Computational Methods for Protein Folding: Advances in Chemical Physics, Volume 120.
Edited by Richard A. Friesner. Series Editors: I. Prigogine and Stuart A. Rice.
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-20955-4 (Hardback); 0-471-22442-1 (Electronic)

AB INITIO PROTEIN STRUCTURE PREDICTION USING A SIZE-DEPENDENT TERTIARY FOLDING POTENTIAL

VOLKER A. EYRICH AND RICHARD A. FRIESNER

Department of Chemistry and Center for Biomolecular Simulation, Columbia University, New York, NY, U.S.A.

DARON M. STANDLEY

Schrödinger Inc., New York, NY, U.S.A.

CONTENTS

I. INTRODUCTION

In previous work [1–4], we have investigated the ability of simple potential functions, derived from statistics in the Protein Data Bank (PDB [5,6]), to generate correct predictions of protein tertiary structure given the native secondary structure as input. Most recently [2], we studied an unbiased sample of 95 proteins in the size range of 30–160 residues, and we were able to locate native-like low energy structures in a significant number of cases. However, there were also many examples of unsatisfactory performance; furthermore, the utilization of native secondary structure derived from PDB coordinates is an obvious limitation in terms of the utility of the method for protein structure prediction. Thus, a significant improvement in the potential function, along with tests under more realistic conditions, were required before one could consider applying the methodology to problems of practical interest.

A principal reason for carrying out the studies described above was to generate a large database of plausibly misfolded structures in the hope of elucidating systematic flaws in the database potential function that we employed, a principal component of which is the pairwise potential of mean force developed by Sippl and co-workers [7]. We have recently uncovered one systematic error in the Sippl formulation of the statistical pair potential, and we remedied this deficiency in a straightforward fashion: The potential function, at least as applied to the problems discussed here, should be dependent upon the size of the protein, a feature that has also been uncovered in other, more theoretical work [8]. To this end, we developed a statistical potential that is derived from proteins that are similar in size to the protein for which a prediction is to be made. The result is a new type of statistical pair potential with qualitatively improved predictive properties in tertiary folding simulations. While the new potential function is still not rigorously predictive of the native structure in all cases, application to actual protein structure prediction problems is now a much more feasible goal.

Having achieved this advance in the potential function, we relaxed the assumption of accurate knowledge of native secondary structure and examined the capabilities of the methodology with more realistic types of input data. In the present chapter, we approach this objective in two stages. First, we carry out simulations using ideal, rather than crystallographic, representations of the

secondary structure elements (while still deriving the location and length of the various elements from the PDB). For α-helices, the use of ideal helices leads in some (but not all) cases to a quantitative degradation of the quality of the results; in general, however, qualitatively similar success is achieved. For all α- and mixed α/β-proteins, there is an occasional substantial diminishment of the ranking of the lowest energy low RMSD structure, when idealized strands are used.

Second, we carry out computational experiments using secondary structure assignments derived from secondary structure prediction methods in conjunction with ideal secondary structural elements. This protocol constitutes an actual attempt at *ab initio* protein structure prediction; no experimental data other than sequence information is input into the calculations (other than, of course, the input of PDB statistics to derive the tertiary folding potential and secondary structure prediction algorithms). Because secondary structure prediction methods have not yet reached a high degree of robustness, we perform calculations using several different predictions generated by a variety of alternative secondary structure prediction methods (which are conveniently available on Web-based servers). While there are nontrivial cases where the native-like fold is uniquely determined by the algorithm, our objective at present is not to demonstrate successful *ab initio* prediction. Instead, we ask whether the protocol is capable of generating a prediction with a good RMSD that is highly ranked (e.g., within the top five predictions, a condition compatible with the rules of the CASP3 prediction contest). For a significant number of cases, this goal has been accomplished. Furthermore, in most cases where our algorithm fails to generate a native-like fold in the top five predictions, we are able to rationalize the results in terms of limitations of our model and propose straightforward extensions to generalize and improve the model. These proposed extensions are briefly discussed at the end of this chapter.

We have chosen in this chapter to focus our efforts on α-helical and mixed α/β-proteins below 100 residues in size. In previous work [2] we showed that β-strand proteins present more of a challenge to our prediction methodology than α-helical or mixed α/β-proteins [9–12]; the modified size-dependent potential function discussed above improves the results of earlier work on β-strand containing proteins, but does not change the basic conclusion. For larger systems, our results are quite promising but not yet at the stage of completeness that we have been able to achieve for the smaller proteins. Consequently, we defer discussion of these cases to a subsequent publication.

The chapter is organized as follows. Section II describes the new potential function, discussing its novel qualitative features and presenting an algorithm for optimization of parameters using a large training set derived from the PDB. Section III briefly reviews the computational methodology used to carry out the tertiary folding simulations (previously described in detail [2]) and then presents simulation results using native secondary structure and ideal secondary

structure. As a test set in this section, we employ a subset of the proteins studied previously [2] so that comparisons can be made with the results reported in that publication, and improvements in the potential functions quantified. In Section IV, we utilize predicted secondary structure lengths and positions and ideal secondary structure elements to carry out *ab initio* prediction experiments; we focus in this chapter on helical proteins, and include, in addition to proteins from the test set of Section IV, two targets from CASP3 [13]. Section V, the conclusion, summarizes our efforts.

II. DEVELOPMENT OF A SIZE-DEPENDENT POTENTIAL ENERGY FUNCTION

A. Identification of Systematic Errors in Previous Tertiary Folding Simulations

Although the tertiary structure prediction protocol employed in our previous work [2] was more or less able to consistently generate native-like structures for α- and mixed α/β-proteins, the energetic rank of these structures was not always satisfactory. An analysis of high-RMSD, low-energy structures obtained from those simulations reveals a systematically incorrect behavior of the statistical potential function of Sippl and co-workers [7] at large separations, most prominently for pairs of hydrophilic residues. This feature of statistical potentials has been uncovered in several other computational experiments [8,14].

The hydrophobicity term developed by Sippl was originally used only for recognition (i.e., threading), so it is not surprising that some modifications would be required for the asymptotic large-distance parts of the energy surface. It remains to be seen whether or not the general type of systematic errors uncovered in our tertiary structure predictions are present in the threading studies of others using similar potentials. A complete derivation of the coefficients by Sippl and co-workers can be found in Ref. 7. The two key elements of interest in the derivation of the hydrophobicity function are the inclusion of proteins of many sizes in the definition of a statistical ''potential of mean force'' (PMF) and the asymptotic behavior of these potentials when they are linearly extrapolated to large distances.

In Ref. 7 an individual PMF for residues i and j, separated by a distance d, is defined as

$$E_{ij} = -kT \ln\left(\frac{p_{ij}^{1}(d)}{p^{2}(d)}\right) \tag{1}$$

where $p_{ij}^{1}(d)$ is the normalized distribution of d for all i,j pairs in a training set and $p^{2}(d)$ is the normalized distribution of d of irrespective of residue pair. The

training set Sippl used consisted of 88 proteins that ranged in length from 46 to 374 residues. Note also that Eq. (1), which is sometimes known as the ''quasi-chemical approximation,'' applies only to residues separated in sequence by more than 20 amino acids (at least in Ref. 7).

Equation (1) is only defined for distances that correspond to nonzero values of both distribution functions. For this set of distances, E_{ij} is well-approximated by a linear function

$$E_{ij}^{\text{hyd}} = (H_{ij} + H_0)\, d \tag{2}$$

where H_{ij} is one of 400 ''pairwise hydrophobicities'' and H_0 is an adjustable ''average hydrophobicity,'' for which Sippl suggest the value 0.36. (In our own simulations, H_0 was increased if local minimization starting from the native structure yielded noncompact structures.)

The basic idea inherent in the development of the Sippl hydrophobicity potential, that of extracting a potential of mean force using PDB statistics, is an essential component of our empirical tertiary folding potential. However, based on our analysis of the low-energy misfolded structures generated in our previous experiments [2] described above, we propose to improve upon the detailed methodology for construction of the PMF by implementing the following modifications:

1. The derivation of an individual PMF for tertiary structure prediction of protein P is to be based only on proteins of roughly the same size as P.
2. In the large and small distance limits, a functional form other than Eq. (1) is to be used. The precise representation of the potential that we use to accomplish this is described below.

The first of these objectives appears rather straightforward to implement. However, a reduction in the number of proteins used to derive the distributions means we will most likely reduce the signal to noise ratio in the PMF. We addressed this problem in the following fashion. At short range, where no systematic errors were observed, we generated the usual distance statistics for each amino acid pair, averaging over proteins of various sizes. In addition to considering amino acid type, we also took into account the secondary structure type (α-helix, β-strand, loop/coil) of the residue pair for short-range statistics. At a pair separation larger than a cutoff distance R_0 (a value of 15 Å was used in all calculations), we grouped the amino acids together according to hydrophobicity. A total of four classes are defined (Table I). The statistics of residue pair i, j were grouped together with those of pair j, i so the total number of pairs was given by $N_{\text{class}}[N_{\text{class}} - 1]/2 + N_{\text{class}}$.

The reduction in the number of pairs from 210 to only 10 offsets the reduction in the number of proteins well enough that we can obtain an adequate

TABLE I
Hydrophobicity Class[a]

Class	Amino Acids
Weakly hydrophobic	Ala, Cys, His, Leu, Met, Phe, Tyr
Strongly hydrophobic	Ile, Trp, Val
Weakly hydrophilic	Asn, Gln, Gly, Pro, Ser, Thr
Strongly hydrophilic	Arg, Asp, Glu, Lys

[a]The definitions used to bin long-range distance statistics according to hydrophobicity are listed.

signal-to-noise ratio. The justification for this approach is that at large separation the probability distribution should not be sensitive to the specifics of the amino acid pair (e.g., the size of the side chain) but only to the propensity to reside on the surface of the protein as opposed to the interior. Support for this idea comes from the work of Yue and Dill [15], who carried out tertiary folding simulations with fixed secondary structure for a series of small proteins, many of which were also studied by us using a Sippl-based potential. What is striking is that, although Yue and Dill used only a two-letter code (hydrophobic and hydrophilic), in many cases their results were qualitatively similar to the ones we obtained using a much higher level of detail in the amino acid pair functions. This suggests that the considerably less drastic simplification we are making (including the retention of a fully detailed pair distribution for short distances, allowing packing effects to be described more accurately) is plausible, although this must of course be validated by the actual results.

The proteins are binned according to radius of gyration using the following formula

$$\text{size} = \text{int}(15Rg^{1/3} - 29) \tag{3}$$

where int(x) is the largest integer that is less than or equal to the real number x. Once the long- and short-range pair statistics are accumulated, they can be spliced together to generate a complete distribution for each amino acid pair. The assumption is that in the region around R_0, the individual pair distributions have already converged toward the hydrophobicity class pair distributions. By appropriately scaling the data, a potential valid over all distance ranges is generated for each amino acid pair in each size class.

The second modification was implemented by setting the PMF to a constant at distances outside of the observable range:

$$E_{ij} = \begin{cases} -kT\ln(\varepsilon_1) & (d < d_{\min}) \\ -kT\ln\left[\frac{p_{ij}^1(d)}{p^2(d)}\right] & (d_{\min} < d < d_{\max}) \\ -kT\ln(\varepsilon_2) & (d > d_{\max}) \end{cases} \tag{4}$$

where $d_{\min}$ and $d_{\max}$ are the lower and upper bounds, respectively, on the distance range over which we were able to collect good distance statistics (the distribution function had to be greater than or equal to 0.001). The parameters ϵ_1 and ϵ_2 were also set to 0.001. In addition to the residue pair potential above, we included a second long-range energy term that is somewhat analogous to the average hydrophobicity H_0 in the linear case, in that it ensures compactness. This term, which we will refer to as the density profile, is given by

$$E_{ij} = \begin{cases} -kT \ln(p^2(d_x)) & (d < d_x) \\ -kT \ln(p^2(d)) & (d_x < d < d_{\max}) \\ -kT \ln(\varepsilon_2) & (d > d_{\max}) \end{cases} \tag{5}$$

where d_x is the distance at which the residue independent distribution function $p^2(d)$ is a maximum. The final long-range energy is a linear combination of Eqs. (4) and (5) (with weights 1 and 0.6, respectively). The optimization of the density profile in the scoring function is a key ingredient in properly constraining the potential in the large separation limit.

