49/40	5.08/5.44	60/56	1.08/1.14
T0075	110	α	50/40	5.24/5.65	60/56	1.08/1.14

[a] The Protein Data Bank (Bernstein *et al.*, 1977) identifier for the initial test set, and target identifier for CASP3 predictions.

[b] For CASP3 proteins, we list the length of the target sequence from which we built our models. In some cases the length of the target sequence is larger than the protein size with experimentally determined coordinates, shown in Table 3.

[c] Walk length is half the number of residues for proteins up to size 76. For larger proteins in the test set, walk length is fixed at 38. For even larger proteins in CASP3, the walk length can be as long as 50.

[d] Edge size is the distance between adjacent vertices in the lattice. The edge size is chosen such that, on average, the volume per residue is 100 A^3. We have found that this estimate usually gives the best fit between the most accurate lattice models and their corresponding native structures.

[e] Number of vertices within the predetermined elliptical bounding volume. It is dependent on protein size. Here four different bounding volumes are used.

[f] R_g cutoff is the upper bound for radius of gyration, relative to that of a sphere with the same volume. R_g cutoff is predetermined and is solely dependent on protein size; the smaller the protein, the larger R_g cutoff is set.

[g] For some CASP3 targets, tetrahedral lattice conformations were generated with two different sets of lattice parameters.

phobic interaction in protein folding, it is able to discriminate native-like lattice structures with well-defined hydrophobic cores from random structures, even at very low resolution.

We characterize the overall performance of the lattice prediction procedure by the dRMS distribution of the low energy subset of all lattice structures. We show the dRMS distribution statistics for the 10,000 low energy structure subset for all test proteins in Table 2. Our lattice prediction procedure is moderately successful as a purifying step to concentrate promising structures. There is a wide spread of dRMS values in the low energy structure subset, and in many cases near-native structures are sampled within this small subset. For example, we are able to sample structures with dRMS as low as 4.10 Å in the 10,000 low energy subset for protein 1aa2 (108 residues). It is unfortunate that, due to its coarse grained nature, our energy function is unable to select out one best structure from the small set of low energy candidates. We, therefore, resort to the detailed structure construction and selection procedure described below.

Detailed structure construction and selection

Encouraged by the results of the lattice prediction procedure, we further purified the low energy subset of lattice structures by constructing all-atom structures and evaluating them using all-atom energy functions.

Even though our all-atom energy functions have previously proven powerful in comparative modelling tests (Samudrala & Moult, 1998), they have not been rigorously tested in an *ab initio* prediction scenario. For our selection scheme to work, it is crucial that the RMS range within which the discriminatory function is most sensitive matches the resolution of decoy structures generated by the lattice prediction method. With this in mind, for each test protein we generate all-atom models from 10,000 lowest energy structures created by the lattice prediction procedure, and test the performance of all-atom energy functions on these decoy sets. Since the lattice prediction generates a pair of mirror images for each chain configuration, we choose the conformation that has lower C^{α} root mean square deviation (cRMS) compared with

Table 2. Performance of lattice structure enumeration and selection

PDB code	Low energy dRMS (Å)[a]			All decoy dRMS (Å)[b]			Mean shift (Å)[d]
	Best	Mean	SD	Best[c]	Mean	SD	
1aa2	4.10	6.06	0.35	2.99	6.79	0.44	0.73
1beo	4.45	6.36	0.48	3.30	7.41	0.47	1.05
1ctf	3.35	5.45	0.45	2.81	6.34	0.44	0.89
1dkt-A	3.90	5.59	0.35	2.86	7.59	0.42	2.00
1fca	3.48	5.16	0.39	2.42	5.65	0.39	0.49
1fgp	4.23	5.98	0.41	3.15	6.62	0.46	0.64
1jer	5.55	7.53	0.41	4.22	8.42	0.39	0.89
1nkl	3.73	5.70	0.42	2.69	6.28	0.43	0.58
1pgb	3.87	5.62	0.39	2.61	6.23	0.39	0.61
1sro	4.67	6.27	0.38	3.11	7.26	0.45	0.99
1trl-A	4.11	5.99	0.48	2.60	6.25	0.51	0.26
4icb	3.58	4.99	0.40	2.76	6.23	0.40	1.24
Average	4.08	5.89	0.41	2.96	6.76	0.43	0.86

Our low detail prediction procedure is moderately successful as a purifying step to concentrate promising structures. In all 12 cases, using energy criteria for threading and selection pushes the structure population towards lower dRMS.

[a] dRMS distribution statistics (best, mean, and standard deviation) of the 10,000 lowest energy structure subset compared with the native structure.

[b] dRMS distribution statistics (best, mean and standard deviation) of all lattice decoy structures compared with the native structure.

[c] dRMS of the lattice structure with the highest number of native contacts. This is the structure that would be selected if a perfect energy function is used.

[d] Difference in the mean of the dRMS distribution between the complete structure set and the low energy structure set. It measures how effective selection by an energy function can push the structure population towards lower dRMS compared with the native structure.

native structure. This particular choice does not significantly affect the cRMS range and distribution of the decoy sets, and any discriminatory function that performs well in our decoy sets is likely to do well in blind prediction experiments. This method cannot be used for the CASP3 where we examined the structure and its mirror image for each decoy, since the experimental structure is not available.

Secondary structure prediction accuracy

We use the secondary structure prediction provided by the PHD PredictProtein Server (Rost *et al.*, 1993) as is, without further tuning of the multiple sequence alignments or the prediction results. The summary table (Table 3) reports the three-state accuracy (Q3) of the secondary structure predictions compared with the DSSP secondary structure assignments of the native structures (Kabsch & Sander, 1983). Q3 ranges from 54% to 97% for the test proteins, and is greater than 72% for five out of the 12 test proteins. On average, secondary structure predictions are better for α proteins than for β proteins.

Secondary structure fitting preserves overall topology of the lattice conformations

We use a greedy algorithm and a simple four-state model to incorporate predicted secondary structure into lattice structures. Figure 1 shows the cRMS difference distribution between structures before and after secondary structure fitting for the protein 1ctf. This distribution has a peak around 4 Å, and a long tail towards large cRMS. Within our lattice prediction scheme (roughly 6 Å cRMS), this simple fitting procedure preserves overall topology of lattice structures to a reasonable degree. The spread of the distribution reflects variation in the extent of agreement between predicted secondary structure and lattice structure topology.

Figure 1 also shows the cRMS distribution for both the lattice structure and all-atom structure sets compared with the native structure. Our fitting procedure preserves the cRMS distribution very well.

