Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach

Yu Xia1, Enoch S. Huang2, Michael Levitt1* and Ram Samudrala1*

1Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305, USA
2Cereon Genomics, 45 Sidney Street, Cambridge MA 02139, USA

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.

© 2000 Academic Press

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\(^2\) 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 discriminatory 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 © 2000 Academic Press
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 (Chepurnov 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 surfaces 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

<table>
<thead>
<tr>
<th>Protein*</th>
<th>Sizeb</th>
<th>Class</th>
<th>Walk lengthc</th>
<th>Edge size (Å)d</th>
<th>Bounding box vertex counte</th>
<th>Rs cutofff</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Test set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1aa2</td>
<td>108</td>
<td>4</td>
<td>38</td>
<td>5.71</td>
<td>51</td>
<td>1.08</td>
</tr>
<tr>
<td>1beo</td>
<td>98</td>
<td>4</td>
<td>38</td>
<td>5.53</td>
<td>51</td>
<td>1.08</td>
</tr>
<tr>
<td>1ctf</td>
<td>68</td>
<td>2 + 4</td>
<td>34</td>
<td>5.08</td>
<td>50</td>
<td>1.10</td>
</tr>
<tr>
<td>1dkt-A</td>
<td>72</td>
<td>4</td>
<td>36</td>
<td>5.08</td>
<td>51</td>
<td>1.10</td>
</tr>
<tr>
<td>1ca</td>
<td>55</td>
<td>4</td>
<td>28</td>
<td>5.08</td>
<td>50</td>
<td>1.12</td>
</tr>
<tr>
<td>1gp</td>
<td>67</td>
<td>4</td>
<td>34</td>
<td>5.08</td>
<td>50</td>
<td>1.10</td>
</tr>
<tr>
<td>1jer</td>
<td>110</td>
<td>4</td>
<td>38</td>
<td>5.75</td>
<td>51</td>
<td>1.08</td>
</tr>
<tr>
<td>1kl</td>
<td>78</td>
<td>4</td>
<td>38</td>
<td>5.12</td>
<td>51</td>
<td>1.08</td>
</tr>
<tr>
<td>1pgb</td>
<td>56</td>
<td>2 + 4</td>
<td>28</td>
<td>5.08</td>
<td>50</td>
<td>1.12</td>
</tr>
<tr>
<td>1so</td>
<td>76</td>
<td>4</td>
<td>38</td>
<td>5.08</td>
<td>51</td>
<td>1.08</td>
</tr>
<tr>
<td>1trl-A</td>
<td>76</td>
<td>4</td>
<td>31</td>
<td>5.08</td>
<td>50</td>
<td>1.10</td>
</tr>
<tr>
<td>4icb</td>
<td>76</td>
<td>4</td>
<td>38</td>
<td>5.08</td>
<td>51</td>
<td>1.08</td>
</tr>
<tr>
<td>B. CASP3 predictions*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0043</td>
<td>158</td>
<td>2 + 4</td>
<td>50/40</td>
<td>5.92/6.37</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
<tr>
<td>T0046</td>
<td>119</td>
<td>4</td>
<td>50/40</td>
<td>5.38/5.80</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
<tr>
<td>T0052</td>
<td>101</td>
<td>4</td>
<td>50</td>
<td>5.08</td>
<td>60</td>
<td>1.08</td>
</tr>
<tr>
<td>T0054</td>
<td>202</td>
<td>2 + 4</td>
<td>51</td>
<td>6.38</td>
<td>60</td>
<td>1.08</td>
</tr>
<tr>
<td>T0056</td>
<td>114</td>
<td>4</td>
<td>50/40</td>
<td>5.31/5.72</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
<tr>
<td>T0059</td>
<td>75</td>
<td>4</td>
<td>38</td>
<td>5.08</td>
<td>56</td>
<td>1.14</td>
</tr>
<tr>
<td>T0061</td>
<td>89</td>
<td>4</td>
<td>45</td>
<td>5.09</td>
<td>60</td>
<td>1.08</td>
</tr>
<tr>
<td>T0063</td>
<td>138</td>
<td>4</td>
<td>50/40</td>
<td>5.66/6.09</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
<tr>
<td>T0064</td>
<td>111</td>
<td>4</td>
<td>50/40</td>
<td>5.26/5.67</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
<tr>
<td>T0065</td>
<td>57</td>
<td>4</td>
<td>29</td>
<td>5.08</td>
<td>60</td>
<td>1.80</td>
</tr>
<tr>
<td>T0074</td>
<td>98</td>
<td>4</td>
<td>49/40</td>
<td>5.08/5.44</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
<tr>
<td>T0075</td>
<td>110</td>
<td>4</td>
<td>50/40</td>
<td>5.24/5.65</td>
<td>60/56</td>
<td>1.08/1.14</td>
</tr>
</tbody>
</table>

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

† 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.

‡ 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.

§ 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 Å³. We have found that this estimate usually gives the best fit between the most accurate lattice models and their corresponding native structures.