In Eqs. (1)–(5) inter-residue distances are defined in terms of a single side-chain interaction point. This point, which we will refer to for simplicity as C_β, is actually the projection of the average side-chain geometric center onto the C_α–C_β bond vector.

The only function that depends on distances other than C_β–C_β is the excluded volume potential, which depends on C_α–C_α, C_α–C_β, and C_β–C_β distances. The functional form of the excluded volume term is the same as in previous work [16]:

$$E_{ij}^{\text{exvol}} = \exp\left(-\left(\frac{d_{ij}}{d_{ij}^0}\right)^{10}\right)$$

where the width of the excluded volume region d^0 is derived from the distance of closest approach for the residue pair in question in the training set.

Equations (4) and (5) are not evaluated explicitly in the minimization program, but are fit using a combination of spline [17] methods, which provide stability, the ability to filter noise easily, and the flexibility to describe an arbitrarily shaped potential curve. Moreover, the final functional form is inexpensive to evaluate, making it amenable to global minimization. The initial step in our methodology is to fit the statistical pair data for each amino acid and for the density profile to Bezier splines [17]. In contrast to local representations such as cubic splines, the Bezier spline imposes global as well as local smoothness and hence effectively eliminates the random oscillatory behavior observed in our data.

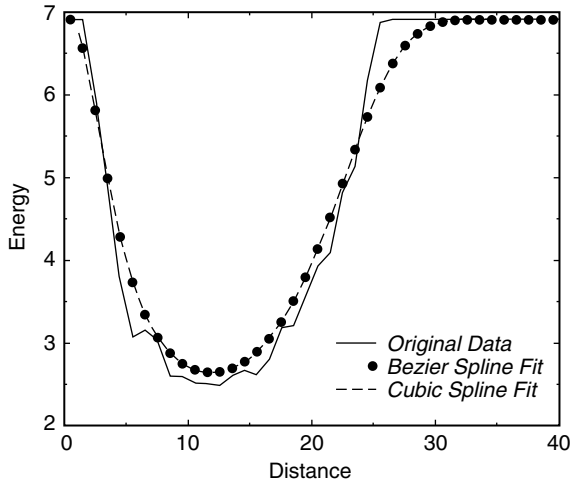


Figure 1. Data smoothing via Bezier and cubic splines. Bezier splines are shown as circular data points which approximate a typical noisy density profile (black line). Cubic splines (dashed line) are then fit to the Bezier data points (at a higher resolution than is shown here).

While Bezier splines are an optimal approach for smoothing noisy data, they cannot be rapidly evaluated using local interpolation methods. We therefore next fit a cubic spline to the Bezier spline curve. Figure 1 compares the Bezier spline and cubic spline curves for the same dataset; it can be seen that there is no meaningful difference between the two. Cubic splines can be evaluated rapidly at an arbitrary value of the residue pair separation using a standard interpolation formula (see, e.g., Ref. 17 for details). The spline coefficients needed for carrying out the interpolation are preprocessed and stored in fast memory during the simulation; the computational effort required to evaluate the spline potential is not much larger than that, for example, to determine the inter-residue distance.

B. Further Improvement of the Potential Energy Function

As Eq. (5) shows, the original form of the PMF used by Sippl and co-workers (1)] remains essentially intact in regions where good statistics are available, although more weight is given to the density distribution. The validity of treating different amino acid pairs as essentially independent, as in Eq. (1), has recently been questioned by Thomas and Dill [18]. They proposed an improved approach based on an iterative algorithm, the goal of which is to have the Boltzmann distribution of distance pairs associated with the potential energy function agree with the distribution derived from native structures. The following are the

components of the iterative cycle:

1. Initialize the potential to values obtained from the quasi-chemical approximation.
2. Use this potential to generate structures; determine the relevant distribution functions (in our case, residue pair separation probabilities) from the simulated data.
3. If there are deviations between the two, the potential is corrected so as to minimize them.
4. A simulation is carried out with the new potential, and a new set of statistics is generated.
5. Steps 3 and 4 are repeated until the deviation between the statistics from the simulated data and the experimental data have been reduced to an acceptable level.

For tertiary folding, there are three major problems in implementing this strategy. First, generation of simulated data is computationally expensive if a large training set is to be used. Second, one has to define the ensemble of simulated structures from which to extract statistics. For example, does one keep, only the lowest-energy structure for each protein or keep an ensemble of low-energy structures? Third, there is the question of how to update the potential function. In what follows, we adopt a heuristic approach to these issues; the protocols presented here represent preliminary explorations of this strategy and no doubt can be improved upon. In the present work we have chosen to optimize the potential function by comparing the distribution of locally minimized native structures with that of the native structure itself. The idea is that if the minimized native structure is as close to the native structure as possible, the basin of attraction associated with the minimized native will yield acceptable low RMSD predictions. From numerous computational experiments that we have carried out, resemblance of the minimized native structure to the native structure is clearly a *necessary* condition for obtaining useful predictive results; if the minimized native structure has, for example, a high RMSD from the native, one typically will fail to locate anything reasonable in a full-scale tertiary folding simulation starting from an unfolded state. Whether this is a *sufficient* condition for robust results in such simulations is one of the principal subjects of the present chapter. We briefly summarize here the entire optimization cycle, drawing on the results of the previous sections as well as on the basic idea described above. The steps of the optimization cycle are outlined as follows:

1. Initialization:
   a. The training set of native structures, with secondary structure assigned by DSSP [19], is read into the optimization program.

   b. Proteins are sorted into size bins according to their radius of gyration using Eq. (3).
   c. The iteration counter is initialized to zero ($it = 0$).
   d. A potential energy function E_0 is computed from distance distribution functions based on the native structures.
2. All native proteins are locally minimized using E_{it}.
3. A potential energy function E_{min} is computed for each size bin based on statistics derived from the minimized structures.
4. The difference between E_{min} and E_0 is calculated:

$$E_{\text{diff}} = E_0 - E_{\text{min}}$$

5. The iteration counter is incremented and E_{it} is updated by adding a correction that is proportional to E_{diff}:

$$it = it + 1$$
$$E_{\text{it}} = E_{\text{it}-1} + k_{\text{diff}} E_{\text{diff}}$$

   (A proportionality constant equal to 0.1 was chosen empirically so as to damp oscillations in the optimization procedure.)
6. Steps 2–5 are repeated until substantive improvements are no longer produced in the RMSDs of the minimized native structures.

In addition to the RMSD, the energy gap between the native and the minimized native was monitored. The smaller this energy gap, the better in general we have observed the performance of the potential to be in tertiary folding simulations. In our initial efforts we utilized a more elaborate short-range potential function that, in addition to the C_β–C_β term described in Section II A (above), included both C_β–C_α and C_α–C_α terms. The additional terms involving C_α were included in the iteration process described above. Subsequently, however, the extra terms in the short-range potential were not used in the tertiary structure predictions, because we did not see an overall improvement in the results when they were included. Another important difference between the potential energy function used in the above iterative procedure and the one used in actual tertiary structure predictions involves the density profile function. In the iterative procedure, this function was not flattened at $d < d_x$ [see Eq. (5)]. However, we found that we could improve the ranking of native-like structures with this simple modification. Thus the improvement of the potential energy function was ultimately achieved by a combination of the iterative algorithm described above and manual inspection of the individual terms after parameter optimization.

Because it is computationally expensive to carry out global minimizations on a large test set, we are unable to objectively determine the amount of improvement with respect to the zeroth-order potential (E_0) realized by the optimization

procedure outlined above. But given the fact that several proteins, which were unstable in local minimizations starting from the native using E_0, yield acceptable RMSDs using the optimized potential, we believe that parameter optimization can effectively remedy some of the deficiencies of reduced model approaches. The issue of parameter optimization along the lines of the procedure outlined above as well as other approaches in the literature (for a review see Ref. 20) will be the subject of future work.

C. Resulting Potential Energy Function

Table II lists the proteins used in the training set, a subset of the PDB Select database of nonhomologous proteins [21]. We avoided inclusion of proteins that form dimers (or other oligomers) in solution because one would expect the distributions in this case to be significantly altered due to the oligomerization process. For each protein we list the PDB code, number of residues, radius of gyration, and classification in our size bin scheme.

Figures 2–4 show the size dependence of three representative terms in Eq. (4) (after being fit to splines, as described below) for the amino acid pairs arginine–arginine, arginine–isoleucine, and isoleucine–isoleucine for the first six size bins (the bins relevant to the prediction results discussed in this chapter). Figure 5 shows the density profile [Eq. (5)] for the same size bins. Note that because the total energy is a linear combination of Eqs. (4)–(6), the oscillatory behavior at large distances (>15 Å) of the potentials in Figs. 2–4 is effectively masked by the density profile; in the short-distance limit, the excluded volume term serves a similar purpose. The energy plots in Figs. 2–4 show clearly that a linear function is a good approximation over the most populated distance ranges (10–20 Å). Moreover, the slopes in these regions can

TABLE II
Training Set[a]

Size Bin	PDB Name	N_{res}	R_g	Size Bin	PDB Name	N_{res}	R_g	Size Bin	PDB Name	N_{res}	R_g
1	1chl	36	8.8	5	1svr	94	12.1	7	1bvh	153	14.5
1	1erd	35	8.4	5	1vcc	77	12.1	7	1c25	154	14.8
1	1ret	37	8.8	5	1wkt	88	12.1	7	1cdb	101	14.0
1	2erl	35	8.2	5	2abd	86	12.6	7	1cfe	135	14.0
1	3bbg	40	8.7	5	2bby	69	12.0	7	1chd	198	15.0
2	1bor	52	9.3	5	2ezh	65	11.9	7	1cur	150	14.2
2	1dec	39	9.7	5	2fow	76	11.8	7	1def	147	14.0
2	1gps	47	9.6	5	2hgf	97	12.5	7	1eal	127	14.3
2	1sco	38	8.9	5	2hp8	68	11.7	7	1hfc	157	14.6
2	1zwa	29	9.1	5	2rgf	93	12.5	7	1ido	184	14.9
2	2bds	43	9.3	5	2sxl	88	12.6	7	1jpc	108	14.1
3	1afp	51	9.8	6	1a1x	106	13.5	7	1lcl	141	14.3

TABLE II (*Continued*)