Combined energy function is able to achieve discrimination at low resolution

We tested a variety of different energy functions on the 12 test decoy sets. Any energy function will only be useful if it tends to assign lower energy to near-native structures. To demonstrate this, for each test decoy set we compute the average energy Z-scores of the top ten near-native conformations with lowest cRMS compared with the native structure. A negative Z-score would indicate that the energy function is capable of discriminating near-native structures from other structures. We find that three energy functions (RAPDF, HCF, Shell) stand out to give negative Z-score for the majority of the test proteins. Moreover, a simple combination of the three normalized energies performs better than any one of them alone (Table 4 and Figure 2). Our explanation for this is that the three energy functions are somewhat complementary: the HCF function favors compact structures, the shell function emphasizes long range hydrophobic interactions, whereas the RAPDF function encodes all-atom details including local geometry and side-chain interactions. As a result, combining the three

Table 3. Summary of overall performance for test set and CASP3 predictions

Protein	Size	Q3[a]	All cRMS range (Å)[b]	Best all cRMS (Å)[c]	Fragment size[d]	Prediction fragment cRMS (Å)[d]	Best fragment cRMS sampled (Å)[d]
A. *Initial test set*							
1aa2[e]	108	76	6.18-15.28	11.08			
1beo[e]	98	54	6.96-15.94	11.13			
1ctf	68	72	5.45-13.54	5.75			
1dkt-A	72	72	6.68-14.79	7.80			
1fca	55	78	5.09-12.06	5.90			
1fgp[e]	67	66	7.80-14.40	10.93			
1jer[e]	110	69	9.55-17.53	13.60			
1nkl[e]	78	78	5.26-14.23	5.70			
1pgb	56	57	5.60-13.30	8.41			
1sro[e]	76	65	7.30-15.42	9.68			
1trl-A	62	97	5.30-13.16	6.35			
4icb	76	86	4.74-13.28	4.95			
B. *CASP3 predictions*							
T0043	158	70	10.0-19.5	14.5	48	6.3	4.6
T0046	119	67	10.1-19.2	13.9	39	6.6	5.1
T0052	98	50	10.6-16.3	13.6	33	6.6	5.1
T0054[f]	202	-	-	15.5	202	15.5	-
T0056	114	100	6.2-17.8	13.0	60	6.8	3.3
T0059	71	80	7.4-15.7	11.6	46	6.7	5.4
T0061	76	62	6.0-14.0	10.1	66	7.4	5.6
T0063	135	60	10.8-22.0	15.1	35	6.4	4.0
T0064	103	90	8.0-18.8	11.2	68	4.8	4.5
T0065	31	90	2.4-7.6	4.1	31	4.1	2.4
T0074	98	88	6.3-16.5	11.3	60	7.0	4.2
T0075	88	78	6.0-17.0	9.8	77	7.7	5.5

[a] Percentage accuracy of the PHD three-state (helix, sheet, other) secondary structure prediction.

[b] The range of cRMS for all the all-atom conformations sampled.

[c] For each protein in the initial test set, we evaluate the cRMS between the experimental structure and the final model for all residues. For each target in CASP3 predictions, we evaluate the cRMS between the experimental structure and the best model out of five for all residues.

[d] For each target in CASP3 predictions, we select a continuous fragment that fits the experimental structure best in at least one of the five models, and compute cRMS between the fragment of the best model and the corresponding part of the experimental structure. We also compute the best cRMS between any fragment with the same size in the all-atom structures sampled and the corresponding part of the experimental structure.

[e] These proteins were targets from the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2).

[f] The experimental coordinates for T0054 were not made available during CASP3. The only data shown here were provided by the CASP3 organizers.

energy functions provides a better balance of different forces responsible for protein folding than any single energy function alone.

We also use other measures to assess the discriminatory power of the combined energy function, for example, correlation coefficient between energy and cRMS, energy rank of near-native structure with a certain cRMS cutoff, and average cRMS Z-scores for low energy conformations (Table 5). We emphasize that even with our best efforts, the combined energy function only achieves moderate success in discriminating near-native structures from other structures.

One consequence of the moderate discriminatory power of the energy function is that cRMS of the lowest energy conformation is very noisy. This is particularly evident in Figure 2(d): a near-native structure with cRMS of 5.3 Å is one of the three lowest energy conformations, but if we simply choose the one lowest energy conformation, its cRMS is almost 11 Å away from the native structure. Since our energy function is noisy and the three lowest energy conformations have very similar energy, it is not clear why we should choose one in favor of the other two conformations. Moreover, the lowest energy conformations share certain structural features of the native protein, though they are in many cases overwhelmed by the high energy of their specific non-native parts. We can enhance this shared structural features among lowest energy conformations by averaging the noise out in a proper way, thereby increasing the chance of finding near-native structures, increasing signal-to-noise ratio and making the prediction more robust. We use consensus-based distance geometry to perform proper averaging, the results of which are described in detail below.

Consensus-based distance geometry improves distribution of near-native structures

Our consensus-based distance geometry procedure is a proper averaging procedure over the set of candidate structures in distance space. Pre-

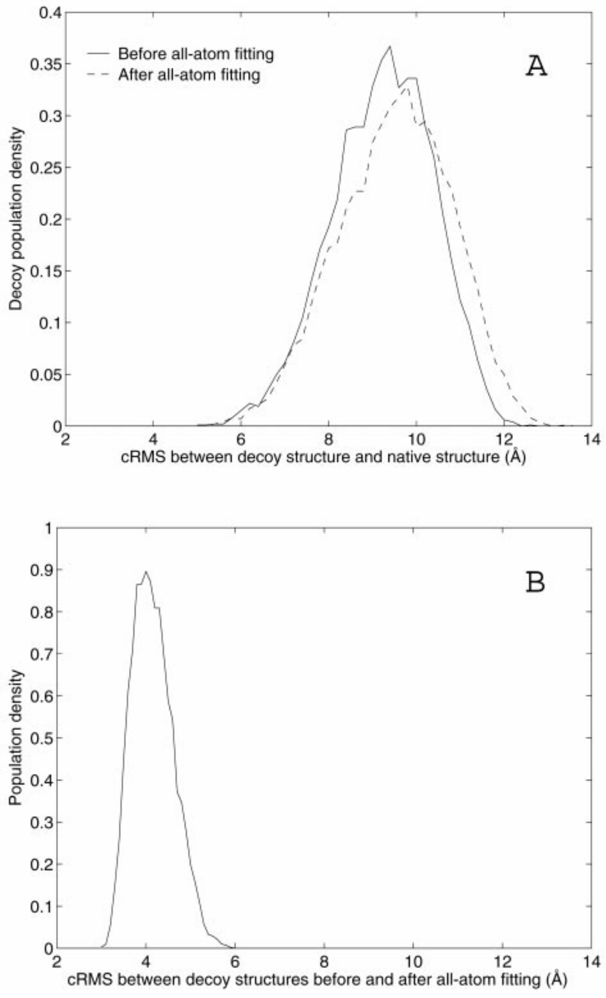

Figure 1. (a) The cRMS distribution of the decoy set compared with the native structure of protein 1ctf. Distributions for two decoy sets are shown: the low energy lattice structure set before secondary structure fitting, and corresponding all-atom structure set after fitting. We see that our secondary structure fitting procedure does not significantly change the cRMS distribution of decoy sets compared with the native structure. (b) Distribution of cRMS between structures before and after all-atom fitting for protein 1ctf. This plot shows that secondary structure fitting procedure preserves the overall chain topology of lattice structures.

