Protein Engineering vol. 16 no. 9 pp. 659–663, 2003
DOI: 10.1093/protein/gzg085

A new computational approach for real protein folding prediction

Ben Zhuo Lu[1,2], Bao Han Wang[3], Wei Zu Chen[1] and Cun Xin Wang[1,4]

[1]College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, [2]Department of Astronomy and Applied Physics, University of Science and Technology of China, Hefei 230026 and [3]Institute of Biophysics, Academia Sinica, Beijing 100101, China

[4]To whom correspondence should be addressed.
E-mail: cxwang@bjut.edu.cn

An effective and fast minimization approach is proposed for the prediction of protein folding, in which the 'relative entropy' is used as a minimization function and the off-lattice model is used. In this approach, we only use the information of distances between the consecutive C_α atoms along the peptide chain and a generalized form of the contact potential for 20 types of amino acids. Tests of the algorithm are performed on the real proteins. The root mean square deviations of the structures of eight folded target proteins versus the native structures are in a reasonable range. In principle, this method is an improvement on the energy minimization approach.

Keywords: minimization/off-lattice model/protein folding prediction/relative entropy

Introduction

Recently, there has been a great deal of interest in studying the prediction of tertiary structure of proteins or protein folding, using theoretical models. Some good prediction results [with root mean square deviation (r.m.s.d.) values ranging from 3 to 7.5 Å for small proteins] can be obtained by using methods such as threading based on information on known structure (Moult *et al*., 1999; Venclovas *et al*., 1999). However, it is well established that for small proteins the information contained in the amino acid sequence is sufficient to determine the folded structure, which is the structure with minimum free energy (Anfinsen, 1973). Thus, the native structure is dictated by the physical interactions between amino acids in the sequence. Generally, some direct methods, such as the Monte Carlo method, molecular dynamics or other methods, so called *ab initio* methods, are used to minimize the system's energy (Hinds and Levitt, 1994; Shakhnovich, 1994; Huang *et al*., 1999; Lee *et al*., 1999; Zhou and Karplus, 1999). However, in the usual minimization method the entropy effect is not taken into account and the predicted structure does not necessarily correspond to the state with the lowest free energy. Here, we propose a new and simple algorithm for protein folding calculations using the off-lattice model, in which a new minimization function called 'relative entropy' (defined below) is used other than the system's Hamiltonian. In the off-lattice model, the folding prediction mainly deals with the tendency of the protein backbone. In this approach, we have only used the distances between the consecutive C_α atoms along the peptide chain and a generalized form of the contact potential for 20 types of amino acids. Unlike the traditional energy minimization methods (Mumenthaler and Braun, 1995; Sun *et al*., 1995) starting from a rigid or semi-rigid secondary structure element and then assembling it into a compact structure, in our prediction procedure we tried to start from a nearly random initial conformation. Tests of the algorithm on real proteins were carried out. The consensus results are of generally good quality, yielding eight sample predictions with r.m.s.d.s in the range 3.9–6.8 Å from their native structures. The relation between free energy and the methodology is discussed.

Theory and computational method

Assume that $H(s,r)$ is the Hamiltonian of a protein molecule with the sequence $S = (s_1, s_2, \ldots, s_n)$ and configuration $r = (\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n)$, where $\vec{r}_i$ is the position coordinate of the C_α atom of the ith amino acid residue. Instead of directly minimizing the system Hamiltonian, the quantity called 'relative entropy' G minimized here is defined as

$$G(\{\vec{r}_i\}) = \sum_{\{s_i\}} P_\alpha \ln(P_\alpha/P_0) \tag{1}$$

where P_α and P_0 mean different probabilities. For a given configuration $r = \{\boldsymbol{r}_i\}$, the probability P_0 that the molecule adopts sequence $S = \{s_i\}$ is written as

$$P_0 = \frac{1}{Z_0}\, e^{-\beta H(s,r)}$$
$$Z_0 = \sum_{\{s_i\}} e^{\beta H(s,r)} \tag{2}$$