Size Bin	PDB Name	N_{res}	R_g	Size Bin	PDB Name	N_{res}	R_g	Size Bin	PDB Name	N_{res}	R_g
3	1afp	51	9.8	6	1a1x	106	13.5	7	1lcl	141	14.3
3	1apf	49	9.7	6	1a2p.A	108	13.6	7	1mak	113	14.0
3	1ark	56	9.9	6	1acz	108	13.8	7	1mup	157	14.7
3	1awo	57	10.4	6	1bea	116	13.6	7	1mut	129	14.6
3	1brf	53	10.1	6	1bfg	126	13.0	7	1poa	118	14.3
3	1cka.A	57	10.1	6	1bkf	107	13.3	7	1rcf	169	14.5
3	1tih	53	10.6	6	1btn	106	13.1	7	1svp.A	155	14.8
3	1zaq	44	9.9	6	1buz	116	13.2	7	1vhh	157	14.5
3	2brz	53	10.5	6	1bw3	125	13.7	7	2a0b	118	14.7
3	5pti	55	10.6	6	1c52	131	13.5	7	2ezl	99	14.7
4	1ab7	89	11.6	6	1exg	110	13.6	7	2hbg	147	14.7
4	1ah9	66	10.9	6	1fna	91	13.4	7	2hfh	93	13.9
4	1c5a	65	11.2	6	1hcd	118	13.4	7	2i1b	153	14.7
4	1ehs	48	11.6	6	1irs.A	108	13.4	7	2sns	136	14.4
4	1hoe	74	11.4	6	1jer	110	13.5	7	2vil	126	14.0
4	1kbs	60	11.3	6	1krt	110	13.6	7	3cyr	102	14.2
4	1leb	72	11.3	6	1ksr	100	13.8	7	5p21	166	14.8
4	1msi	66	10.7	6	1kte	105	13.2	8	1amx	150	15.4
4	1nkl	78	11.3	6	1kuh	132	13.6	8	1aqb	175	15.8
4	1opd	85	11.6	6	1lit	131	13.4	8	1atl.A	200	15.9
4	1pih	73	10.9	6	1lou	97	13.2	8	1ble	161	15.1
4	1pou	71	11.2	6	1mai	119	13.7	8	1cex	197	15.2
4	1tpn	45	11.0	6	1pne	139	13.8	8	1cto	109	15.1
4	1ubi	71	10.9	6	1rie	123	13.6	8	1kid	189	16.2
4	1uxd	59	11.5	6	1sfp	111	13.4	8	1knb	186	16.1
4	1vif	60	10.9	6	1tit	89	12.9	8	1np4	184	15.5
4	1vig	67	11.2	6	1tul	102	13.5	8	1pkp	145	15.1
4	2ech	49	11.1	6	1whi	122	13.6	8	1ra9	159	15.5
4	2hqi	72	10.7	6	1wiu	93	13.0	8	1rlw	126	15.3
4	2igd	57	10.7	6	2bb8	71	12.9	8	1sfe	165	15.7
4	2sn3	65	10.8	6	2mcm	112	13.4	8	1std	162	16.0
5	1aba	87	12.5	6	2phy	125	13.3	8	1vhr.A	178	15.5
5	1ag4	103	12.5	6	2pld.A	101	13.7	8	1xnb	185	15.2
5	1aoy	74	12.0	6	2tbd	128	13.3	8	1yua	122	15.2
5	1awd	94	11.7	6	3chy	128	13.3	8	2gdm	149	15.1
5	1awj	77	11.7	6	3nll	138	13.6	8	2pth	193	15.4
5	1bdo	80	11.9	7	153l	185	14.9	8	2rn2	155	15.3
5	1bxa	105	12.6	7	1ahk	129	14.8	8	2sak	121	15.4
5	1cyo	88	12.6	7	1ax3	156	14.3	9	119l	162	16.5
5	1mb1	98	12.3	7	1ayo.A	125	14.9	9	1asx	152	16.6
5	1mzm	86	11.9	7	1b10	104	13.9	9	1gky	186	16.4
5	1put	106	12.2	7	1bc4	110	14.5	9	1pbw.B	195	17.3
5	1spy	85	12.2	7	1be1	137	13.9	9	2ucz	164	16.5

[a]The training set listed was used to derive the size-dependent potential. Size bins are defined in terms of radius of gyration (R_g) rather than number of residues (N_{res}).

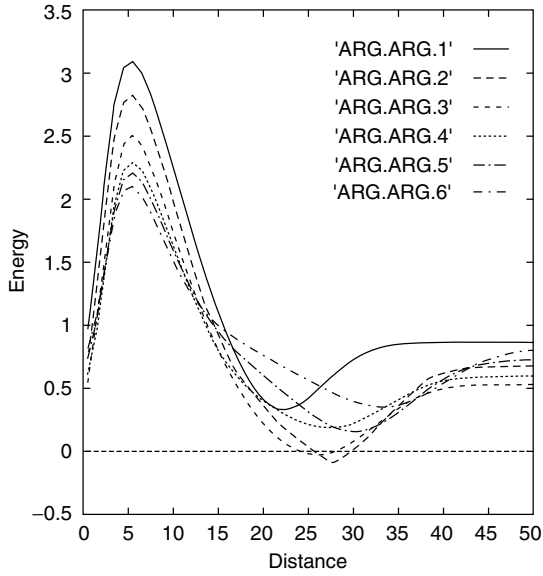


Figure 2. Size dependence of three representative terms in Eq. (4) for the amino acid pair arginine–arginine–arginine. Data for the first six size bins are shown.

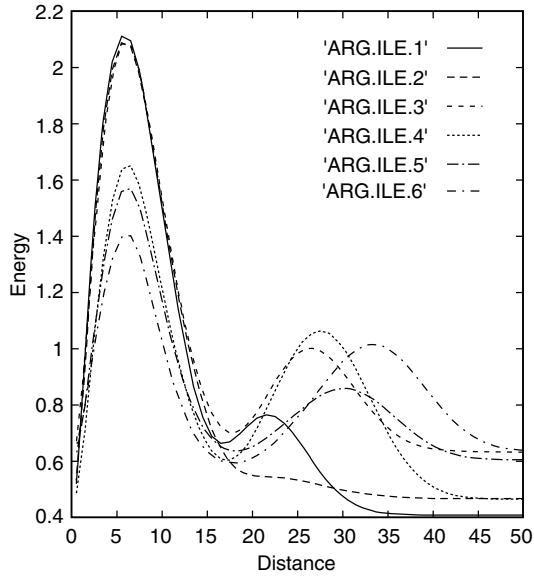


Figure 3. Size dependence of three representative terms in Eq. (4) for the amino acid pair arginine–isoleucine. Data for the first six size bins are shown.

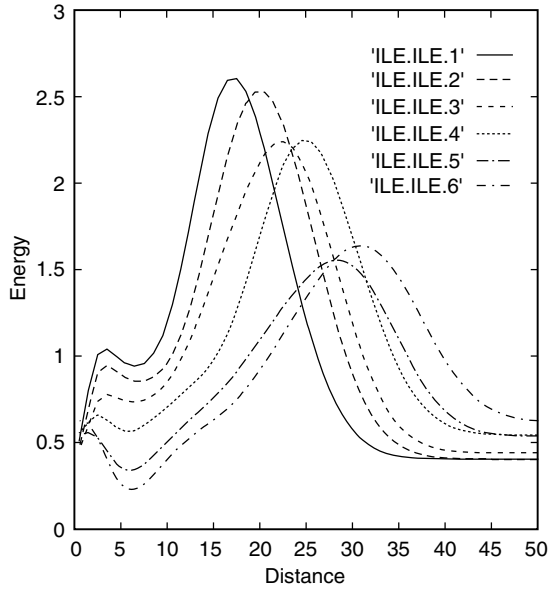


Figure 4. Size dependence of three representative terms in Eq. (4) for the amino acid pair isoleucine–isoleucine. Data for the first six size bins.

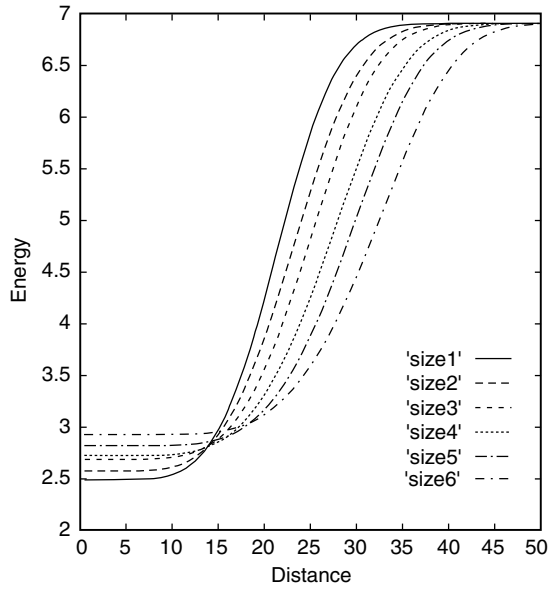


Figure 5. Density profiles for the first six size bins.

be easily rationalized: The arginine–arginine residues are pushed apart, while the isoleucine–isoleucine interaction is attractive. The arginine–isoleucine term is repulsive is well, but the minimum values occur at shorter distances than in the corresponding arginine–arginine plots, consistent with our intuitive picture of a spheroid with hydrophilic residues residing primarily on the surface. Not surprisingly, the basic effect of the density profile is to restrict the interresidue separation as a function of protein size. Note also that the density profile is the most sensitive to protein size (although the isoleucine–isoleucine pair potential clearly decreases with size).

Figure 6 illustrates the effect of adding the excluded volume and density profile to the arginine–arginine, arginine–isoleucine, and isoleucine-isoleucine potentials, respectively, for size bin 6. We see here that the linear portions of the potential are now restricted to a small range in distance (about 6–12 Å), outside of which the density profile and excluded volume become the dominant terms. The energies of each of the three residue pairs at large separation (e.g., 25 Å) relative to their minimum values increase in the expected order ($E_{\text{Ile-Ile}} > E_{\text{Arg-Ile}} > E_{\text{Arg-Arg}}$).

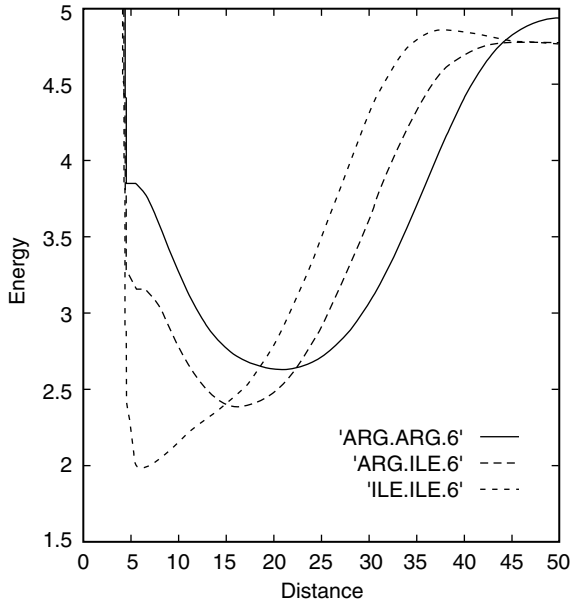


Figure 6. Total energy for three representative residue pairs: arginine–arginine, arginine–isoleucine, and isoleucine–isoleucine. The data corresponds to size bin 6.

III. TERTIARY FOLDING SIMULATIONS: PDB DERIVED AND IDEAL SECONDARY STRUCTURES

A. Physical Model

The physical model of the polypeptide chain we use has been described previously [2]; a few minor modifications are introduced as noted below. All bond angles and bond lengths are fixed at ideal values. The variables in the optimization are the torsional angles ϕ and ψ of the peptide backbone. Each residue is represented by a C_α atom and a C_β-like atom. The C_β atom position is given by the average projection of the side-chain center of mass onto the C_α–C_β bond vector.

We employ three different methods to describe the location and three-dimensional structure of secondary structure elements (i.e., α-helices and β-strands). The first is to take both the sequence location and backbone angles (which are frozen during the simulation) directly from the PDB entry. This is obviously not a realistic data set in a predictive situation, but is an essential computational experiment in that it indicates what level of accuracy is possible with ''perfect'' secondary structure information. The second is the replacement of PDB backbone angles with ideal backbone angles; this separates the effects of distortion of secondary structural elements from ideal geometries from errors in location in the sequence or in length. For these two types of calculations the correct size-dependent potential is selected by evaluating the radius of the gyration of the corresponding native structure. The third is to employ predicted, rather than PDB, secondary structure (along with the use of ideal geometries for the predicted elements) and to select the correct potential by predicting the radius of gyration from the number of residues of the target [22]. We have carried out an extensive investigation in this regard, using secondary structure prediction from various secondary structure prediction servers that are available over the Internet. These results are then combined to produce genuine *ab initio* structural prediction. The results, while far from a robust *ab initio* methodology over all protein types, yield important insights into the key obstacles to *ab initio* prediction and are in many cases surprisingly accurate. Predictions from the CASP3 contest are also included so that comparisons can be made with the work of others. While we are not generating these predictions as a blind test, it is the case that our CASP3 calculations were carried out using our software in a completely automated fashion, with no readjustment of parameters after obtaining results for the CASP3 targets.

B. Simulation Methodology

Our simulation methodology is identical to that presented in previous publications [2], so we will describe it only briefly here. The algorithm is based

on the Monte Carlo plus minimization (MCM) strategy proposed by Li and Scheraga [23]. This approach has proven to be extraordinarily efficacious in our previous work, and the present results reinforce our conclusions concerning its robustness and efficiency in enumerating the low-energy basins of attraction for low-resolution models such as those employed here. As in previous work [2], we have incorporated several key modifications of the algorithm, the most important of which is that the number of minimization steps is annealed as a function of the simulation temperature (i.e., more steps are taken later in the simulation), which yields a factor of 5–10 times reduction in computational effort. Finally, calculations are performed using a parallelized version of the code (an MPI implementation) on a network of PCs using Intel microprocessors and also on a large SGI Origin at the National Center for Supercomputing Applications.