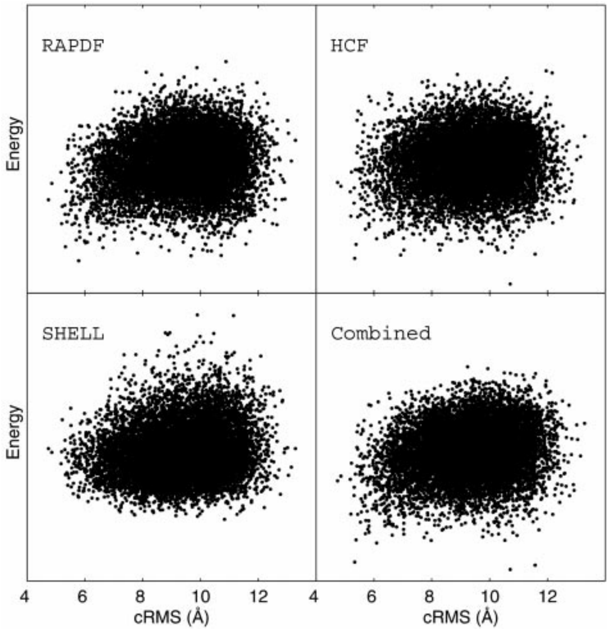

Figure 2. Energy *versus* cRMS plot for 10,000 lowest energy lattice structures of protein 4icb. Four energy functions are evaluated: RAPDF, HCF, SHELL, and the combined energy function. Even though the combined energy function seems to work best, the discriminatory power is limited for all energy functions tested.

vious studies have shown that consensus-based distance geometry procedure can generate a final structure that is better than one chosen randomly from a set of promising candidates generated in an *ab initio* manner, and even more so with the help of a more discriminating energy function (Huang *et al.*, 1998, 1999). We further test this approach on our decoy sets. We pick out 50, 100, and 500 lowest energy conformations (as ranked by the combined energy function) as input to the consensus-based distance geometry routine that produces one final structure for each set. Table 6 shows the cRMS of the output structures from distance geometry procedure, compared with random selection without distance geometry. Consensus-based distance geometry improves the concentration of near-native structures for six out of the 12 test proteins. In cases where sampling is ineffective, however, distance geometry does not improve prediction results, for example, for proteins 1aa2, 1beo, 1fgp, 1jer and 1sro.

The three output structures from consensus-based distance geometry only have C^{α} atoms; we then generate all-atom models for them and choose the lowest energy conformation as the final selection according to the RAPDF energy function.

Overall performance of all-atom construction and selection

Figure 3 shows both the cRMS distribution of all-atom structure decoy set and the cRMS of the final selection for the 12 test proteins. For nine out of 12 proteins, we sample the conformational space adequately to ensure that at least one conformation representing the native topology is included. For eight out of 12 proteins, the selection procedure is able to select out a final structure that is significantly better than a random selection from the decoy set. Overall, for five out of 12 proteins our final selection has the correct native topology and is roughly 6 Å in cRMS compared with the experimental structure. We show six of our best predictions in Figure 4.

Table 4. Performance of different discriminatory functions

Protein	RAPDF	HCF	Shell	Combined
1aa2	0.04	−0.10	0.02	−0.02
lbeo	0.02	−0.41	−0.55	−0.52
lctf	−0.44	−0.51	−1.01	−1.14
ldkt-A	−0.46	0.00	−0.46	−0.52
1fca	0.28	−0.38	0.07	−0.02
1fgp	0.23	−0.86	0.19	−0.23
1jer	−0.49	0.38	0.07	−0.03
1nkl	0.02	−0.23	0.10	−0.07
1pgb	0.62	−1.07	0.26	−0.12
1sro	0.02	−1.07	−0.07	−0.63
1trl-A	−0.16	−0.50	−0.77	−0.79
4icb	−1.69	−0.11	−1.49	−1.80
Average	−0.17	−0.40	−0.30	−0.49

This Table shows the average energy Z-scores for the ten lowest cRMS conformations with different discriminatory functions. Z-score is defined as the difference between the energy of the target structure and the average energy over the population, measured in units of standard deviation. The combined energy function on average has better discriminating power than other energy functions. However, the performance of these energy functions varies greatly depending on the protein.

We note that the cRMS of the final selection does not depend strongly with protein size, and we are able to make successful predictions for proteins that span a variety of different structural classes (all-α, α + β, and all-β). However, for large proteins (1aa2, 1beo, and 1jer) and some all-β proteins, like 1fgp, our procedure fails to select a structure with native-like topology. This is largely due to the poor sampling of the initial lattice walks. Indeed, large all-β proteins are poorly represented by simple lattice models. On the other hand, for the protein 1pgb, all-atom structures with cRMS less than 6 Å are sampled, but our selection procedure was unable to select them out.

Our predictions can tolerate relatively large errors in secondary structure predictions in terms of both sampling and final selection. For instance, we are able to sample structures with 5.6 Å cRMS for protein 1pgb even though the secondary structure prediction accuracy (Q3) is only 57 %; and our prediction for protein 1ctf has a cRMS of 5.75 Å when Q3 is 72 %. This is because we only use secondary structure information to generate all-atom models based on existing lattice structure topologies that are generated without secondary structure information. As a result, the cRMS distribution of low energy structures will not change much even when the secondary structure prediction error is large.

Ab initio prediction on CASP3 targets

Encouraged by these test results, we decided to participate in the CASP3 experiment, where our method was tested against target proteins in a double blind manner. We made *ab initio* predictions for 13 targets for the CASP3 experiment. Twelve out of the 13 predictions were made by the combined approach described here. CASP3 targets are in general larger than proteins in the test case,

Table 5. Performance of combined energy function (RAPDF + HCF + Shell)

Protein	top 50[a]	top 100[a]	top 500[a]	cRMS (Å)[b]	rank[c]	c.c.[d]
1aa2	0.06	0.11	−0.01	8.64	7	0.01
1beo	−0.45	−0.25	−0.10	9.11	7	0.02
1ctf	−0.73	−0.49	−0.33	5.76	33	0.19
1dkt-A	−0.56	−0.44	−0.33	6.97	3	0.15
1fca	−0.22	−0.34	−0.20	6.05	10	0.06
1fgp	−0.23	−0.26	−0.16	8.90	13	0.09
1jer	−0.23	−0.25	−0.16	10.80	17	0.06
1nkl	0.05	0.02	0.01	8.01	21	0.07
1pgb	−0.43	−0.53	−0.32	7.31	3	0.11
1sro	−0.48	−0.43	−0.35	9.10	27	0.14
1trl-A	−0.26	−0.16	−0.26	5.97	23	0.14
4icb	−0.71	−0.62	−0.44	5.33	3	0.18

Our combined energy function only achieves moderate success in discriminating near-native structures from other structures. Since cRMS of the lowest energy conformation is very noisy, consensus-based distance geometry is employed to obtain the final predicted structure.

[a] Average cRMS Z-scores for top 50, 100, and 500 lowest energy conformations.

[b] Native-like cRMS cutoff.

[c] Energy rank of best native-like structure using the combined energy function.

[d] Correlation coefficient between cRMS and energy for the whole population (10,000 structures).