The probability P_α that the molecule has a specially assigned sequence (the target sequence) $S^\alpha = \{s_i^\alpha\}$ is defined as

$$P_\alpha = \frac{1}{Z_\alpha}\, e^{\beta H(s,r)} \prod_i \delta_{s_i,s_i^\alpha}$$
$$Z_\alpha = \sum_{\{s_i\}} e^{-\beta H(s,r)} \prod_i \delta_{s_i,s_i^\alpha} \tag{3}$$

From the definition, it is easily shown that the 'relative entropy' $G \geqslant 0$. Obviously, the measure G has minimum value when $P_0 = P_\alpha$. The property of G determines that it is a measure to evaluate the difference between the distribution functions P_0 of random sequence and the distribution P_α with the specially assigned sequence in sequence space. Thus, the difference between the distributions P_0 and P_α is not directly evaluated by $P_0 - P_\alpha$, but by the mean of the difference of their logarithms. When the protein conformation closes to the native structure adopted by the given sequence, the difference, in the sense of measure G, between the distribution P_0 and the distribution P_α tends to be small. This is because on the native conformation of

the given sequence S_α, P_0 has a very large value at least on this sequence S_α from the evolutionary point of view, which results in a small value of G. Our aim is to find a 'good' distribution function P_0 closest to P_α through minimizing G, rather than a minimum energy or free energy of the system through energy minimization. Therefore, the folding prediction can be done by searching the conformational space to find an optimal structure $\{\vec{r}_i\}$ for a sequence $\{s_i^\alpha\}$ through minimizing G. The gradient descent algorithm for minimization can be expressed as

$$\frac{\mathrm{d}\vec{r}_i}{\mathrm{d}t} = -\eta\frac{\partial G}{\partial \vec{r}_i} = -\eta\sum_{\{s_i\}}\left[\frac{\partial P_\alpha}{\partial \vec{r}_i}\ \ln\ \frac{P_\alpha}{P_0} + P_0\frac{\partial}{\partial \vec{r}_i}\left(\frac{P_\alpha}{P_0}\right)\right]$$
$$= -\eta\sum_{\{s_i\}}\left[\frac{\partial P_\alpha}{\partial \vec{r}_i}\ \ln\ \frac{P_\alpha}{P_0} + \frac{\partial P_\alpha}{\partial \vec{r}_i} - \frac{P_\alpha}{P_0}\frac{\partial P_0}{\partial \vec{r}_i}\right] \quad (4)$$

where η is an adjustable parameter with a value between 0 and 1 for controlling the convergence speed and the subscript i denotes the ith C_α atom. As shown in Equation 4, the measure G substitutes the system's potential $<H>_\alpha$ used in energy minimization.

Moreover, using the definition of P_α in Equation 3, it is found that the spatial derivative

$$\frac{\partial P_\alpha}{\partial \vec{r}_i} = -\beta\frac{\mathrm{e}^{-H(s,r)}\prod_k \delta_{s_k\cdot s_k^\alpha}}{\mathrm{e}^{-H(s^\alpha,r)}}\left[\frac{\partial H(s,r)}{\partial \vec{r}_i} - \frac{\partial H(s^\alpha,r)}{\partial \vec{r}_i}\right]$$

is zero, because the last two terms in the brackets are equal owing to the limitation of the Kronecker delta functions. Then, using P_0 and P_α defined in Equations 2 and 3, Equation 4 can be reduced to

$$\frac{\mathrm{d}\vec{r}_i}{\mathrm{d}t} = -\eta\sum_{\{s_i\}}\left\{-\frac{P_\alpha}{P_0}\frac{\partial}{\partial \vec{r}_i}\left[\frac{\mathrm{e}^{-\beta H(s,r)}}{\sum_{\{s_i\}}\mathrm{e}^{-\beta H(s,r)}}\right]\right\}$$
$$= -\eta\beta\left[<\frac{\partial H}{\partial \vec{r}_i}>_\alpha - <\frac{\partial H}{\partial \vec{r}_i}>_0\right] \quad (5)$$

where the second term in the brackets, $<\frac{\partial H}{\partial \vec{r}_i}>_0$, is a mean value that is independent of the amino acid sequence, but dependent on conformation space $\{\boldsymbol{r}_i\}$.