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

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

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

In the selection process, we found that, due to its coarse grained nature, our lattice prediction method generates a wide spread of dRMS values in 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 modeling 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 Cα root mean square deviation (cRMS) compared with
natural 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.

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
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-
Previous 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 Ca 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.
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 1gp, 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,

<p>| Table 4. Performance of different discriminatory functions |
|-----------------------------|------------------|----------------|----------------|------------------|</p>
<table>
<thead>
<tr>
<th>Protein</th>
<th>RAPDF</th>
<th>HCF</th>
<th>Shell</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1aa2</td>
<td>0.04</td>
<td>-0.10</td>
<td>0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>1beo</td>
<td>0.02</td>
<td>-0.41</td>
<td>-0.55</td>
<td>-0.52</td>
</tr>
<tr>
<td>1ctf</td>
<td>-0.44</td>
<td>-0.51</td>
<td>-1.01</td>
<td>-1.14</td>
</tr>
<tr>
<td>1dkt-A</td>
<td>-0.46</td>
<td>0.00</td>
<td>-0.46</td>
<td>-0.52</td>
</tr>
<tr>
<td>1fca</td>
<td>0.28</td>
<td>-0.38</td>
<td>0.07</td>
<td>-0.02</td>
</tr>
<tr>
<td>1gp</td>
<td>0.23</td>
<td>-0.86</td>
<td>0.19</td>
<td>-0.23</td>
</tr>
<tr>
<td>1jer</td>
<td>-0.49</td>
<td>0.38</td>
<td>0.07</td>
<td>-0.03</td>
</tr>
<tr>
<td>1nkI</td>
<td>0.02</td>
<td>-0.23</td>
<td>0.10</td>
<td>-0.07</td>
</tr>
<tr>
<td>1pgb</td>
<td>0.62</td>
<td>-1.07</td>
<td>0.26</td>
<td>-0.12</td>
</tr>
<tr>
<td>1sro</td>
<td>0.02</td>
<td>-1.07</td>
<td>-0.07</td>
<td>-0.63</td>
</tr>
<tr>
<td>1trl-A</td>
<td>-0.16</td>
<td>-0.50</td>
<td>-0.77</td>
<td>-0.79</td>
</tr>
<tr>
<td>1icb</td>
<td>-1.69</td>
<td>-0.11</td>
<td>-1.49</td>
<td>-1.80</td>
</tr>
<tr>
<td>Average</td>
<td>-0.17</td>
<td>-0.40</td>
<td>-0.30</td>
<td>-0.49</td>
</tr>
</tbody>
</table>

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.

<p>| Table 5. Performance of combined energy function (RAPDF + HCF + Shell) |
|-----------------------------|------------------|----------------|------------------|</p>
<table>
<thead>
<tr>
<th>Protein</th>
<th>top 50*</th>
<th>top 100*</th>
<th>top 500*</th>
<th>cRMS (Å)*</th>
<th>rank*</th>
<th>c.c.†</th>
</tr>
</thead>
<tbody>
<tr>
<td>1aa2</td>
<td>0.06</td>
<td>0.11</td>
<td>-0.01</td>
<td>8.64</td>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td>1beo</td>
<td>-0.45</td>
<td>-0.25</td>
<td>-0.10</td>
<td>9.11</td>
<td>7</td>
<td>0.02</td>
</tr>
<tr>
<td>1ctf</td>
<td>-0.73</td>
<td>-0.49</td>
<td>-0.33</td>
<td>5.76</td>
<td>33</td>
<td>0.19</td>
</tr>
<tr>
<td>1dkt-A</td>
<td>-0.56</td>
<td>-0.44</td>
<td>-0.33</td>
<td>6.97</td>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>1fca</td>
<td>-0.22</td>
<td>-0.34</td>
<td>-0.20</td>
<td>6.05</td>
<td>10</td>
<td>0.06</td>
</tr>
<tr>
<td>1gp</td>
<td>-0.23</td>
<td>-0.26</td>
<td>-0.16</td>
<td>8.90</td>
<td>13</td>
<td>0.09</td>
</tr>
<tr>
<td>1jer</td>
<td>-0.23</td>
<td>-0.25</td>
<td>-0.16</td>
<td>10.80</td>
<td>17</td>
<td>0.06</td>
</tr>
<tr>
<td>1nkI</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>8.01</td>
<td>21</td>
<td>0.07</td>
</tr>
<tr>
<td>1pgb</td>
<td>-0.43</td>
<td>-0.53</td>
<td>-0.32</td>
<td>7.31</td>
<td>3</td>
<td>0.11</td>
</tr>
<tr>
<td>1sro</td>
<td>-0.48</td>
<td>-0.43</td>
<td>-0.35</td>
<td>9.10</td>
<td>27</td>
<td>0.14</td>
</tr>
<tr>
<td>1trl-A</td>
<td>-0.26</td>
<td>-0.16</td>
<td>-0.26</td>
<td>5.97</td>
<td>23</td>
<td>0.14</td>
</tr>
<tr>
<td>1icb</td>
<td>-0.71</td>
<td>-0.62</td>
<td>-0.44</td>
<td>5.33</td>
<td>3</td>
<td>0.18</td>
</tr>
</tbody>
</table>

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.