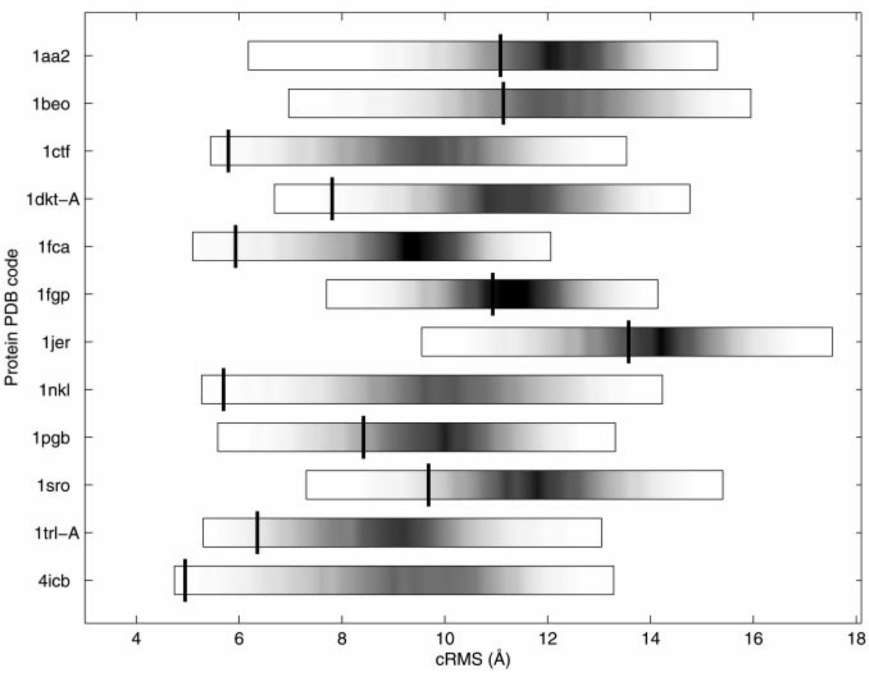

Figure 3. cRMS distribution of the sampled all-atom structures and the final selection for the 12 test proteins. For each protein, the population of 10,000 all-atom structures that we constructed from low energy lattice structures are shown by a box. Within the box, the distribution of cRMS for this structure population is represented by a shaded density bar, where the density of the shading at a given cRMS is proportional to the fraction of conformations present. The thick vertical bar indicates the cRMS of our final structure. We observe that we generally have better selection when the sampling is better (i.e. has more low cRMS structures).

and the prediction protocol was slightly modified to accommodate this change. For each target, up to 40,000 all-atom structures were sampled. We submitted up to five models for evaluation for each target. For six proteins, we are able to predict models that capture the global topology for large or all parts of the sequence. A summary of our prediction results is shown in Table 3. Four of the best non-trivial predictions are shown in Figure 5. Results of the individual performance of the method on each target was published in the CASP3 proceedings (Samudrala *et al.*, 1999). Here, as with the initial test set, we comprehensively analyze the performance within the framework of our hierarchical methodology. Our results at CASP3 represent a marked progress in *ab initio* prediction relative to what was achieved at CASP1 and CASP2.

The performance of our method at CASP3 is consistent with the previous tests. Both results follow the same trend: α proteins are easier to predict,

Table 6. Performance of the distance geometry procedure

Protein	Random[a]	SD[b]	Top 50[c]	Top 100[c]	Top 500[c]
1aa2	12.09	1.05	13.51	11.08	11.36
1beo	12.01	1.33	11.54	11.50	11.14
1ctf	9.20	1.29	5.79	6.32	6.94
1dkt-A	10.98	1.20	7.81	9.28	9.22
1fca	9.00	1.11	8.21	7.99	5.93
1fgp	11.16	0.92	10.93	11.52	11.36
1jer	13.79	1.17	14.16	13.57	15.31
1nkl	10.13	1.25	5.70	9.66	8.71
1pgb	9.45	1.21	8.42	8.77	9.49
1sro	11.42	1.08	9.68	9.88	12.47
1trl-A	8.62	1.17	6.36	7.83	7.83
4icb	8.78	1.63	8.95	9.67	4.95

In half the cases, consensus-based distance geometry procedure is able to generate one structure out of three trials that is much more native-like than a structure chosen at random.

[a] cRMS between the native structure and a structure chosen randomly from 500 lowest energy conformations.

[b] SD is the standard deviation associated with the above (a).

[c] cRMS between the native structure and the structure produced by distance geometry procedure from top 50, 100, and 500 lowest energy conformations.

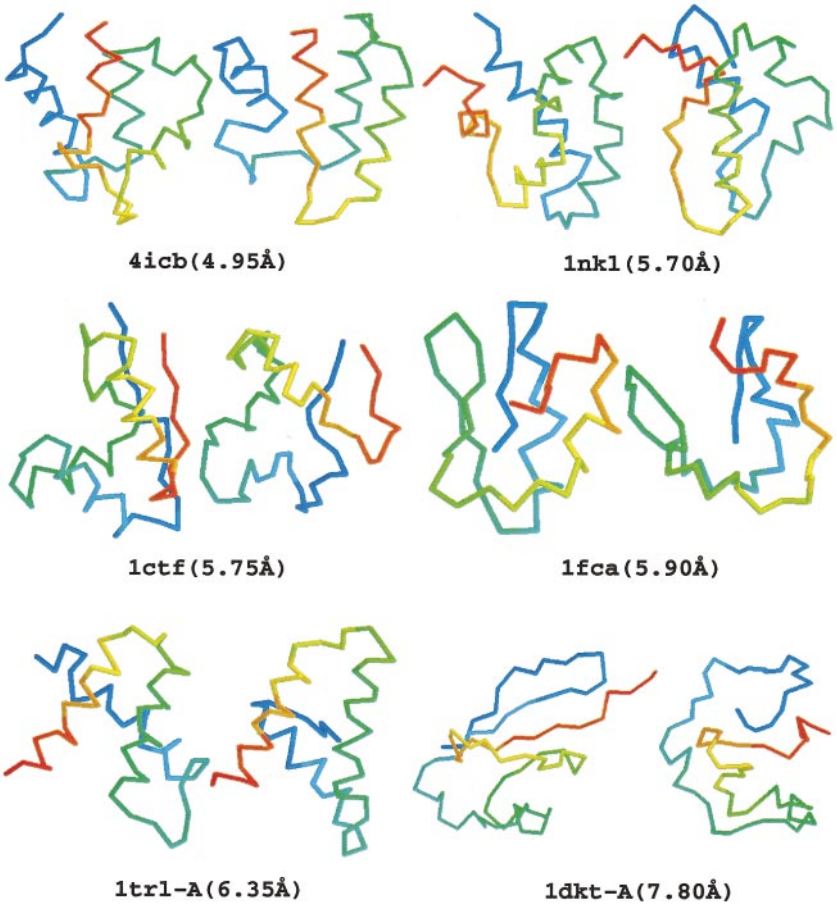

Figure 4. Native structures (left) and our final structures (right) for the following test proteins: 4icb, 1nkl, 1ctf, 1fca, 1trl-A, and 1dkt-A. The structures are colored according to sequence order (from the N terminus, blue, to C terminus, red).

and all-β proteins are hardest to predict. The sampling efficiency is similar in both cases: we are able to sample native-like all-atom structures for nine out of 12 test proteins, compared with for six out of 12 CASP3 targets. In both cases the final selection is typically 6-7 Å cRMS for 60 to 70 residues.