A simple Hamiltonian of protein system using the type of the contact potential is written as

$$H = \frac{1}{2}\sum_{i,j\neq i} U(s_i,s_j)A(\vec{r}_i - \vec{r}_j) \quad (6)$$

where $A(\vec{r}_i - \vec{r}_j)$ denotes the contact strength depending on the distance r_{ij} between the ith and jth residues, $U(s_i,s_j)$ is the contact potential between residues s_i and s_j. For a real protein, we have chosen a simple form of $U(s_i,s_j)$ given by Li *et al.* (Li *et al*., 1997), who showed that a simple equation could be used as a good approximation of Miyazawa and Jernigan's 20×20 potential matrix elements (Miyazawa and Jernigan, 1985, 1996), which are statistically deduced pairwise interaction potential energies among the 20 types of amino acids:

$$U(s_i,s_j) \cong c_2 q_i q_j + c_1(q_i + q_j) + c_0 \quad (7)$$

where the subscripts i and j label the 20 amino acid residues and the three coefficients are set to $c_0 = -1.38$, $c_1 = 5.08$ and $c_2 = -7.40$, in units of RT, the gas constant times room temperature. Each s_i corresponds to a value q_i [see Table I in Wang and Lee (Wang and Lee, 2000)].

Using Equations 5 and 6, the final numerical iteration equation can be deduced from Equation 4 as

$$\vec{r}_i^{\,k+1} - \vec{r}_i^{\,k} = -\eta\beta\sum_{j\neq 1}[U(s_i^\alpha,s_j^\alpha) - \langle U(s_i,s_j)\rangle_0]\frac{\partial}{\partial \vec{r}_i^{\,k}}A(\vec{r}_i^{\,k} - \vec{r}_j^{\,k}) \quad (8)$$

where the superscript k represents the kth iteration, $\beta = 1/RT$ and $<U(s_i,s_j)>_0$ denotes the average contact potential with respect to the probability distribution P_0. In order to search the configurations in real space, a continuous form of the function $A(\vec{r}_i - \vec{r}_j)$ [denoted $A(r_{ij})$ for simplicity] is adopted in our work:

$$A(r_{ij}) = \frac{1}{\sqrt{2n\pi}}\ \mathrm{e}^{-(r_{ij}^2 - d^2)/2n} + \varepsilon\left(-\frac{\sigma^2}{r_{ij}^2} + \frac{\sigma^4}{r_{ij}^4}\right) \quad (9)$$

where d is a value around the contact distance between residues (it is set to 6.0 Å) and n, ε and σ are adjustable parameters. The distance between the connective residues is constrained by the SHAKE algorithm as a bond (Ryckaert *et al*., 1977), and therefore the interaction between any two connective residues is skipped. The first exponential term in Equation 9 can be considered as a continuous approximate form of a delta function, which vanishes quickly when $r_{ij} > d$. This term together with the $U(s_i,s_j)$ factor counts for the major driving force of protein folding: hydrophobic and hydrophilic interactions, in which the d value (6.0 Å) just corresponds to the contact distance. The second term in Equation 9 can be considered as an additional term used to prevent some residues, such as hydrophobic residues, from moving closely. In this work, a van der Waals-like potential is used, which has a smoother distance dependence than the ordinary van der Waals function. This term results in a potential barrier at a small contact distance between two residues around 2 Å.