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

† Native-like cRMS cutoff.

‡ Energy rank of best native-like structure using the combined energy function.

§ Correlation coefficient between cRMS and energy for the whole population (10,000 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,

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).

Table 6. Performance of the distance geometry procedure

<table>
<thead>
<tr>
<th>Protein</th>
<th>Random*</th>
<th>SD*</th>
<th>Top 50*</th>
<th>Top 100*</th>
<th>Top 500*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1aa2</td>
<td>12.09</td>
<td>1.05</td>
<td>13.51</td>
<td>11.08</td>
<td>11.36</td>
</tr>
<tr>
<td>1beo</td>
<td>12.01</td>
<td>1.33</td>
<td>11.54</td>
<td>11.50</td>
<td>11.14</td>
</tr>
<tr>
<td>1cf</td>
<td>9.20</td>
<td>1.29</td>
<td>5.79</td>
<td>6.32</td>
<td>6.94</td>
</tr>
<tr>
<td>1dkt-A</td>
<td>10.98</td>
<td>1.02</td>
<td>7.81</td>
<td>9.22</td>
<td>9.22</td>
</tr>
<tr>
<td>1ca</td>
<td>9.00</td>
<td>1.11</td>
<td>8.21</td>
<td>7.99</td>
<td>5.93</td>
</tr>
<tr>
<td>1gp</td>
<td>11.16</td>
<td>0.92</td>
<td>10.93</td>
<td>11.52</td>
<td>11.36</td>
</tr>
<tr>
<td>1jer</td>
<td>13.79</td>
<td>1.17</td>
<td>14.16</td>
<td>13.57</td>
<td>15.31</td>
</tr>
<tr>
<td>1nkd</td>
<td>10.13</td>
<td>1.25</td>
<td>5.70</td>
<td>9.66</td>
<td>8.71</td>
</tr>
<tr>
<td>1pgb</td>
<td>9.45</td>
<td>1.12</td>
<td>8.42</td>
<td>8.77</td>
<td>9.49</td>
</tr>
<tr>
<td>1sro</td>
<td>11.42</td>
<td>1.08</td>
<td>9.68</td>
<td>9.88</td>
<td>12.47</td>
</tr>
<tr>
<td>1trl-A</td>
<td>8.62</td>
<td>1.17</td>
<td>6.36</td>
<td>7.83</td>
<td>7.83</td>
</tr>
<tr>
<td>4icb</td>
<td>8.78</td>
<td>1.63</td>
<td>8.95</td>
<td>9.67</td>
<td>4.95</td>
</tr>
</tbody>
</table>

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.

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

* SD is the standard deviation associated with the above (a).

* cRMS between the native structure and the structure produced by distance geometry procedure from top 50, 100, and 500 lowest energy conformations.
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
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
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 Cα-Cα 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 are 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:

$$\varepsilon_{uv} = -kT\ln \frac{\sum_p C_{uvp}}{\sum_p T_{uvp}}$$

where $\varepsilon_{uv}$ 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
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 ($\psi, \phi$) model that has been shown to represent protein structures well (Park & Levitt, 1995). The four states have ($\psi, \phi$) 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 $C^\alpha$ 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 $\gamma$ 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:

$$e_{ab} = -\ln \frac{N(d_{ab})/\Sigma dN(d_{ab})}{\Sigma abN(d_{ab})/\Sigma ab\Sigma abN(d_{ab})}$$

(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, $\Sigma d$ is summation over all distance bins $d$, and $\Sigma ab$ 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 a conformation is native-like given a set of inter-atomic contacts, and $\Sigma \Sigma ab$ 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:

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

(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 1 Å from the Cα atom along the Cα-Cα 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:

$$e_{uv} = -kT \ln \frac{\sum_p C_{uvw}}{\sum_p C_p T_{uvw}}$$

(4)

where for each protein $p$, $C_{uvw}$ is the number of contact counts for residue types u, v, $N_p$ is the total number of contacts, and $T_{uvw}$ 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:

$$T_p = (N_p - 2)(N_p - 1)/2$$

(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-Cα distance in 1 Å bins. The upper and lower bounds for each distance are determined by a jury process. Each distance receives a weight equal to the Boltzmann weight of the structure from which it was measured, i.e.:

$$w_i = \exp(-E_i/kT) \sum \exp(-E_i/kT)$$

(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 Cα-Cα 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 Cα-Cα 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 Cα-Cα distances (dRMS) (Cohen & Sternberg, 1980). Our lattice enumeration procedure only generates low resolution Cα 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 proteins using 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 rMDS 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 to Y.X., a Jane Coffin Childs Memorial Fund Fellowship to R.S. Overall support was provided by NIH grant GM 41455 to M.L.

**References**


Edited by B. Honig

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