The performance of our method is comparable to other best methods at CASP3. We made six predictions out of the 11 medium and hard targets selected by the CASP3 assessors. Five of these predictions are among the five best of all groups judged by various fragment analyses (Orengo *et al.*, 1999). It is hard to make precise comparisons because the methods that performed well at CASP3 are very different from one another. For example, two methods use known sequence and structure information that is dependent on the specifics of current databases (Simons *et al.*, 1999; Ortiz *et al.*, 1999), and one method constructs tertiary structures by manually docking secondary structure elements (Lomize *et al.*, 1999). Our method is generally automated and does not rely on additional database information other than for creating the multiple sequence alignments used for the secondary structure predictions and for compiling the knowledge-based energy functions. However, the method is fairly tolerant of secondary structure prediction accuracy, and is not very sensitive to the specifics of the database of known structures. Comprehensive reviews on the CASP3 performances can be found elsewhere (Orengo *et al.*, 1999; Koehl & Levitt, 1999).

Computation times

For small proteins less than 80 residues, the computation time for each protein is roughly three CPU days on a 533 MHz alpha processor for the entire process. For the larger proteins in CASP3 experiments, the computation time is about one week. Our procedure can be trivially made to run in a massively parallel manner.

Advantages of this approach

Our method is relatively insensitive to the details of current protein sequence and structure databases. We only use these databases to compile knowledge-based energy functions and perform secondary structure predictions. Our prediction results are likely to be better on specific proteins by incorporating additional constraints derived from experiments or other statistical analysis performed on the data.

Our method is also tolerant with respect to errors in secondary structure prediction. Because we start with a complete low resolution enumeration, we are able to sample all chain topologies within a certain resolution and the prediction

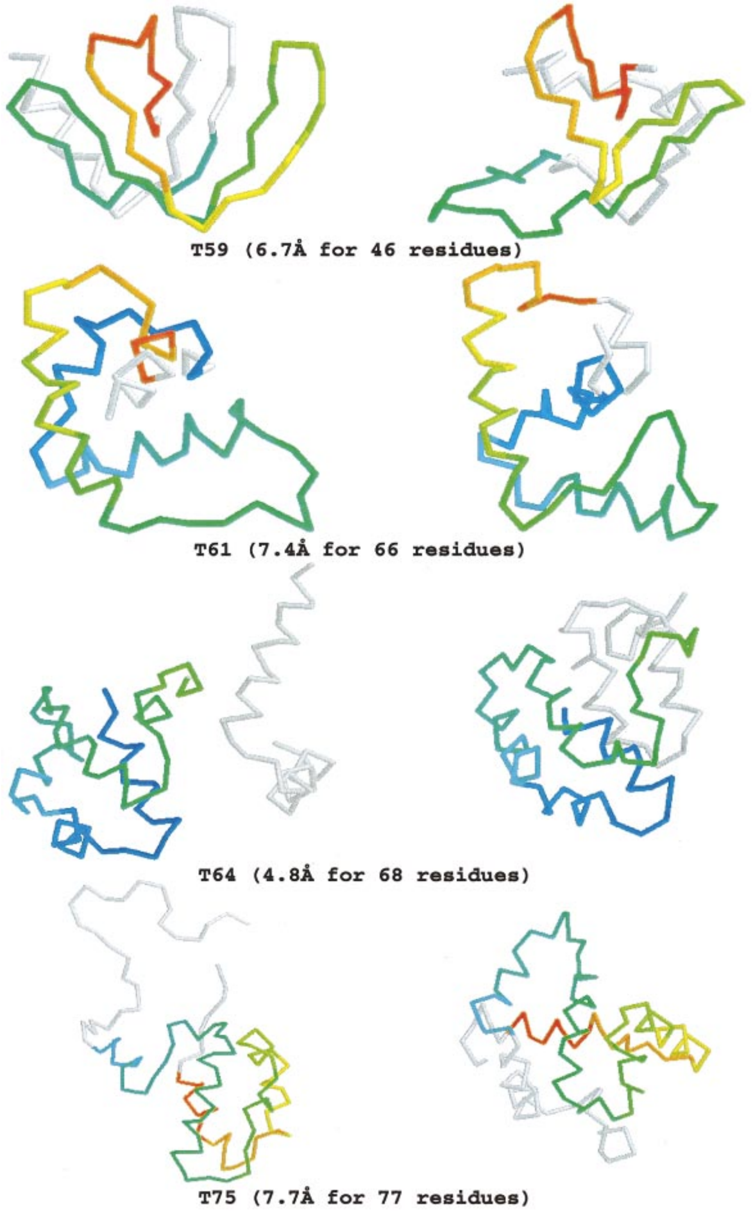
T59 (6.7Å for 46 residues)
T61 (7.4Å for 66 residues)
T64 (4.8Å for 68 residues)
T75 (7.7Å for 77 residues)

Figure 5. More successful CASP3 prediction results. Experimental structures (left) and our best predictions (right) for the following targets are shown: T0059, T0061, T0064, and T0075. For each target, we color according to sequence order the major fragment that gives best cRMS agreement between predicted and experimental conformations (T59: 25-70; T61: 4-69; T64: 1-68; T75: 27-103). Other parts of the structures are light gray. Additional results are shown in Table 3.

results are consistent over a wide range of protein folds.

Since our method decouples the complicated problem of protein structure prediction into small independent parts, we are able to evaluate and improve the performance of each part independently. Even though the specifics need to be much improved, we feel that the general paradigm employed in our approach, i.e. proceeding from low to high detail, from decoy generation to evaluation, and finally producing one conformation from many promising candidates, will remain useful in the design of better *ab initio* methods. We show that the discriminating power of all-atom knowledge-based energy functions extends beyond comparative modelling and threading to the *ab initio* folding scenario. We emphasize the importance of building all-atom conformations; simplified models have a very distorted energy surface that it is unlikely to consistently select one near-native structure from a decoy set for a wide range of proteins.

Our study also highlights the general applicability of consensus distance geometry method as an effective way to generate one final conformation from a set of promising candidates. We feel that it is an integral part of our approach to deal with the noise in the energy functions. We note that methods with similar philosophy have also been proposed by other recent *ab initio* studies, for example selecting structures with greatest number of neighbors (Simons *et al*., 1997) and clustering of structures (Eyrich *et al*., 1999).

Limitations of this approach

Our approach can predict to an accuracy of about 6-7 Å in cRMS for protein fragments of up to 80 residues, and is not restricted to particular structural classes. However, it fails for proteins with complicated supersecondary structure topologies. Sampling appears to be the bottleneck of our approach: the low resolution of the lattice model ultimately limits the sampling quality, thus putting an upper limit on the predictive power of this approach.

Our procedure generates protein models with low resolution. Such rough models are not very likely to be useful for functional studies in general (Wei *et al.*, 1999). However, when treated with caution and combined with experimental studies, our models may provide insights for further experiments in specific cases (Samudrala *et al.*, 2000a).