In order for a sequence to fold into a stable native structure, it is reasonable to suggest that the native state has an energy that is much lower than the energies of the bulk of misfolded states, especially of the denatured states (Shakhnovich and Gutin, 1993; Deutsch and Kurosky, 1996; Shakhnovich, 1998). It also holds that there is a large energy gap between the energy of the ground state and the average energy of all possible conformations. As treated below, the average energy of a protein is considered to be approximately equal to that of the denatured state, because the conformational space is mainly occupied by the denatured states. The average energy is related to the term $<U(s_i,s_j)>_0$ in Equation 8. Because the energy of the native state is less than the average, minimizing the quantity in Equation 8 is likely to result in a large energy gap. In addition, our method is essentially to find a structure with higher occupation probability, and therefore the predicted structure should correspond to the conformation with nearly the lowest free energy. With some modifications of this minimization function, it is just the form of the difference in free energy used in the work on reverse protein folding (Deutsch and Kurosky, 1996). Our method is nearly identical with those used in the learning theory of 'neural networks', the distinction being that the probability functions are Gibbs distributions in this work.

A difficult task is to estimate the value of the term $<U(s_i,s_j)>_0$ in Equation 8. It can be estimated from the mean field theory. However, in order to test our computational approach, we

prefer to estimate $<U(s_i,s_j)>_0$ by a more simplified approach. Because the conformational space is mainly occupied by the denatured or misfolded states, the mean energy $<H>_0$ can be considered as the mean energy of denatured states, which can be estimated by calculating the energy of the denatured state of the given sequence. $<H>_0$, the ensemble average of $H(s,r)$ over the sequence S, is expressed as

$$< H >_0 = \frac{1}{2}\sum_i \sum_{j\neq i} < U(s_i, s_j) >_0 A(r_{ij}) = \frac{N(N-1)}{2} < U(s_i, s_j) >_0 \tilde{A} \quad (10)$$

where $<U(s_i,s_j)>_0$ can be assumed to be a constant independent of s_i, s_j for simplicity and $\tilde{A}$ denotes the average of $A(r_{ij})$.

The energy of the native state for the given sequence $s^\alpha = \{s_i^\alpha\}$ can be divided into two parts:

$$H^\alpha = \frac{1}{2}\sum_{i,j\neq i} U(s_i^\alpha, s_j^\alpha)A(r_{ij}) = \sum_i^{N-1} U(s_i^\alpha, s_{i+1}^\alpha)A(r_{i,i+1}) + \frac{1}{2}\sum_{i,j\neq i\pm 1} U(s_i^\alpha, s_j^\alpha)A(r_{ij}) \quad (11)$$

where the second term corresponds to the long-range interaction. The denatured state energy H^α is estimated by neglecting all long sequence-range interactions, all those terms whose sequence separation is >1 and taking some average distance between residues. From the above, $<H>_0$ can be taken into account as the denatured state energy H^α, hence Equation 10 can be written as

$$< H >_0 = \frac{N(N-1)}{2} < U(s_i, s_j) >_0 \tilde{A} = \sum_i^{N-1} U(s_i^\alpha, s_{i+1}^\alpha)A(r_{i,i+1}) = \overline{A}\sum_i^{N-1} U(s_i^\alpha, s_{i+1}^\alpha) = \overline{A}(N-1)\overline{U} \quad (12)$$

where we used two definitions,

$$\overline{A} = \sum_i^{N-1} U(s_i^\alpha, s_{i+1}^\alpha)A(r_{i,i+1}) \Big/ \sum_i^{N-1} U(s_i^\alpha, s_{i+1}^\alpha)$$

and

$$\overline{U} = \frac{1}{N-1}\sum_i^{N-1} U(s_i^\alpha, s_{i+1}^\alpha)$$

Obviously, $\overline{U}$ can be calculated from the given sequence. Thus,

$$\langle U(s_i, s_j)\rangle_0 = \frac{\overline{A}}{\tilde{A}}\frac{2}{N}\overline{U} = k_m\frac{2}{N}\overline{U} \quad (13)$$

where we define an adjustable parameter $k_m = \overline{A}/\tilde{A}$ to avoid calculating $\tilde{A}$ and $\overline{A}$ directly. Now $<U(s_i,s_j)>_0$ can be estimated from the given sequence under a condition of an adjustable parameter k_m.