The MCM procedure produces a large number of low-energy structures. The structurally unique predictions are extracted from the raw simulation data by a clustering algorithm. Figure 7 illustrates this process for the protein 1ACP. The raw simulation data (red dots) are combined into structurally similar clusters using a procedure discussed in Ref. 24. The criterion for separating structures into clusters is that the average RMSD between clusters (calculated over all structures in a particular cluster) be at least 5 Å. Clusters are represented by their lowest energy structure (black circles), which means that energies and RMSDs reported for clusters are based on their lowest-energy structure. The

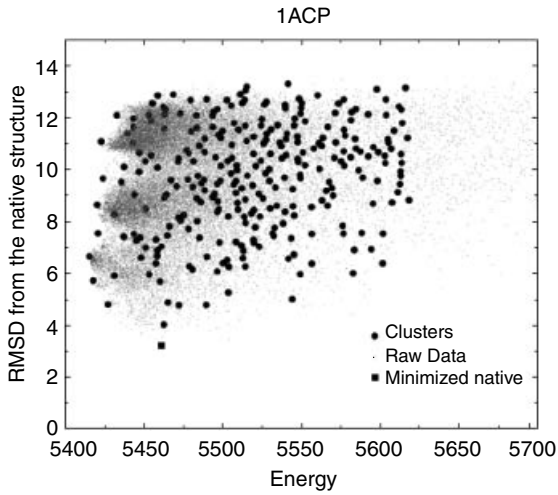


Figure 7. (See also color insert.) Comparison of raw data and clustered results (red dots: raw simulation data, black circles: cluster representatives, green square: locally minimized native structure).

TABLE III
RMSD with Respect to the Native of the ten Lowest Energy Clusters (Represented by Their Lowest Energy Member) for the Protein 1acp[a]

Cluster #	RMSD	Energy	*N*
1	6.65	5415.43	43
2	5.72	5417.83	58
3	8.64	5420.17	23
4	7.54	5420.80	25
5	11.08	5422.82	31
6	9.65	5424.04	34
7	4.79	5427.21	22
8	5.91	5431.10	16
9	8.27	5431.23	12
10	12.10	5432.68	19

[a]*N* gives the number of structures combined into a cluster.

RMSD between resulting representative structures is usually at least 5 Å, but this is not guaranteed by the clustering algorithm because we use the average RMSD as the clustering criterion. For the 10 representative structures lowest in energy we list energy, RMSD with respect to the native and number of structures combined into a cluster in Table III, and the RMSD between the representative structures themselves in Table IV. Derivation of the ranks of structures (discussed below) is straightforward given the data in Table III.

The lowest-energy structure obtained from the simulations is generally highly refined, meaning that its energy cannot be lowered significantly by performing more extensive searches. Refinement of higher-energy structures, structures that do not rank first, is possible though and in some of the cases,

TABLE IV
RMSD Between the Representative Structures from the Ten Lowest-Energy Clusters for the Protein 1acp

Cluster #	1	2	3	4	5	6	7	8	9	10
1	0.00	2.23	8.73	2.31	9.91	5.01	7.18	9.35	8.20	12.23
2		0.00	8.70	3.41	9.94	6.04	6.08	8.35	8.16	11.93
3			0.00	8.85	11.86	8.63	8.46	8.24	3.04	8.63
4				0.00	9.70	4.08	7.97	9.91	8.53	12.14
5					0.00	8.69	11.31	11.02	11.77	8.40
6						0.00	10.54	11.91	8.45	11.75
7							0.00	3.01	8.27	10.81
8								0.00	8.46	9.91
9									0.00	9.04
10										0.00

especially the larger proteins, actually results in improved ranks. We have not yet developed the optimal refinement strategy though and therefore do not report results for this approach.

C. Comparison of the Size-Dependent Potential with Previous Results Using PDB-Derived Secondary Structure

As a test set, we employed the subset of the 95 proteins used in Ref. 2 which are less than 100 residues and are not all β-strand. There is some overlap with the training set; but in tertiary folding, this is less of a concern than in secondary structure prediction because the three-dimensional phase space of the protein is so large that as long as an adequate number of proteins are used to generate the pair potential statistics, systematic bias of the results coming from the training set is unlikely to be large. In fact, we see little difference in performance for proteins depending upon whether they were included in the training set or not (or for the CASP3 targets we examined). By retaining the test set used in the previous chapter, we are able to directly compare our new potential with the older potential lacking size dependence, and thus assess the degree of progress that has been made by incorporating size dependence into the potential function.

As discussed above, after the tertiary folding simulations are completed, we group the resulting structures into clusters (without any reference to the native structure, which is presumed to be unknown during clustering) and report the highest-ranking clusters with RMSD from the native below 4 Å , 5 Å, 6 Å, and 7 Å, respectively.

In Table V, we compare these results for our test set with those obtained in Ref. 2. Note that Ref. 2 also included postsimulation screening algorithms; we have not developed such methods for the new potentials because some of the ideas have been incorporated directly into the energy function. Consequently we compare only with results taken directly from the simulations in Table V. However, we note that the overall quality of the results from the new potential is substantially better than those from the old, even when screening is employed in the latter. Table VI summarizes performance for various types of proteins and size classes.

The performance of the new potential function is particularly striking for proteins in the 50–100 residue size. For α-helical proteins in this category, the average rank of the best structure less than 7 Å is 3.6; furthermore, in the overwhelming majority of cases, the rank is 5 or better. This is a sufficient reduction in the number of possible structures that discrimination among the resulting structures via more expensive calculations at an atomic level of detail [25] becomes feasible. The reliability of the results demonstrates that the basic physics of the low-resolution model have been qualitatively improved as compared to previous efforts.

TABLE V
Comparison to Previous Results[a]

				"Old" Potential			Size-dependent Potential									
				PDB—X-RAY			PDB—X-RAY					PDB—IDEAL				
	N_{res}	N_α	N_β	<5 Å	<6 Å	<7 Å	<4 Å	<5 Å	<6 Å	<7 Å	LER	<4 Å	<5 Å	<6 Å	<7 Å	LER
Alpha Proteins ($N_{res} < 50$)																
1ajj	17	6	0	—	1	1	—	1	1	1	4.0	—	1	1	1	4.9
1bgk	27	18	0	4	2	1	2	2	2	1	6.5	2	2	2	1	6.2
1erd	29	25	0	1	1	1	1	1	1	1	3.8	1	1	1	1	3.3
2erl	35	29	0	2	2	2	—	1	1	1	4.9	1	1	1	1	2.8
1res	35	27	0	3	1	1	1	1	1	1	3.5	1	1	1	1	3.8
1roo	17	14	0	1	1	1	1	1	1	1	3.7	1	1	1	1	3.7
1uxd	43	31	0	1	1	1	4	4	4	1	6.0	—	4	4	1	6.4
Mixed Alpha/Beta Proteins ($N_{res} < 50$)																
1aho	31	10	10	5	3	1	7	5	2	1	6.8	3	3	2	2	7.5
1ayj	46	11	15	33	1	1	—	—	2	2	7.7	—	3	3	2	8.6
1cmr	26	8	10	3	1	1	3	2	2	1	6.6	4	4	3	1	6.8
1gpt	47	13	19	23	2	2	13	13	12	3	8.1	—	—	2	2	8.9
1hev	25	7	11	1	1	1	3	1	1	1	5.0	—	3	2	2	7.1
2ktx	34	11	14	1	1	1	1	1	1	1	3.6	—	1	1	1	4.2
1pce	30	12	10	2	2	2	1	1	1	1	2.8	—	—	1	1	5.1
1ptq	43	6	8	732	21	18	—	—	20	11	8.6	—	—	16	1	6.8
2sn3	48	8	15	94	21	7	—	29	2	2	8.5	—	13	3	3	8.9
2vgh	34	6	12	126	61	21	—	—	—	4	7.1	—	—	—	3	8.2
1vtx	36	7	10	—	78	2	—	—	34	3	7.8	—	—	9	1	7.0
5znf	25	12	11	1	1	1	1	1	1	1	2.6	—	—	1	1	6.0
Alpha Proteins ($50 N_{res} < 100$)																
1acp	73	45	0	256	115	30	—	7	2	1	6.7	—	—	11	11	11.3
1ail	67	60	0	5	5	2	1	1	1	1	3.0	1	1	1	1	3.9
1aj3	95	86	0	2	2	2	2	2	2	2	9.3	2	1	1	1	4.6
1am3	57	45	0	—	8	8	—	6	6	2	10.7	—	24	5	1	6.1
1c5a	62	49	0	1	1	1	—	3	3	2	8.2	10	3	3	3	8.0
1cc5	76	41	0	—	78	21	—	6	6	2	8.5	—	18	6	3	7.2
1ddf	87	66	0	—	7	7	—	63	3	2	12.7	—	58	8	8	7.1
2ezh	59	45	0	16	5	2	1	1	1	1	3.8	3	3	3	2	9.7
2ezk	76	64	0	28	8	1	—	—	1	1	5.7	—	—	1	1	5.9
2hp8	56	44	0	—	4	2	—	2	2	2	9.7	—	2	2	2	7.1
1hsn	62	46	0	88	88	67	—	—	19	19	11.4	—	—	98	17	8.3
1jvr	74	59	0	5	5	5	31	31	1	1	5.3	—	10	9	7	10.4
1Ifb	69	48	0	—	94	94	—	—	5	5	10.4	—	15	11	11	10.6
1mzm	71	54	0	—	8	8	—	5	4	4	10.7	—	3	2	2	11.0
1nkl	70	56	0	—	—	2	1	1	1	1	3.9	2	2	2	2	9.6
1nre	66	55	0	22	22	22	22	1	1	1	4.9	19	1	1	1	4.6
2pac	77	26	0	—	—	136	—	—	53	1	6.4	—	—	76	5	11.2
1pou	70	57	0	—	6	6	1	1	1	1	2.3	4	4	4	4	11.2
1r69	61	41	0	46	9	8	—	6	6	3	11.3	—	23	12	5	10.7

TABLE V (*Continued*)

				"Old" Potential			Size-dependent Potential									
				PDB—X-RAY			PDB—X-RAY					PDB—IDEAL				
	N_{res}	N_α	N_β	<5 Å	<6 Å	<7 Å	<4 Å	<5 Å	<6 Å	<7 Å	LER	<4 Å	<5 Å	<6 Å	<7 Å	LER
1utg	62	53	0	4	2	1	—	21	1	1	5.6	—	14	1	1	5.3
5icb	72	52	0	—	—	—	8	8	2	1	6.1	—	—	8	1	6.2
				Mixed Alpha/Beta Protein ($50 < N_{res} < 100$)												
1aa3	56	31	8	—	—	—	19	19	6	3	8.4	7	7	7	5	9.4
2acy	92	24	41	—	—	16	—	—	5	5	12.0	—	—	—	—	13.0
1ag2	97	58	8	—	—	349	—	—	—	87	10.9	—	—	—	187	12.3
1bor	52	9	14	187	22	8	—	—	17	6	7.2	—	—	40	12	8.3
1btb	89	45	19	—	274	24	1	1	1	1	3.8	—	—	31	28	8.1
1ctf	67	38	19	15	12	4	1	1	1	1	3.0	—	—	4	4	11.1
2fdn	53	8	6	123	4	4	—	—	38	6	8.1	—	—	—	30	10.3
2fow	66	29	8	181	56	8	—	—	23	8	10.6	—	—	69	4	7.9
1fwp	66	22	17	484	2	2	—	3	3	3	10.3	—	42	10	10	10.3
1gb1	54	13	16	1	1	1	—	—	15	1	6.5	—	—	2	1	6.5
1pgx	57	15	33	4	4	4	2	2	2	2	9.5	—	35	28	11	8.1
1leb	63	36	6	142	27	4	—	3	3	3	10.9	—	6	6	6	8.7
1orc	56	25	17	2	2	1	8	6	6	6	7.1	46	2	2	1	6.2
5pti	55	16	14	109	16	16	—	—	14	4	10.1	—	—	47	14	7.1
2ptl	60	15	34	1	1	1	1	1	1	1	3.4	—	35	4	4	8.2
1ris	92	25	42	—	180	11	9	9	9	9	11.1	—	—	129	11	11.7
1svq	90	22	34	—	—	—	—	119	117	32	12.5	—	—	462	43	9.0

[a]Following global energy minimization, structures are clustered without reference to the native; the energetic ranks of clusters that have an RMSD close to the native (for old results, three RMSD cutoffs—5 Å, 6 Å, and 7 Å—were used; for new results, four RMSD cutoffs—4 Å, 5 Å, 6 Å, and 7 Å—were used). Energetic rank was defined so that the lowest-energy structure ranks 1, the second-lowest ranks 2, and so on. LER refers to the RMSD of the lowest-energy structure. The column "PDB—X-Ray" list's results of runs using location and configuration of secondary structure derived from the PDB entry. Column "PDB-Ideal" lists results for calculations where the location of secondary structure was derived from the PDB, but configuration of secondary structural elements was assumed to be ideal.