Directions for future work

Our prediction results can be improved by using predicted tertiary contacts (Ortiz *et al.*, 1998) and more accurate secondary structure predictions (Jones, 1999; http://globin.bio.warwick.ac.uk/psipred). In a more fundamental way, we need to overcome the sampling limit of the lattice model. This can be achieved by replacing the lattice model by knowledge-based off-lattice models, and replacing exhaustive enumeration by Monte Carlo minimization (Simons *et al.*, 1997).

Another area of improvement is the energy function. Knowledge-based energy functions have outperformed physical energy functions in many discriminatory tests. However, some promising physical energy functions have recently been proposed with discriminatory power comparable to knowledge-based energy functions, and with the advantage of clear underlying physics (Lazaridis & Karplus, 1999). Further testing is required to find the optimal energy function that works best for *ab initio* prediction.

We hope ultimately to generate better decoy sets that can fool the best energy functions and better energy functions that can discriminate the hardest decoy sets. We believe that this is the most powerful way to approach realistic *ab initio* protein structure prediction.

Methods

Overview

A typical flow chart of our procedure for protein tertiary structure prediction is shown in Figure 6. We describe the individual components of our combined hierarchical approach in detail below.

Lattice enumeration and selection

We represent the simplified chain topology of protein structure as a self-avoiding walk on a tetrahedral lattice. A full description of the methodology is given elsewhere (Hinds & Levitt, 1992, 1994). For small proteins with no more than 76 residues, we choose a walk length such that on average each vertex represents two residues. For larger proteins, we fix the walk length to an upper limit of 38 vertices. Lattice spacing between vertices is scaled based on the mean C^{α}-C^{α} distance obtained from a database of protein conformations. We also construct predefined elliptical bounding volumes. To ensure diversity of the lattice walks, these bounding volumes contain 20 % to 50 % more vertices than will be used by any particular structure.

We exhaustively enumerate all possible bounded lattice walks and pick out walks that are reasonably compact judged by a radius of gyration of no more than 1.14 times that of a sphere with the same volume. The criteria of compactness is also predefined: we allow less compact structures for short lattice walks because small proteins tend to be more irregular in shape. The total number of such compact lattice structures depends on the chain length but is never more than twenty million.

Since there are more residues than vertices, we thread the residues into every lattice walk using an iterative dynamic programming method that quickly converges to a locally optimal arrangement: no more than three residues are positioned between each pair of lattice points along the walk and each lattice point is occupied by a specific residue. After threading optimization, we calculate the energy for each lattice walk and the subset of structures with lowest energies are then selected for subsequent all-atom analysis.

The energy function we use is a residue-residue contact function. We count residue-residue contacts in a lattice structure in such a way that the total numbers of long-range contacts in lattice and actual structures are approximately the same. Contact energy parameters are derived from pairwise amino acid contact frequencies in a database of experimentally determined structures as:

$$e_{\mathrm{uv}} = -kT \ln \frac{\sum_{\mathrm{p}} C_{\mathrm{uvp}}}{\sum_{\mathrm{p}} \frac{C_{\mathrm{p}}}{T_{\mathrm{p}}} T_{\mathrm{uvp}}} \qquad (1)$$

where e_{uv} is the effective energy of a contact between amino acid types u and v, and p varies over all proteins in the database. For each protein p, C_{p} is the number of tertiary contacts, C_{uvp} is the number of u-v contacts, T_{p} is the total number of possible tertiary contacts, and T_{uvp} is the total number of possible u-v tertiary contacts. A tertiary contact is defined between two residues wherever a non-hydrogen atom of one residue approaches within 4.5 Å of a non-hydrogen atom of the other residue, and the two residues are at least five sequence positions away from one another.

Secondary structure prediction

We use the PHD PredictProtein Server (http://www.embl-heidelberg.de) (Rost *et al.*, 1993) to predict secondary structures of the sequences to be modeled. No manual adjustment was made to the predictions. We assign helix or sheet conformation to those residues with high confidence prediction from the PHD server (>5), and do not impose secondary structures on any other residues.

Secondary structure fitting and all-atom structure generation

The lattice structures from our simple lattice prediction only capture the overall chain topology and completely lack secondary structure and side-chain detail. We

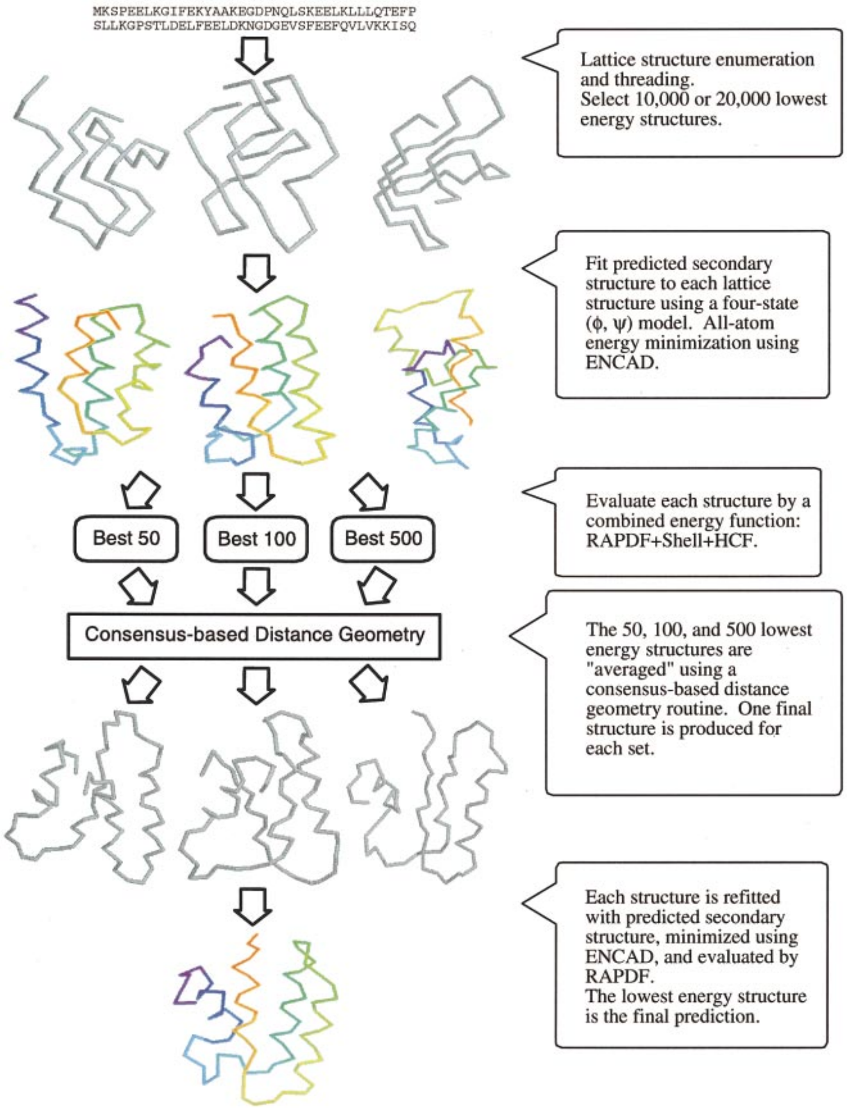
MKSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFP
SLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ
Lattice structure enumeration and threading. Select 10,000 or 20,000 lowest energy structures.
Fit predicted secondary structure to each lattice structure using a four-state (ϕ, ψ) model. All-atom energy minimization using ENCAD.
Best 50
Best 100
Best 500
Evaluate each structure by a combined energy function: RAPDF+Shell+HCF.
Consensus-based Distance Geometry
The 50, 100, and 500 lowest energy structures are "averaged" using a consensus-based distance geometry routine. One final structure is produced for each set.
Each structure is refitted with predicted secondary structure, minimized using ENCAD, and evaluated by RAPDF. The lowest energy structure is the final prediction.