Equation 8 shows that our approach differs from the energy minimization method by the additional term $<U(s_i,s_j)>_0$. The appearance of this term gives our approach close similarity to other work (Deutsch and Kurosky, 1996) using free energy as a minimization function.

Table I. Properties of the proteins and the results of folding predictions

PDB code	No. of residues	NC of native conformation	NC (and r.m.s.d.) of initial conformation	NC (and r.m.s.d.) of final conformation
1bpi	58	180	0 (16.4)	92 (6.8)
1fcl	56	179	0 (10.0)	96 (5.6)
1ejg	46	144	0 (10.3)	76 (5.2)
1a7f	50	141	0 (18.6)	94 (3.9)
1kde	65	215	6 (20.5)	124 (6.3)
1stu	68	257	11 (17.3)	133 (6.7)
1e68	70	236	0 (19.8)	162 (4.0)
1ubq	76	229	18 (15.6)	135 (5.5)

NC is the number of native contacts. The distance criterion for contact of two residues is 7.5 Å. The values of r.m.s.d. (Å) were obtained from the structures versus the native structures of the PDB.

In all performance tests, we set η in Equation 4 equal to 0.13 and the temperature $T = 1$. To decide simply on the values of the unknown parameters n, ε and σ, the approach was first tested on bovine pancreatic trypsin inhibitor (BPTI). We adjusted the three parameters to achieve the most accurate prediction result and they were finally set as $n = 360$ Å^2, $\varepsilon = 0.17$ and $\sigma = 1.35$ Å. The parameter k_m was assigned a value of 7.2 for our target proteins. The iteration of our algorithm was considered to be convergent when the positional difference between two consecutive iterations for each bead was <0.001 Å. Therefore, the tolerance for SHAKE was set to 0.001 Å. It should be noted that when the initial bond length deviated too far from the constrained length, a harmonic potential was used to constrain the bond length first, followed by the SHAKE constraint method. For a more efficient conformational search, simulated annealing was adopted.

Results and discussion

A simple representation was adopted as the popular off-lattice model, in which a residue was reduced to a bead and its coordinate was in the position of the C_α atom of the residue. Eight small proteins were selected as the tested targets from the Protein Data Bank (PDB) (1bpi, 1fcl, 1ejg, 1a7f, 1kde, 1stu, 1e68 and 1ubq), in which 1a7f has two chains A and B. Each protein was fully denatured as coil to be the initial structure in the folding process, in which the secondary structures and the disulfide bonds in their native structures are obviously broken. The prediction results and the properties of these proteins are given in Table I. It is found that the r.m.s.d. data are in the range of recently reported results for *ab initio* protein fold prediction (Reva *et al*., 1998; Bonneau *et al*., 2001). A value of ~6 Å for the r.m.s.d. has been suggested as a target value for a small protein (Reva *et al*., 1998). The improvement in accuracy of folding prediction is in effect in current *ab initio* protein folding; however, the correct protein folding is still very difficult, especially for folding prediction with a very simplified model, simple residue contact potential and with minimum statistical information on known protein structures (Osguthorpe, 2000). As a comparison, the CASP3 meeting indicated that the absolute accuracy of all *ab initio* methods is still low compared with solving the structure experimentally, with over 90% of predictions for the 'hard' targets having a global r.m.s.d. for C_α >10 Å (Orengo *et al*., 1999).