For mixed α/β-proteins, the absolute quality of the results is somewhat diminished, but the improvement as compared to previous work is even larger. There are two cases, 1ag2 and 1svq, where the rank obtained for the best low RMSD structure is above 10, with the 1ag2 result being particularly problematic. We have investigated this case further and show improved results for 1ag2 below. On the other hand, there is a significant number of cases for which no reasonable structures were recovered previously which now rank in the top 10.

The energies of structures located by the global optimization algorithm are lower than the native and locally minimized native structures in all cases, a

TABLE VI
Summary of Ranks Listed in Table V[a]

(**a**)

	RMSD < 4 Å				RMSD < 5 Å			RMSD < 6 Å			RMSD < 7 Å		
Class	N_{prot}	N_{conv}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank
Small α	7	—	—	—	6	2	4	7	1	2	7	1	2
Small α/β	12	—	—	—	11	93	732	12	16	78	12	5	21
Medium α	21	—	—	—	11	43	256	18	26	115	20	21	136
Medium α/β	17	—	—	—	11	114	484	13	46	274	15	30	349

(**b**)

	RMSD < 4 Å				RMSD < 5 Å			RMSD < 6 Å			RMSD < 7 Å		
Class	N_{prot}	N_{conv}	Ave Rank	Max Rank	N_{prot}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank	N_{prot}	Ave Rank	Max Rank
Small α	7	5	2	4	7	2	4	7	2	4	7	1	1
Small α/β	12	7	4	13	8	7	29	11	7	34	12	3	11
Medium α	21	8	8	31	17	10	63	21	6	53	21	3	19
Medium α/β	17	7	6	19	10	16	119	16	16	117	17	10	87

(**c**)

	RMSD < 4 Å				RMSD < 5 Å			RMSD < 6 Å			RMSD < 7 Å		
Class	N_{prot}	N_{conv}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank	N_{conv}	Ave Rank	Max Rank
Small α	7	5	1	2	7	2	4	7	2	4	7	1	1
Small α/β	12	2	4	4	6	5	13	11	4	16	12	2	3
Medium α	21	7	6	19	16	11	58	21	13	98	21	4	17
Medium α/β	17	2	26	46	6	21	42	14	60	462	16	23	187

[a]Part a lists old results; part b lists results using the size-dependent potential and X-ray-derived secondary structure; part c lists results using the size-dependent potential and ideal secondary structure. The number of proteins N_{prot} is listed in column 2; the number of cases that converged within a specified RMSD from the native (< 4 Å, < 5 Å, < 6 Å, or < 7 Å) N_{conv} is listed in columns 3, 6, 9, and 12. (Note that the rank < 4 Å was not calculated for the old results, so a "—" is shown). Also listed are the average and maximum rank of converged clusters within each RMSD range.

feature that other groups using similar approaches have also observed [25]. A very important aspect of the results though, not apparent in the data presented here, is that for all simulations discussed above, the energy gap between the lowest-energy misfolded structures and low-energy native-like structures is quite small, on the order of 5–30 energy units where the energy scale is

TABLE VII
Comparison of Rankings for PDB Secondary Structure and DSSP Secondary Structure for Several Cases from the Test Set

Protein	<4 Å	<5 Å	<6 Å	<7 Å	Comments
1ag2	—	—	—	87	PDB secondary structure, terminal loops deleted
1ag2	—	—	—	11	DSSP secondary structure, terminal loops included
1hsn	—	—	19	19	PDB secondary structure, terminal loops deleted
1hsn	—	—	23	10	DSSP secondary structure, terminal loops deleted
1orc	8	6	6	6	PDB secondary structure, terminal loops deleted
1orc	1	1	1	1	DSSP secondary structure, terminal loops included
1ris	9	9	9	9	PDB secondary structure, terminal loops deleted
1ris	—	—	3	3	DSSP secondary structure, terminal loops included

typically thousands of energy units. This is in sharp contrast to the results obtained with our previous tertiary folding potential, which routinely generated energy gaps between misfolded and native-like structures that were 5–10 times larger than those seen here.

D. Effects of Secondary Structure Definition and Truncation of Terminal Loops

The results presented above employ PDB-defined secondary structure and in some cases involve truncation of terminal loops, primarily carried out here to facilitate direct comparisons with the results of Ref. 2. However, the process of defining secondary structure even with X-ray crystallographic or NMR coordinates in hand is not entirely unambiguous, and the effects of terminal loops could be favorable or unfavorable. To examine these issues, we selected several proteins in Table V for which the results with the new potential appeared less accurate than would have been expected given the difficulty of the case being considered. Table VII presents results for these selected cases, listing the protein and identifying what experiments were carried out. Most of the cases examined are mixed α/β because these displayed the most significant problems. It can be seen that in some cases the use of a different secondary structure definition (e.g., DSSP rather than PDB) and the inclusion or deletion of a terminal loop has a substantial effect on the ranking of low RMSD structures. Clearly, more work needs to be done in understanding these effects.

E. Effects of Using Ideal Rather than PDB-Derived Three-Dimensional Topologies for Secondary Structure Elements

Having established that our new size-dependent potential is quite effective for generating low-resolution structures of proteins below 100 residues using secondary structure derived from PDB coordinates, we next ask what the effect is of using ideal torsional angles for helices and strands as opposed to PDB-derived

torsion angles. Tables V and VIc summarize results for the entire test set of proteins utilizing ideal secondary structure elements. The results are surprisingly good; while there are certainly cases in which quantitative degradation of the rank of the best low-RMSD structure occurs (particularly with α/β-proteins—for example, the proteins 2fdn, 1fwp, and 5pti), in general the simulations are able to find such structures successfully and to rank them reasonably well in terms of total energy. Even in the case of 5pti, where there is severe distortion of the β-strands in the native structure, the use of ideal strands produces reasonable results. While incorporation of strand distortion is possible in our methodology [4], the reasonable predictive capability using ideal elements is likely to save considerable computational effort because one can carry out such simulations initially and then use the results as a starting point from which to incorporate distortions and other detailed effects.

IV. USE OF PREDICTED RATHER THAN PDB-DERIVED SECONDARY STRUCTURE ELEMENTS

A. Overview

Secondary structure prediction methods, while they have improved significantly over the past decade (principally via the use of multiple sequence analysis), still have nontrivial error rates. The best method at present appears to be the PSIPRED approach developed by Jones [26], which is claimed to achieve an accuracy between 76% and 78% on a reasonably large training set (it also outperformed other methods in the CASP3 contest). This level of reliability appears to be sufficient for low-resolution *ab initio* structure prediction and suggested to us that it was now worth experimenting with tertiary folding calculations based entirely on predicted, rather than PDB-derived, secondary structure [27–32]. Using servers set up on the World Wide Web, we are able to obtain predictions from PSIPRED and other secondary structure prediction algorithms for proteins in our test set. We have obtained results from a variety of servers to see what happens in cases where their predictions disagree; it is likely that *ab initio* prediction will involve trying a number of secondary structures, because in some cases the tertiary fold will be critical in selecting among plausible secondary structures predicted exclusively from sequence data.

Our calculations in this section endeavor to answer the following questions:

1. Can we for some percentage of cases make a successful *ab initio* prediction? We explore two different approaches below.
2. What are the effects of small errors in secondary structure—for example, elimination or addition of small elements, incorrect lengths of major elements and so on?

3. What is the impact of a major error—for example, replacing a long helix by a similar strand or missing an important loop?

In the present chapter, we have chosen to focus our *ab initio* prediction efforts primarily on α-helical proteins, although one mixed α/β-protein is also examined. The *ab initio* prediction calculations presented below are considerably more computationally intensive than those using PDB-derived secondary structure, because we have investigated a substantial number of secondary structure predictions for each protein. By studying helical systems intensively, we are able to draw conclusions concerning the necessary and sufficient conditions for success for such systems from a significant database of results. In addition to the α-helical proteins in the 50- to 100-residue range from the data set above, we also include two helical proteins from the CASP3 prediction contest. Our results for the CASP3 test cases are similar to those from the PDB-derived test suite.

B. Secondary Structure Prediction Methods

We use the following secondary structure prediction methods in our *ab initio* predictions:

- PSIPRED [26]: A two-stage neural network that predicts protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST (available at *http://insulin.brunel.ac.uk/psipred/*). Average three-state prediction accuracy is between 76.5% and 78.3%. Currently the most accurate method.
- PhD [33,34]: Secondary structure is predicted by a system of neural networks (available at *http://cubic.bioc.columbia.edu/pp/*). Overall three-state prediction accuracy is 72.1%. The default secondary structure prediction settings were used in all predictions.
- JPRED [35,36]: A methodology that combines a total of six secondary structure predictions into one consensus prediction (available at http://jura.ebi.ac.uk:8888/ at the time of this writing). Average ''real world'' accuracy is 72.9%. Note that the PhD predictions generated by JPRED differ from the original PhD predictions (denoted: orig_phd) mentioned above. In addition to using the consensus prediction, we also report results for the six individual prediction methods included in the JPRED server.

By default, secondary structure prediction accuracies reported here are determined with DSSP as the reference (for details see Ref. 26). The secondary structure assignments used in the actual calculation differ from the original predictions in that helices and strands of less than three residues are eliminated. N- and C-terminal loops are deleted.

C. Simulation Protocols

The amino acid sequence of the target represents the only input data for our methodology. We do not carry out explicit database searches (i.e., threading) of any sort. Secondary structure predictions from the sources listed above are parsed and used directly in the structure predictions. In the case of JPRED we examine individually the results of all predictions that contribute to the consensus prediction (DSC [37], PhD [33,34], PREDATOR [38,39], NNSSP [40], Mulpred, and Zpred [41]). Because we do not assume any knowledge of approximate radius of gyration of the target, which is important for the selection of the correct potential energy parameters, we predict the radius of gyration via a simple formula [22] and use this prediction to assign the size bin for the tertiary folding simulation.

The first stage of our prediction algorithm applies the MCM-based approach described above to each of the nine secondary structure predictions for each target. Simulations are usually carried out on two to four nodes of a multiprocessor machine and take between 12 and 24 hours depending on protein size. To extract the structurally unique predictions, we apply the clustering algorithm discussed above. Table VIII shows the results of this procedure for the three targets discussed in more detail below. We list results for every secondary structure prediction (unless predictions consist only of loop or coil, in which case we did not believe it worthwhile to carry out the simulation).