Figure 6. Flowchart of our protein structure prediction procedure illustrated by a real example (4icb). Structures with C^α atoms only are shown in gray, whereas all-atom structures are shown in color.

use a greedy chain growth algorithm and a four-state off-lattice model to generate all-atom structures that closely resemble the chain topology of the lattice structure templates, while at the same time having the predicted secondary structure and side-chain detail.

We specify the main-chain conformation for each residue by a four-state (ϕ, ψ) model that has been shown to represent protein structures well (Park & Levitt, 1995). The four states have (ϕ, ψ) equal to (−57, −47) for helix, (−129, 124) for sheet, (−36, 108) and (108, −36) for two different turn conformations. Each residue with a high confidence secondary structure prediction is set to idealized helix or sheet conformation as described by the four-state model. For other residues we allow for all four conformations. Starting from the N terminus of the protein, we first enumerate all possible conformations for the first ten non-fixed residues using the four-state model, then select the 600 best conformations with lowest cRMS relative to the corresponding C^α atoms of the lattice structure. At each iteration, we add an additional residue in all four possible conformations at the C terminus of each of the 600 candidate fragments, and then again select the 600 best conformations. This is repeated until the entire lattice model is fitted.

All bond lengths and bond angles are fixed at idealized values. We build up side-chain conformations with χ angles fixed to those that are most frequently observed in a database of protein native structures. This has been shown to work surprisingly well for near-native template structures (Samudrala *et al.*, 2000b).

Energy minimization procedures

All-atom structures after secondary structure and side-chain fitting are minimized for 200 steps using ENCAD (Levitt & Lifson, 1969; Levitt, 1974; Levitt *et al.*, 1995).

Energy functions for all-atom models

We evaluate all-atom structures by a combination of three energy functions: (1) an all-atom distance-dependent conditional probability discriminatory function (RAPDF); (2) a hydrophobic compactness function (HCF); and (3) a residue-residue contact function (Shell).

We first normalize the energies of each function respectively, and then combine all three energies with equal weights.

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

RAPDF is an all-atom distance dependent knowledge-based energy function that indicates the probability of a conformation being native-like given a set of inter-atomic distances (Samudrala & Moult, 1998). We use a set of 312 unique folds from the SCOP database (Hubbard *et al.*, 1997) to compile the RAPDF energy function. We divide all non-hydrogen atoms into a total of 167 residue-specific atom types. We divide distances into a total of 18 distance bins: 1 Å bins from 3 Å to 20 Å, and one separate bin for the 0-3 Å range. The energy e_{ab} for a particular pair of atom types, a and b, is computed thus:

$$e_{ab} = -\ln \frac{N(d_{ab})/\Sigma_d N(d_{ab})}{\Sigma_{ab} N(d_{ab})/\Sigma_d \Sigma_{ab} N(d_{ab})} \tag{2}$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d in the database of experimental protein structures, Σ_d is summation over all distance bins d, and Σ_{ab} is summation over all pairs of atom types a and b.

The total RAPDF energy, evaluated by summing the energies for all distances and corresponding atom pairs, represents the negative log conditional probability that we are observing a native conformation. A complete description of RAPDF can be found elsewhere (Samudrala & Moult, 1998).

Hydrophobic compactness function (HCF)

Hydrophobic compactness function (HCF) measures the compactness of a structure. It is calculated using the following formula:

$$HCF = \frac{\Sigma_i((x_i - \bar{x})^2 + (y_i - \bar{y})^2 + (z_i - \bar{z})^2)}{N} \tag{3}$$

where N is the number of carbon atoms in the protein, and x, y, z are the Cartesian coordinates of the carbon atoms.

Residue-residue contact function (Shell)

The shell energy function (Park *et al.*, 1997) is a pairwise residue contact function. Two residues are said to be in contact if their interaction centers, located 3 Å from the C$^\alpha$ atom along the C$^\alpha$-C$^\alpha$ vector, are within 7 Å. The total energy for a conformation is then the sum of contact energies for all residue pairs that are in contact. The contact energy e_{uv} for residue types u and v is computed in a similar way to the energy function that we use for selecting lattice structures:

$$e_{uv} = -kT \ln \frac{\sum_p C_{uvp}}{\sum_p \frac{C_p}{T_p} T_{uvp}} \tag{4}$$

where for each protein p, C_{uvp} is the number of contact counts for residue types u, and v, C_p is the total number of contacts, and T_{uvp} is the number of residue pairs of type u and v separated by at least two residues in sequence. T_p, the total number of possible tertiary contacts, is calculated in the following way:

$$T_p = (N_p - 2)(N_p - 1)/2 \tag{5}$$

where N_p is the number of residues for protein p.

Consensus-based distance geometry

We use consensus-based distance geometry to produce a single Cartesian structure from a set of lowest energy conformations. Restraints for metric matrix distance geometry are taken directly from the lowest energy conformation sets by measuring and storing inter-C$^\alpha$ distance in 1 Å bins. The upper and lower bounds for each distance are determined by a jury process. Each distance receive a weight equal to the Boltzmann weight of the structure from which it was measured, i.e.:

$$w_i = \frac{\exp(-E_i/kT)}{\Sigma_i \exp(-E_i/kT)} \tag{6}$$

where E_i is the energy for the ith structure in the lowest energy set, and kT is set to 10. In the jury process, the distance bin that received the most weighted votes was used to set the upper and lower bounds for a given C$^\alpha$-C$^\alpha$ distance.

Distance geometry calculations are performed using the program distgeom from the TINKER suite (http://dasher.wustl.edu/tinker/) to compute a single Cartesian structure consistent with the most frequently observed C$^\alpha$-C$^\alpha$ distances in the lowest energy subset of conformations. The generated structure is refined *via* 10,000 steps of simulated annealing against a set of penalty functions to enforce local geometry, chirality, excluded volume, and the input distance restraints. Additional details can be found elsewhere (Huang *et al.*, 1998; Samudrala *et al.*, 1999).

Structure comparison

In our study of lattice prediction, we compare structures using the rmsd of corresponding C$^\alpha$-C$^\alpha$ distances (dRMS) (Cohen & Sternberg, 1980). Our lattice enumeration procedure only generates low resolution C$^\alpha$ structures with no secondary structure or side-chain information, and our energy function is based entirely on distance. As a result, the lattice prediction procedure does not discriminate between a structure and its mirror image. dRMS, which is based on distance and also does not discriminate between mirror images, is therefore a good measure of the performance of our lattice prediction procedure.