In our work, the representation of native fold does not depend on known structural information about the target, such

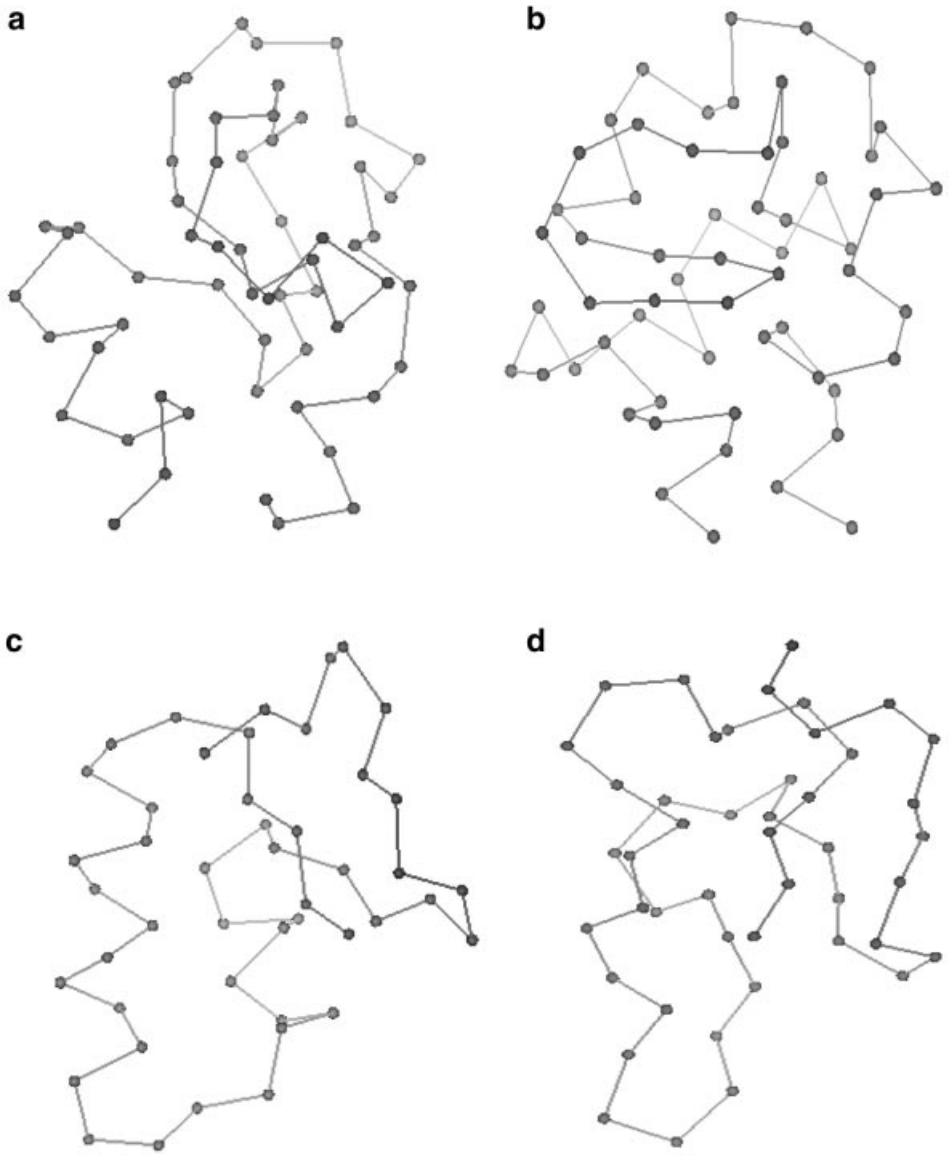

Fig. 1. (**a**) The native structure of 1bpi; (**b**) the final folded structure of 1bpi; (**c**) the native structure of 1ejg; (**d**) the final folded structure of 1ejg.

as the native secondary structure, disulfide bonds and the radius of gyration, except for the distance between any two sequential C_α atoms, which varies slightly around 3.8 Å. Figure 1a–d show the native and final folded structures of 1bpi and 1ejg. The native structures are taken from X-ray structures. Their initial structures before folding are coils without any secondary structure. However, for the peptide 1bpi as seen in Figure 1a and b, the folded structure obviously restores two helices located at two ends and the part sheets that are held in the native structure. For the case of 1ejg (see Figure 1c and d), two segments at the left side of the folded structure are partial helical, and are just helices in the native structure. Therefore, some native secondary structures can be obtained and kept in the folded structure in our folding calculation.