Because it is quite possible that simulations utilizing different secondary structure predictions results in very similar representative low energy structures, we apply a second level of filtering which basically tries to eliminate structurally similar predictions and ranks the resulting "unique" predictions on a absolute energy scale. The first step in this process is the determination of the subset of residues common to all predictions (regardless of whether they belong to helices or strands). Secondary structure predictions for which the number of residues included in the simulation is substantially smaller than the average (due

TABLE VIII
Individual Clustering Results for the *ab initio* prediction Targets Discussed in More Detail in the Text (Stage 1 of the Composite Prediction Method)[a]

Protein	SSP	Q3	N_{res}	<4 Å	<5 Å	<6 Å	<7 Å
1aj3	cons	94.90	93	—	—	—	—
1aj3	dsc	87.76	93	—	—	—	—
1aj3	mul	86.73	92	—	—	—	—
1aj3	nnssp	95.92	98	—	—	2	2
1aj3	orig_phd	89.80	89	—	—	3	2
1aj3	phd	88.78	88	—	—	—	—
1aj3	pred	88.78	92	—	—	—	—

TABLE VIII (*Continued*)

Protein	SSP	Q3	N_{res}	<4 Å	<5 Å	<6 Å	<7 Å
1aj3	psipred	93.88	94	—	—	1	1
1aj3	zpred	93.88	98	—	—	2	2
1am3	cons	92.86	58	—	—	24	3
1am3	dsc	88.57	57	—	—	11	2
1am3	mul	77.14	59	—	—	11	11
1am3	nnssp	88.57	60	—	8	8	8
1am3	orig_phd	92.86	58	—	22	8	6
1am3	phd	94.29	58	—	4	4	2
1am3	pred	72.86	57	—	—	—	—
1am3	psipred	88.57	57	—	35	1	1
1am3	zpred	67.14	68	—	—	—	—
1mzm	cons	44.09	44	—	—	2	2
1mzm	dsc	66.67	67	—	—	26	3
1mzm	mul	37.63	68	—	—	243	4
1mzm	nnssp	38.71	93	—	—	—	—
1mzm	orig_phd	59.14	74	—	—	50	3
1mzm	phd	38.71	49	—	—	9	1
1mzm	pred	55.91	44	—	—	18	6
1mzm	psipred	78.49	78	—	1	1	1
1mzm	zpred	34.41	82	—	—	—	—
1eh2	cons	87.37	68	—	12	5	5
1eh2	dsc	80.01	73	—	—	—	—
1eh2	mul	74.74	43	—	3	3	3
1eh2	nnssp	86.32	67	—	6	3	1
1eh2	orig_phd	86.32	68	—	15	1	1
1eh2	phd	85.26	67	15	4	4	4
1eh2	pred	88.42	68	—	1	1	1
1eh2	psipred	95.79	72	—	3	2	1
1eh2	zpred	66.32	72	—	—	32	13
1bg8.A	cons	57.89	57	—	—	11	3
1bg8.A	dsc	42.11	55	—	—	64	6
1bg8.A	mul	51.32	67	—	—	—	24
1bg8.A	nnssp	63.16	67	—	—	31	9
1bg8.A	orig_phd	57.89	57	—	—	52	5
1bg8.A	phd	57.89	57	—	—	34	8
1bg8.A	pred	38.16	56	—	—	321	3
1bg8.A	psipred	50.01	52	—	92	17	5
1bg8.A	zpred	46.05	68	—	—	—	—

[a]Here N_{res} refers to the number of residues actually considered for every prediction. (cons: JPRED consensus prediction; dsc: DSC; mul: MULPRED; nnssp: NNSSP; orig_phd: PhD in its most current implementation; phd: PhD as run by JPRED; pred: PREDATOR; psipred: PSIPRED; zpred: ZPRED). Q3 refers to the three-state accuracy of a given prediction.

to deletion of terminal loops) are not considered at this stage. This set of residues is then extracted from the 50 clusters lowest in energy for every secondary structure prediction, and the energies of the resulting substructures are evaluated. After a second round of clustering, we obtain the final set of clusters (Table IX). At this point the RMSDs with respect to the native structures are reevaluated over the subset of common residues to allow a fair comparison of the tertiary folding results obtained from different secondary structure predictions. We refer to this method below as the composite energy prediction method.

D. Final Rankings of Structures for Fully *Ab Initio* Predictions

We examine the use of two different approaches for producing fully *ab initio* predictions for the 22 proteins studied in this section. One approach is simply to use the secondary structure prediction with the highest calibrated prediction—accuracy—in this case, PSIPRED. Results for this approach are summarized in

TABLE IX
Final Clustering Results for the Subset of Common Residues for All *Ab Initio* Prediction Targets (Stage 2 of the Composite Energy Prediction Method)[a]

Protein	N_{res}	<4 Å	<5 Å	<6 Å	<7 Å
1acp	70	—	—	10	5
1aj3	88	—	89	89	89
1am3	56	—	17	17	2
1bg8.A	52	—	—	92	1
1c5a	57	—	—	4	4
1cc5	68	—	22	12	2
1ddf	85	—	—	—	7
1eh2	65	—	4	4	3
1hsn	61	—	—	—	46
1jvr	66	39	12	12	2
1lfb	55	—	114	22	4
1mzm	66	—	—	65	4
1nkl	63	—	—	31	1
1nre	65	—	89	50	50
1pgx	53	—	—	—	35
1pou	64	30	3	3	1
1r69	57	—	5	5	1
1utg	56	—	—	4	3
2ezh	57	—	—	7	4
2ezk	67	—	—	—	—
2hp8	49	—	58	8	7
2pac	53	—	25	7	3

[a]Here N_{res} refers to the number of residues for which RMSD and energy are evaluated. We omitted predictions that were too short as compared to all others and the length of the sequence (1eh2: mul; 1jvr: psipred; 1mzm: cons, phd, pred; 1r69: mul, orig_phd, pred; 2ezh: orig_phd).

TABLE X
Individual Clustering Results for All *Ab Initio* Prediction Targets Using the PSIPRED Secondary Structure Predictions[a]

Protein	Q3	N_{res}	<4 Å	<5 Å	<6 Å	<7 Å
1acp	83.12	72	—	—	17	5
1aj3	93.88	94	—	—	1	1
1am3	88.57	57	—	35	1	1
1bg8.A	50.01	52	—	92	17	5
1c5a	93.94	63	4	1	1	1
1cc5	74.7	75	—	6	4	4
1ddf	81.1	86	—	—	43	9
1eh2	95.79	72	—	3	2	1
1hsn	87.34	62	—	—	—	20
1jvr	72.26	3	—	—	—	—
1lfb	58.97	59	—	—	—	34
1mzm	78.49	78	—	1	1	1
1nkl	94.87	71	—	—	4	1
1nre	60.49	65	—	—	—	380
1pgx	77.14	60	—	—	—	—
1pou	73.24	67	7	5	5	5
1r69	84.13	59	—	6	6	3
1utg	85.71	62	—	32	1	1
2ezh	81.54	57	—	15	1	1
2ezk	51.61	77	—	—	—	—
2hp8	64.71	53	—	6	2	1
2pac	70.73	77	—	19	2	2

[a]Here N_{res} refers to the number of residues actually considered for every prediction.

Table X. As above we list the rank of structures below a certain RMSD cutoff. The second is the composite energy prediction method discussed above. We summarize statistics for the success rate of each of these two approaches on the entire test set and CASP3 prediction targets below.

E. Results

1. *Summary and Overall Success of Fully Ab Initio Prediction*

We begin by summarizing the results for all of the secondary structure prediction methods (including the composite energy prediction method described above) and all of the target proteins. As in previous sections of this chapter, the ranks of the lowest-energy cluster with RMSDs from the native structure of 4 Å, 5 Å, 6 Å, and 7 Å are reported for both approaches. The first, and most striking, observation is that both approaches provide a surprisingly good success rate for *ab initio* prediction based on criteria used in CASP3. We have observed that for proteins in the 50–100 residue range, an RMSD below 7 Å typically provides a

qualitatively reasonable folding topology at low resolution. Similar conclusions have been reached by Skolnick and co-workers [42] and by Cohen and Sternberg [43], whose estimates show that the probability of achieving a structure below 6 Å RMSD by chance is vanishingly small. Note also that for a significant fraction of proteins, structures below 6 Å are found; at this level, the correspondence with the native structure is quite satisfactory in agreement with the chapters cited above.

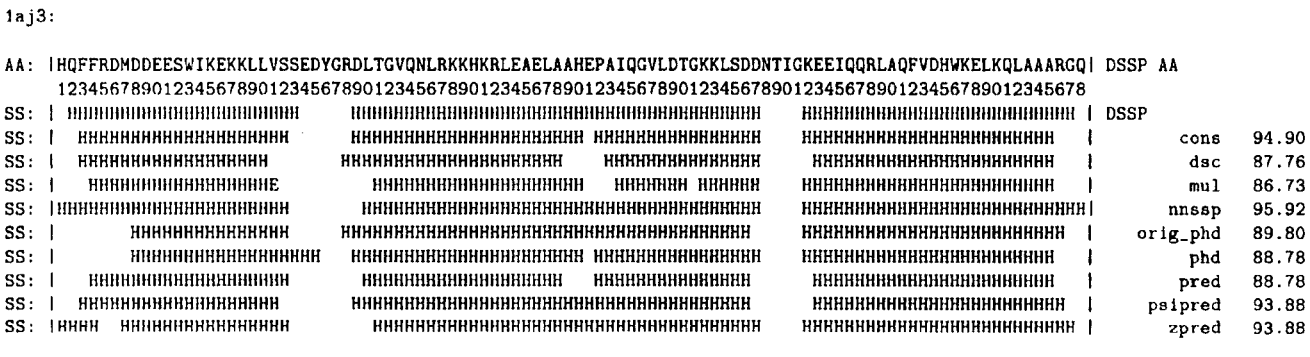
Both proposed fully *ab initio* prediction methods (composite energy method and exclusive use of PSIPRED predictions) yield a number of cases in which a low RMSD structure is ranked first; this would count as a successful prediction under any criterion. Using the assessment criteria of CASP3—that is, a maximum of five predictions—the composite energy method would achieve an RMSD of less than 7 Å in 68% of the cases; there are also four cases where the RMSD is less than 6 Å. Reliance entirely on PSIPRED would lead to an RMSD under 7 Å in 64% of the cases; however, 11 of those would have an RMSD under 6 Å. Thus, the use of the composite energy method appears to succeed slightly more often, however, the use of PSIPRED exclusively generates highly accurate predictions in significantly more cases.

We have employed the protocol described above in a completely automated fashion; but only in an actual blind test can one be sure that the results suffer from no unconscious bias. If these results hold up under truly blind test conditions, this would represent a significant advance in *ab initio* prediction methodology as judged by other *ab initio* efforts in CASP3.

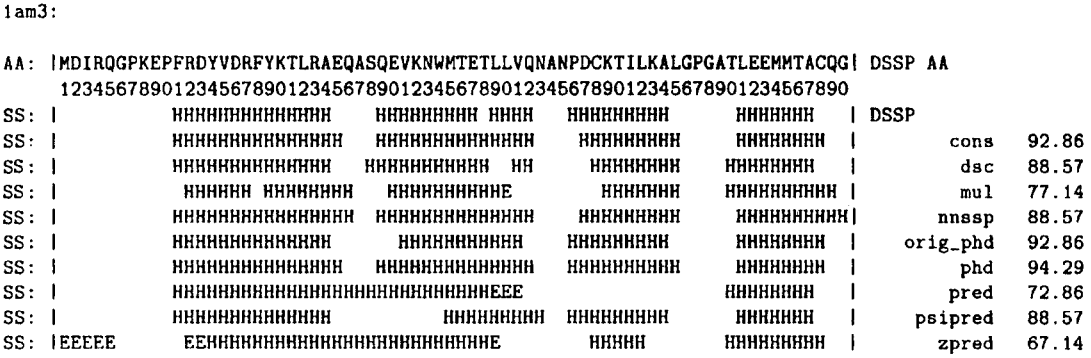
While our new potential energy function certainly represents a step forward, there are also obviously areas where more work needs to be done. Primarily, the causes of failure to routinely achieve a low-RMSD structure in the top five predictions in some cases must be analyzed and understood. These failures are thus more interesting at this point than the successes because they point the way to development of an improved methodology. We therefore analyze a number of these cases in detail below so as to reveal the underlying difficulties and directions in which solutions must be developed.

2. *Detailed Analysis of Specific Cases*

Figure 8 presents the detailed secondary structure predictions for each of the cases that we analyze below. In conjunction with the tertiary folding results summarized in Table VIII, as well as the results using PDB-derived secondary structure presented above, we can extract insight into how various types of errors in secondary structure prediction affect tertiary folding accuracy. Due to the large amount of data, we have selected a subset of interesting examples to analyze in detail, however, the conclusions, summarized in the discussion following consideration of individual examples, reflect an examination of the results for all 22 of the proteins studied.