For our subsequent study of all-atom prediction, the symmetry in supersecondary structure level is broken due to handedness of secondary structure elements and side-chain conformations, hence mirror image lattice structures are readily discernible by the all-atom energy function. To evaluate the performance of our all-atom prediction procedure, we use the more familiar cRMS of two structures with best superposition (McLachlan, 1971).

Selection of test proteins

We select as a test set 12 small globular proteins with less than 110 residues representing different fold classes (Table 1). We choose half of these proteins from targets for the CASP2 meeting because they represent more realistic test cases. Test proteins were not used in compilation of the energy functions, i.e. our procedure is properly jack-knifed.

Differences in the CASP3 strategy

Target proteins in the CASP3 experiment were generally larger than our test proteins. For these larger proteins, we used a longer walk length of 50 within a bounded volume that contains 60 vertices, and only considered compact conformations with relative radius of gyration no larger than 1.08. To account for possible non-globular shapes, we also prepared another set of lattice models using walk length of 40 within a bounded volume that contains 56 vertices, and considered conformations with relative radius of gyration up to 1.14 (Table 1). Subsequently we sampled all compact lattice structures exhaustively, the total number of which is more than two billion for one protein. Each low energy conformation generated by our lattice prediction procedure is a pair of mirror images, and all-atom structures were generated for both of them. The resulting all-atom decoy set contains up to 40,000 structures for each target.

For secondary structure prediction, instead of taking the prediction result from PHD server alone, we generated 20 multiple sequence alignments of a homologous set of sequences to the target protein with a bootstrapping procedure, and used them as input for three secondary structure prediction methods: PHD (Rost *et al.*, 1993), DSC (Ross & Sternberg, 1996), and Predator (Frishman, 1995). The consensus of the 20 predictions for each method was taken as the final prediction.

Acknowledgments

We are extremely grateful to David A. Hinds for the original lattice enumeration and selection program, Patrice Koehl for providing us with efficient FORTRAN code to construct models given a set of $\phi/\psi/\chi$ angles and to calculate cRMS values, and Jay Ponder for the TINKER suite. This work was supported in part by a Howard Hughes Medical Institute Predoctoral Fellowship to Y.X., a Jane Coffin Childs Memorial Fund Fellowship to E.H., and a Burroughs Wellcome Fund Fellowship from the Program in Mathematics and Molecular Biology to R.S. Overall support was provided by NIH grant GM 41455 to M.L.

References

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science,* **181**, 223-230.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tsumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Chelvanayagam, G., Knecht, L., Jenny, T., Benner, S. & Gonnet, G. (1998). A combinatorial distance-constraint approach to predicting protein tertiary models from known secondary structure. *Folding Des.* **3**, 149-160.

Cohen, F. & Sternberg, M. (1980). On the prediction of protein structure: the significance of root-mean-square deviation. *J. Mol. Biol.* **138**, 321-333.

Duan, Y. & Kollman, P. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science,* **282**, 740-744.

Eyrich, V., Standley, D. & Friesner, R. (1999). Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J. Mol. Biol.* **288**, 725-742.

Frishman, D. (1995). Knowledge-based secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566-579.

Hansmann, U. & Okamoto, Y. (1999). New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.* **9**, 177-183.

Hinds, D. & Levitt, M. (1992). A lattice model for protein structure prediction at low resolution. *Proc. Natl Acad. Sci. USA,* **89**, 2536-2540.

Hinds, D. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668-682.

Huang, E., Samudrala, R. & Ponder, J. (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., Samudrala, R. & Ponder, J. (1999). *Ab initio* fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.* **290**, 267-281.

Hubbard, T., Murzin, A., Brenner, S. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236-239.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Koehl, P. & Levitt, M. (1999). A brighter future for protein structure prediction. *Nature Struct. Biol.* **6**, 108-111.

Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.* **35**, 133-152.

Lee, J., Liwo, A., Ripoll, D., Pillardy, J. & Scheraga, H. (1999). Calculation of protein conformation by global optimization of a potential energy function. *Proteins: Struct. Funct. Genet.* **S3**, 204-208.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44-45.

Levitt, M. (1974). Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.* **82**, 393-420.

Levitt, M. & Lifson, S. (1969). Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **46**, 269-279.

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. *Comp. Phys. Comm.* **91**, 215-231.

Lomize, A., Pogozheva, I. & Mosberg, H. (1999). Prediction of protein structure: the problem of fold multiplicity. *Proteins: Struct. Funct. Genet.* **S3**, 199-203.

McCammon, J. & Harvey, S. (1987). *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.

McLachlan, A. (1971). Test for comparing related amino acid sequences. *J. Mol. Biol.* **61**, 409-424.

Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **7**, 194-199.

Orengo, C., Bray, J., Hubbard, T., LoConte, L. & Sillitoe, I. (1999). Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction. *Proteins: Struct. Funct. Genet.* **S3**, 149-170.

Ortiz, A., 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**, 419-448.

Ortiz, A., Kolinski, A., Rotkiewicz, P., Ilkowski, B. & Skolnick, J. (1999). *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins: Struct. Funct. Genet.* **S3**, 177-185.

Osguthorpe, D. (1999). Improved *ab initio* predictions with a simplified, flexible geometry model. *Proteins: Struct. Funct. Genet.* **S3**, 186-193.

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

Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367-392.

Park, B., Huang, E. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831-846.

Pedersen, J. T. & Moult, J. (1996). Genetic algorithms for protein structure prediction. *Curr. Opin. Struct. Biol.* **6**, 227-231.

Ross, D. & Sternberg, M. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5**, 2298-2310.

Rost, B., Sander, C. & Scheider, R. (1993). 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., Huang, E. & Levitt, M. (1999). *Ab initio* protein structure prediction using a combined hierarchical approach. *Proteins: Struct. Funct. Genet.* **S3**, 194-198.

Samudrala, R., Xia, Y., Levitt, M., Cotton, N., Huang, E. & Davis, R. (2000a). Probing structure-function relationships of the DNA polymerase alpha-associated zinc-finger protein using computational approaches. *Proceedings of the Pacific Symposium on BioComputing*, 179-190.

Samudrala, R., Huang, E. S., Koehl, P. & Levitt, M. (2000b). Constructing side-chains on near-native main-chains for *ab initio* protein structure prediction. *Protein Eng.* In the press.

Scheraga, H. (1996). Recent developments in the theory of protein folding: searching for the global energy minimum. *Biophys. Chem.* **59**, 329-339.

Simons, K., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.

Simons, K., Bonneau, R., Ruczinski, I. & Baker, D. (1999). *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **S3**, 171-176.

van Gunsteren, W. (1998). Validation of molecular dynamics simulation. *J. Comput. Phys.* **108**, 6109-6116.

Wei, L., Huang, E. & Altman, R. (1999). Are predicted structures good enough to preserve functional sites? *Structure*, **7**, 643-650.

Edited by B. Honig

(Received 28 January 2000; received in revised form 3 April 2000; accepted 3 April 2000)