Of our eight target proteins, 1bpi, 1a7f and 1ejg each have three disulfide bonds and the others have no disulfide bonds. Here is an example of the changes in distance between the disulfide-bonded cysteines in the folding process of 1a7f. The A7–B7, A20–B19 and A6–A11 distances are 4.8, 6.3 and 5.5 Å in the crystal structures and increase to 35.7, 34.0 and 18.2 Å in their initial structure and finally become 6.6, 7.4 and 12.4 Å in their corresponding folded structures, respectively. When the disulfide bonds are constrained, the accuracy of prediction can be improved.

The feature of our approach lies in the additional term of the mean contact potential $<U(s_i,s_j)>_0$ in Equation 8. Its value is fairly small and generally much less than the $U(s^\alpha_i,s^\alpha_j)$ for the small proteins. This additional term imposes obvious effects on the predicted results and the convergence speed for real proteins. A comparison between the performances of the two algorithms with and without the term was made for 1ejg and 1bpi. For 1ejg, the algorithm with this term converges after 2257 iteration steps and produces 76 native contacts with an r.m.s.d. of 5.2 Å (144 contacts in total in the native structure). However, the algorithm without the mean potential term of Equation 8 converges after 3673 iteration steps and produces 71 native contacts with an r.m.s.d. of 5.9 Å. For 1bpi, the algorithm with the mean potential term converges after 2728 iteration steps and produces 92 native contacts with an r.m.s.d. of 6.8 Å (180 contacts in total in the native structure).

Table II. Results of folding predictions on 1bpi with different initial conformations

R.m.s.d. of initial conformation	R.m.s.d. of folded conformation	NC of folded conformation	Total contact potential of folded structure
12.6	6.4	91	–286
13.3	6.6	89	–277
15.4	7.1	87	–325
16.5	6.9	90	–295
17.1	7.4	85	–269
19.3	7.9	83	–294
22.3	8.1	82	–346

NC is the number of native contacts. The values of r.m.s.d. (Å) were obtained from the corresponding structures versus the native structures of the PDB. The contact potentials are in units of *RT* (see the text).

However, the algorithm without the term converges after 4390 iteration steps and produces 91 native contacts with an r.m.s.d. of 7.0 Å. In most cases, the algorithm with the mean potential results in an increase in the native contacts and therefore the prediction accuracy. The previous minimization method aims to find a structure with the minimum energy. However, the essence of our method is to search the conformational space with a Boltzmann distribution, namely, to find a structure with a maximum occupying probability. This treatment is able to be more consistent with the argument that the native state of a protein does not just stay on the state with the minimum energy; rather, it corresponds to the state with the lowest free energy.

However, as in other minimization methods, there is still a common problem in this method that the predicted result is dependent on the initial conformation. The folded structure can be trapped in the state with the local energy minimum. Table II lists the predicted results for 1bpi with the different initial conformations. Generally, the larger the deviation of the initial conformation from the native conformation, the worse the prediction accuracy will become. To reveal further the difference between our relative entropy minimization method and energy minimization method, the contact potentials of each folded structure are also shown in Table II. It is interesting that the better predicted conformation (with smaller r.m.s.d.) does not necessarily have the lower potential. This indicates that the energy minimization method might fail to determine the native conformation of a protein. The fact that this method tends to select the folded conformation with lower r.m.s.d., instead of with a lower potential, implies that the entropy effect plays an important role in the determination of the native conformation of a protein. The relative entropy of each folded structure is not given in Table II, because the calculation cannot be achieved directly from our minimization method. Generally, the calculation of relative entropy defined in this work needs very expensive sampling in conformational space for real proteins, which is not the aim of our work. In fact, the advantage of this method in practice is just that the calculation of entropy or free energy is avoided and the degrees of freedom in sequence space are averaged out.