(a)



(b)

Figure 8(a–e). Secondary structure predictions for 1aj3, 1am3, 1mzm, 1eh2, 1bg8 Chain A (cons: JPRED consensus prediction; dsc: DSC; mul: MULPRED; nnssp: NNSSP; orig_phd: PhD in its most current implementation; phd: PhD as run by JPRED; pred: PREDATOR; psipred: PSIPRED; zpred: ZPRED). References for the secondary prediction algorithms are given in the text.

```
1mzm:

AA: |AISCGQVASAIAPCISYARGQGSGPSAGCCSGVRSLNNAARTTADRRAACNCLKNAAAGVSGLNAGNAASIPSKCGVSIPYTISTSTDCSRVN| DSSP AA
     123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123
SS: |   HHHHHHHHHHHHHHHH       HHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHH     HHHHHHHHHHH             HHHH | DSSP
SS: |                                   HHHHH      HHHHHH                       EEEE              |      cons   44.09
SS: |            HHHHHH           HHHHHHHHHHH   HHHHHHHHHHHHHHH                 EEEE              |       dsc   66.67
SS: |             EEEEE          EEE E                EEEHHHH                    EEEEE E          |       mul   37.63
SS: |                                                                                             |     nnssp   38.71
SS: |          HHHHHHHH         HHHHHHHHHHHHH      HHHHHHHHHHHHHHH              EEEEEEEEE         |  orig_phd   59.14
SS: |     EE                            HHHHH      HHHHHH                      EEEEE EEEE         |       phd   38.71
SS: |              EEE                 HHHHHHH    HHHHHHHHHHHHH                                   |      pred   55.91
SS: |     HHHHHHHHHHHHH        HHHHHHHHHHH       HHHHHHHHHHHH        HHHHHHH      EEEEEE          |   psipred   78.49
SS: |  EEEEEEEEEEEEEEEEE         EEEEEEEEE        HHHHHEEHHHHEE        EEEEEE  EEEEEEEEEE       E |     zpred   34.41
```

(c)

```
1eh2:

AA: |PWAVKPEDKAKYDAIFDSLSPVNGFLSGDKVKPVLLNSKLPVDILGRVWELSDIDHDGMLDRDEFAVAMFLVYCALEKEPVPMSLPPALVPPSKR| DSSP AA

     12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345
SS: |     HHHHHHHHHHH          EHHHHHHHHH     HHHHHHHHHHH      E  HHHHHHHHHHHHHHHH                                   | DSSP
SS: |   E    HHHHHHHHHH          HHHHHHHHH    HHHHHHHHHHHH        HHHHHHHHHHHHHHH                                    |      cons   87.37
SS: |   HHHHHHHHHHHHHHHH         HHHHHHHHHH   HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH                                    |       dsc   80.00
SS: |            HHEE               HHEEEH      HHHHHHHH                 HHHHHHHHHHHHHHHH                           |       mul   74.74
SS: |        HHHHHHHHHH           HHHHHHH       HHHHHHHHH        HHHHHHHHHHHHHH                                     |     nnssp   86.32
SS: |  E     HHHHHHHHHH      EE   HHHHHHH     HHHHHHHHHHH        HHHHHHHHHHHHHHH                                    |  orig_phd   86.32
SS: |  EE     HHHHHHHHH          HHHHHHHHH    HHHHHHHHHHHH       HHHHHHHHHHHHHHH                                    |       phd   85.26
SS: |        HHHHHHHH             HHHHHHHH     HHHHHHHHH         HHHHHHHHHHHHHHH                                    |      pred   88.42
SS: |     HHHHHHHHHHHH          HHHHHHHHHH    HHHHHHHHHHH        HHHHHHHHHHHHHHHH                                   |   psipred   95.79
SS: |  HHHHHHH    EEEE       EE   HHHEEEE      EEEEHHHHEH       HHHHHHHHHHHHHHH                                     |     zpred   66.32
```

(d)

Figure 8(a–e) *(Continued)*

```
1bg8 - Chain A:

AA: |KKPVNSWTCEDFLAVDESFQPTAVGFAEALNNKDKPEDAVLDVQGIATVTPAIVQACTQDKQANFKDKVKGEWDKI| DSSP AA

     12345678901234567890123456789012345678901234567890123456789012345678901234567890123456
SS: |   HHH EHHHHH      HHHHHHHHHHH           HHHHHHHHHHHHHHHH     EHHHHHHHHH  | DSSP
SS: |       HHHHH         HHHHHHHH         EE         HHHHHHHHHHHHHHH           |       cons   57.89
SS: |       HHHHH         HHHHHHHHH       HHEEE              HHHHH              |        dsc   42.11
SS: |        HHHHHHH         HHHH H      HH    H EEEE EEEEEEEH  H HHHH HHHHHHHH |        mul   51.32
SS: |        HHHHHHHH     HHHHHHHH         EE         HHHHHHHHH  HHHHHHHHHHHHHHH|      nnssp   63.16
SS: |       HHHHH         HHHHHHHH         EE   E EE  HHHHHHHHHHHHHHH           |    orig_phd  57.89
SS: |       HHHHH         HHHHHHHH         EE   E EE  HHHHHHHHHHHHHHH           |        phd   57.89
SS: |          EEE         HHHHHHHH       EEEE         EEEEE HHHHHHHHHH         |       pred   38.16
SS: |          HHHH      HHHHHHHHH        EEEEE   EEEHHHHHHHHHHHHHH            |     psipred  50.00
SS: |       EHHHHHHHHHH   HHHHHHHH    HHHHHHHHHEEEEEEEEEEEEHHHHHHHHHHHHHHHHHH|      zpred   46.05
```

(e)

Figure 8(a–e) (*Continued*)

1aj3: This is a case for which the average three-state prediction accuracy of all of the secondary structure prediction methods is quite good, typically in excess of 85%. However, only four of the secondary structure predictions yield reasonable tertiary folding results (NNSSP, original PhD, PSIPRED, and ZPRED). The reason in this case is quite obvious; the successful methods correctly predict that the region between residues 29 and 67 is a single long helix, whereas the remaining predictions insert a short loop in the middle of this part of the sequence. The short loop allows the two helical pieces surrounding it to fold, producing a very different shape than is enforced by the single long helix.

As we shall see below, in many cases the composite energy scoring method is capable of selecting the better tertiary architecture where there are qualitative differences between predictions. In the present example, however, the simple algorithm that we use to combine the predictions does not work well, for a completely understandable reason. By introducing a loop into the long helix, the protein is given greater flexibility. Because we have not explicitly included any sort of scoring function for secondary structure [44], the only discriminant is the energy of the tertiary fold, which in this case must favor the more flexible structure. In the present system, the non-native structures have energies far below the native-like and native structures.

The problem observed here will be potentially significant whenever the correct secondary structure is a long helix, and prediction methods have trouble distinguishing this from a pair of helices with a short loop in the middle—a very common motif in secondary structure prediction codes. In order to rectify this problem, it will probably be necessary to combine local energies, which determine secondary structures, with long-range energy terms. One approach is to replace fixed secondary structures by torsion angle energy wells, the depth and breadth of which are functions of the secondary structure prediction confidences. It may be possible to optimize the balance of torsion and long-range energy parameters such that correct helix assignments are favored. An alternative approach is to use an atomic level potential function and continuum solvation model to compare the energies of the predictions with different secondary structure assignments. We intend to explore both of these strategies is future work.

1am3: This example contains the other side of the long helix problem observed in 1aj3. Again, all of the secondary structure prediction three-state percentages are reasonable. However, three of the methods (PRED, and Zpred) predict a single long helix between residues 11 and 42, whereas the DSSP-derived secondary structure (and the remaining predictions) specify two short helices. In this case, the methods that incorrectly predict the long helix are unable to obtain reasonable RMSD structures from the native structure. However, here the composite prediction method easily eliminates the qualitatively

incorrect predictions, in this case benefiting from the lower energies obtained due to greater flexibility of the two helical segments as opposed to a single long helix. Also of interest here is the result obtained from the Mulpred prediction, which inserts an incorrect short loop splitting the single helix between residues 12 and 26 into two shorter helices. This leads to a degradation in the rank of the best native-like structure, but does not eliminate the possibility of obtaining a reasonable prediction. Presumably, the magnitude of the effects of this sort of insertion are qualitatively larger when the size of the helix in question is large compared to the radius of gyration of the protein (as is the case in the two instances discussed here). It is also interesting that this error does not qualitatively degrade the results of the composite prediction method; it may be that structures with a significant bend at the short loop are energetically disfavored in this specific case.

1mzm: This protein is a startling example indicating that in some cases the tertiary folding potential can survive very large qualitative errors in secondary structure prediction. The only prediction that is satisfactory in terms of predicting major elements correctly is that of PSIPRED (and even here, a β-strand is incorrectly added on at the end), and indeed the PSIPRED results are certainly the best, particularly in terms of the RMSD of the low-energy structure which is below 5 Å. However, numerous other predictions are capable of achieving reasonable results, despite gross errors in the secondary structure of many different types. We have not analyzed in detail why this is the case; an initial speculation would be that this protein does not have a large number of alternative approaches to forming a good hydrophobic core. Also, because the potential energy function does not include explicit β-strand pairing terms, incorrect prediction of a strand is a local effect.

3. Summary of the Results for All Proteins

The following is a brief analysis of how the various types of errors identified in the secondary structure predictions affected the proteins in the test set:

1. *Incorrect Prediction of Long Helices.* This problem, which amounts to missing a critical loop, affected at least some predictions in most of the proteins studied. Fortunately, in most cases at least one of the secondary structure prediction methods correctly identified the loop in question. Because the composite energy ranking protocol favors flexibility over long helices, the presence of several incorrectly predicted long helices was not, in general, a fatal error.
2. *Incorrect Replacement of a Helix by a Strand.* This problem most significantly affected the proteins 1jvr and 1lfb. In some cases, good low-energy tertiary folds are obtained despite the replacement of a helix with a

strand; in other cases, the replacement eliminates any good predictions. More work is needed to determine under what conditions this type of error can be overcome, and when it is fatal.

3. *Incorrect Replacement of an Important Helix by Loop*. Given our current composite energy ranking scheme, which favors flexibility, this error is in general fatal. Fortunately, in all but one case (1nre) at least one (and usually more than one) secondary structure prediction method correctly identified the important helix. As discussed above, a composite energy that combines local and long-range energy terms appears necessary in order to treat long helices. In the short term, simply preventing one secondary structure assignment from dominating the composite ranking may sufficiently diversify the resulting low-energy structures.
4. *Small Errors in the Prediction (Incorrect Lengths of Secondary Structure, Small Helix, or Strand Incorrectly Present or Missing)*. Generally, these types of errors led to quantitative degradation in the ranking of low-RMSD structures as opposed to complete elimination of these structures.

4. Results from the CASP3 Prediction Contest

In addition to the test cases discussed above, we have also studied two small helical proteins that were targets in the CASP3 prediction contest. These studies allow us to compare our results with those of other groups [11]. The two targets we have investigated are target T0061 (PDB-code: 1bg8) and target T0074 (PDB code: 1eh2). Each is a helical protein between 50 and 100 residues and hence is part of the same general category as most of the proteins in the test set. The results for these two proteins are presented in detail in Tables VIII and IX and discussed below. We make explicit comparisons with the results of the Scheraga [25,45] and Samudrala [29] groups, both of whom carried out *ab initio* folding on these targets and used methods similar in spirit to what we present here. Those of the latter group are in fact quite analogous, because prediction methods are used to determine secondary structure, followed by tertiary folding simulations to generate a three-dimensional topology.

It should be noted that a nontrivial aspect of making these comparisons is that the proteins were truncated differently in the various calculations; we present all of the relevant information below so that the reader can draw his or her own conclusion. We do, however, wish to make one point with regard to the manner in which the comparison sequence is truncated. In our approach, truncation of terminal loop regions is done automatically using the secondary structure prediction, without reference to the native structure. In several of the comparisons we report below, truncation was carried out with the native structure in hand, presumably to minimize the RMSD obtained. While such results do indicate partial success of the folding algorithm, from a statistical

point of view it is much easier to achieve an RMSD of 6–7 Å when an extensive choice of fragments are available to be optimized as opposed to when a single fragment is chosen *a priori*. This is particularly the case when the fragment is relatively small compared to the total length of the sequence.

1eh2: The secondary structure prediction methods generally performed well on this protein. The tertiary folding simulations were also quite effective, with the best results yielding an RMSD of less than 5 Å as the lowest energy prediction. The composite energy method provides a prediction ranking 3 with an RMSD of 6.02 Å, a respectable result for a protein in the 50 to 100-residue range. If the PSIPRED secondary structure method were to be used exclusively, the best prediction among five submitted predictions would be 4.84 Å; this is an excellent result, competitive with the best results obtained from threading methods [46]. We note that in both predictions, a long terminal loop of the protein was truncated, so that the total number of residues predicted was 72 in the PSIPRED simulation and 65 in the composite energy method.

In CASP3, results for 1eh2 varied greatly with prediction method. Several groups were able to identify a remote homolog and hence utilize threading approaches to structure prediction [46], whereas others use methods based more on *ab initio* approaches. When only ~80% of the protein structure was predicted, the best results were in the 5 Å RMSD range; as the percentage of the protein predicted increased to 100%, the prediction accuracy degraded to 6.01 Å. Our results using PSIPRED secondary structure are comparable to the former results; in this case 74% of the residues were predicted to an accuracy of 4.84 Å.

The Scheraga group submitted a prediction for this target; however, they included the long terminal loop in their prediction which it is extremely difficult to predict correctly with *ab initio* methods. Consequently, their reported RMSD of 9.99 Å for the entire protein does not constitute a fair evaluation of the capabilities of their methodology. They also report a 5.8 Å RMSD for a 53-residue fragment of the protein. The calculations would most likely have been more successful had the terminal loop been deleted during the simulation, as was done in our approach. The Samudrala group, who achieved an RMSD of 11.3 Å, also included the terminal loop in their calculations. Their post-CASP3 analysis yielded an optimal fragment prediction of 7.0 Å for a 60-residue fragment. The results reported above (4.84 Å RMSD for 72 residues predicted) is qualitatively superior to either of these results, particularly as the truncation was carried out prior to the simulation.

1bg8—Chain A: 1bg8 is a target for which none of the predictors successfully located a remote homolog. The best results (and indeed the only ones that could be considered even partially successful for a protein this size) were those

of Scheraga and co-workers, who achieved an RMSD of 7.27 Å (for all 76 residues reported experimentally) as the best result of four submitted predictions (their remaining predictions had RMSDs of 8.91 Å, 9.08 Å, and 9.23 Å). Their best results for a postprocessed fragment are 4.2 Å for a 61-residue fragment. Using the composite energy method, our lowest energy prediction achieves an RMSD of 6.69 Å, but for only 52 residues obtained after truncating to allow energetic comparisons among all of the secondary structure predictions. The PSIPRED calculations yield a 6.07 Å RMSD, again for 52 residues (PSIPRED incorrectly predicts a long terminal loop, which we truncate). These results are respectable in terms of RMSD but involve significant truncation in a region where there is actual secondary structure.

The Samudrala group achieved an RMSD of 10.1 Å for all 76 residues and 7.4 Å for 66 residues after postprocessing. The Scheraga group results in this case have to be considered best. Much of their success can be attributed to an impressive 79% accuracy in the secondary structure assembled in their most successful simulation; in this case, the standard neural-network-based secondary structure prediction methods that we (and Samudrala and co-workers) employed have a much poorer performance than they do for the test set, with accuracies below 65% in all cases.

V. CONCLUSION

We have demonstrated that the inclusion of size dependence in the derivation of a statistical potential for tertiary protein folding yields substantially improved results, as compared to previous efforts, for a substantial number of proteins of less than 100 residues in size. The new potential reliably yields highly ranked structures with low RMSDs as compared to the native structure (in contrast to earlier results that displayed occasional failures in this regard) and also provides a significant quantitative improvement in the energetic ranking of the best low RMSD cluster. There remain in most cases a small number (5–10) of competing misfolded structures with low energies; discrimination of these from the native-like topology, necessary for truly reliable tertiary structure prediction, will be a major objective of subsequent work. The reduction of the huge phase space of possible tertiary assemblies to a short list of discrete alternatives does, however, clearly represent progress in the nature and parameterization of the potential function.

We next examined the effect of replacing secondary structure elements derived from the PDB with idealized strands and helices, at the same locations. This substitution examines the effects of helix and strand distortion from ideal geometry on the predicted tertiary fold. Our conclusion is that, while there are occasional cases where substantial effects are observed, particularly for β-strands where a major distortion in length is manifested, the quality of the

results is in general comparable to that obtained using PDB-derived secondary structure elements. This suggests that a folding protocol that initially uses idealized geometries and subsequently refines these geometries by allowing distortions is likely to be successful; furthermore, even if it is necessary in some cases to incorporate distortion directly into the initial simulations, the perturbations induced are relatively small and hence handling them should be computationally tractable.

Finally, we attempted genuine *ab initio* prediction by using predicted, rather than PDB-derived (in either geometry or location), secondary structure, focusing on small helical proteins. Recent improvements in secondary structure prediction, as exemplified by the PSIPRED code of Jones [26], allowed impressively accurate secondary structure predictions to be generated in many cases. When errors in secondary structure were made, the most difficult to deal with were cases in which a long helix was incorrectly predicted to be two short helices, or when two short helices were incorrectly predicted to be a single long helix. Reliable prediction of tertiary structure for α-helical proteins will clearly require secondary structure prediction methods than can robustly discriminate these two cases. Other types of large errors, such as replacement of a helix by a strand or a loop, produced variable results; in some cases, the predictions were surprisingly good despite such major errors. Smaller errors—for example, in length or position of a predicted helix—generally led to relatively minimal quantitative degradations in accuracy as compared to the use of PDB-derived secondary structure. Results for two small, helical CASP3 targets were presented which compared well with the work of other groups [11], including those employing fold recognition methods [46].

While there is still clearly a lot of work to be done, the above results are encouraging with regard to the possibility of developing reliable *ab initio* methods for protein structure prediction to low resolution, at least for small helical proteins. A different direction to pursue is the combination of these methods with fold recognition techniques (threading) and with experimental data, specifically NMR and X-ray crystallographic information. We have demonstrated in previous work [16] that the combination of a tertiary folding potential with sparse NMR constraints can successfully produce structures in the 2–4 Å resolution regime even for large systems; improvements in the folding potential will enhance the utility of such methods.

Acknowledgments

This work was supported in part by grants from the NIH: National Institute of General Medical Sciences (GM-52018) and National Center for Research Resources (P-41 RR06892). We also thank the Intel Corporation for donation of a large cluster of high-end Pentium-based workstations, without which many of these calculations could not have been carried out. Finally, many other computations were performed using SGI Origin machines at NCSA, via the NPACI program run by the NSF.

References

1. V. A. Eyrich, D. M. Standley, A. K. Felts, and R. A. Friesner, *Proteins Struct. Funct. Genet.* **35**, 41–57 (1999).
2. V. A. Eyrich, D. M. Standley, and R. A. Friesner, *J. Mol. Biol.* **288**, 725–742 (1999).
3. J. R. Gunn, A. Monge, R. A. Friesner, and C. H. Marshall, *J. Phys. Chem.* **98**, 702–711 (1994).
4. D. M. Standley, J. R. Gunn, R. A. Friesner, and A. E. McDermott, *Proteins Struct. Funct. Genet.* **33**, 240–252 (1998).
5. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, in *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987, pp. 107–132.
6. F. C. Berstein, T. F. Koetzle, G. J. B. Williams, E. F. j. Meyer, M. D. Brice, J. R. Rodjers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112**, 535–542 (1977).
7. G. Casari, and M. J. Sippl, *J. Mol. Biol.* **224**, 725–732 (1992).
8. P. D. Thomas and K. A. Dill, *J. Mol. Biol.* **257**, 457–469 (1996).
9. K.-C. Chou, M. Pottle, G. Nemethy, Y. Ueda, and H. Scheraga, *J. Mol. Biol.* **162**, 89–112 (1982).
10. A. Kolinski, A. Godzik, and J. Skolnick, *J. Chem. Phys.* **98**, 7420–7433 (1993).
11. C. A. Orengo, J. E. Bray, T. Hubbard, L. LoConte, and I. Sillitoe, *Proteins Struct. Funct. Genet.* **34**, 149–170 (1999).
12. L. Wang, T. Oconnell, A. Tropsha, and J. Hermans, *J. Mol. Biol.* **262**, 283–293 (1996).
13. J. Moult, T. Hubbard, K. Fidelis, and J. T. Pedersen, *Proteins Struct. Funct. Genet.* **34**, 2–6 (1999).
14. J. P. A. Kocher, M. J. Rooman, and S. J. Wodak, *J. Mol. Biol.* **235**, 1598–1613 (1994).
15. K. Yue and K. A. Dill, *Protein Sci.* **5**, 254–261 (1996).
16. D. M. Standley, V. A. Eyrich, A. K. Felts, R. A. Friesner, and A. E. McDermott, *J. Mol. Biol.* **285**, 1691–1710 (1999).
17. M. E. Mortenson, *Geometric Modeling*, 2nd ed., Wiley, New York, 1997.
18. P. D. Thomas and K. A. Dill, *Proc. Nat. Acad. Sci. USA* **93**, 11628–11633 (1996).
19. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577–2637 (1983).
20. M. H. Hao and H. A. Scheraga, *Curr. Opin. Struct. Biol.* **9**, 184–188 (1999).
21. U. Hobohm and C. Sander, *Protein Sci.* **3**, 522–524 (1994).
22. J. Kuszewski, G. A. M., and G. M. Clore, *J. Am. Chem. Soc.* **121**, 2337–2338 (1999).
23. Z. Q. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84**, 6611–6615 (1987).
24. H. C. Romesburg, *Cluster Analysis for Researchers,* Lifetime Learning Publications, Belmont, CA, 1984.
25. J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, *Proteins Struct. Funct. Genet.* **34**, 204–208 (1999).
26. D. T. Jones, *J. Mol. Biol.* **292**, 195–202 (1999).
27. A. Kolinski and J. Skolnick, *Proteins Struct. Funct. Genet.* **18**, 353–366 (1994).
28. A. R. Ortiz, A. Kolinski, and J. Skolnick, *J. Mol. Biol.* **277**, 419–448 (1998).
29. R. Samudrala, Y. Xia, E. Huang, and M. Levitt, *Proteins Struct. Funct. Genet.* **34**, 194–198 (1999).
30. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, *J. Mol. Biol.* **268**, 209–225 (1997).

31. K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, *Proteins Struct. Funct. Genet.* **34**, 171–176 (1999).
32. J. Skolnick, A. Kolinski, and A. R. Ortiz, *J. Biomol. Struct. Dyn.* **16**, 381–396 (1998).
33. B. Rost and C. Sander, *J. Mol. Biol.* **232**, 584–599 (1993).
34. B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235**, 13–26 (1994).
35. J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, *Bioinformatics* **14**, 892–893 (1998).
36. J. A. Cuff and G. J. Barton, *Proteins Struct. Funct. Genet.* **34**, 508–519 (1999).
37. R. D. King and M. J. E. Sternberg, *Protein Sci.* **5**, 2298–2310 (1996).
38. D. Frishman and P. Argos, *Protein Eng.* **9**, 133–142 (1996).
39. D. Frishman and P. Argos, *Proteins Struct. Funct. Genet.* **27**, 329–335 (1997).
40. A. A. Salamov and V. V. Solovyev, *J. Mol. Biol.* **247**, 11–15 (1995).
41. M. J. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. E. Sternberg, *J. Mol. Biol.* **195**, 957–961 (1987).
42. B. A. Reva, A. V. Finkelstein, and J. Skolnick, *Fold. Des.* **3**, 141–147 (1998).
43. F. E. Cohen and M. J. E. Sternberg, *J. Mol. Biol.* **138**, 321–333 (1980).
44. D. J. Osguthorpe, *Proteins Struct. Funct. Genet.* **34**, 186–193 (1999).
45. A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **96**, 5482–5485 (1999).
46. A. G. Murzin, *Proteins Struct. Funct. Genet.* **34**, 88–103 (1999).