Conclusions

We have presented a new and efficient algorithm for protein folding, which essentially searches the conformation space obeying a Boltzmann distribution to find a conformation on which the probability P_0 is close to P_α of a given sequence. As a reasonable result, the found conformation is near the native structure of the given sequence. This approach is based entirely on physical principles and is fundamentally different from other structure prediction methods that employ homology modeling, threading and statistical comparisons with the known crystal structures. Moreover, this method only adopts a simple, generalized contact potential that does not include angle, torsion and other forms of potential (Lee *et al*., 1999). As a result, this method just predicts the frame of the protein backbone, and the conformation obtained does not contain the detailed structure and image of some parts of the native structure. An additional van der Waals-like potential is required in our method. This term is effective in the short range and actually acts as a repulsion force on closely located residues. Of course, other function forms can replace this term. However, as suggested elsewhere (Li *et al*., 1997), Equation 7 underestimates the attractions between positively and negatively charged amino acids and in contact between both Cys residues. In addition, owing to the statistical nature of Miyazawa and Jernigan's matrix, certain features of interresidue interactions (such as orientational dependence of the interaction, side-chain packing, etc.) are averaged out. Specific features may be necessary for building a more realistic potential for protein folding. In principle, this approach can be applied as a uniform frame for both folding and inverse folding of proteins (Wang *et al*., 1999). With some changes in the definition of P_0 and P_α, in which the sum on $\{s_i\}$ space is changed on $\boldsymbol{r}$ space, Equation 1 can lead to the algorithm for the reverse folding problem (Wang *et al*., 1999).

Acknowledgements

This work was supported in part by the Chinese Natural Science Foundation (No.s 10174005, 30170230 and 29992590–2) and the Beijing Natural Science Foundation (No. 5032002).

References

Anfinsen,C.B. (1973) *Science*, **181**, 223–230.
Bonneau,R., Strauss,C.E.M. and Baker,D. (2001) *Proteins*, **43**, 1–11.
Deutsch,J.M. and Kurosky,T. (1996) *Phys. Rev. Lett.*, **76**, 323–326.
Hinds,D.A. and Levitt,M. (1994) *J. Mol. Biol.*, **243**, 668–682.
Huang,E.S., Samudrala,R. and Ponder,J.W. (1999) *J. Mol. Biol.*, **290**, 267–281.
Lee,J., Liwo,A. and Scheraga,H.A. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2025–2030.
Li,H., Tang,C. and Wingreen,N.S. (1997) *Phys. Rev. Lett.*, **79**, 765–768.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.
Miyazawa,S. and Jernigan,R.L. (1996) *J. Mol. Biol.*, **256**, 623–644.
Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J.T. (1999) *Proteins*, **3** (Suppl.), 2–6.
Mumenthaler,C. and Braun,W. (1995) *Protein Sci.*, **4**, 863–871.
Orengo,C.A., Bray,J.E., Hubbard,T., LoConte,L. and Sillitoe,I. (1999) *Proteins*, **3** (Suppl.), 149–170.
Osguthorpe,D.J. (2000) *Curr. Opin. Struct. Biol.*, **10**, 146–152.
Reva,B.A., Finkelstein,A.V. and Skolnic,J. (1998) *Fold. Des.*, **3**, 141–147.
Ryckaert,J.P., Ciccotti,G. and Berendsen,H.J.C. (1997) *J. Comput. Phys.*, **23**, 327–341.
Shakhnovich,E.I. (1994) *Phys. Rev. Lett.*, **72**, 3907–3910.
Shakhnovich,E.I. (1998) *Fold. Des.*, **3**, R45–R58.
Shakhnovich,E.I. and Gutin,A.M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
Sun,S.J., Thomas,P.D. and Dill,K.A. (1995) *Protein Eng.*, **8**, 769–778.
Venclovas,C., Zemla,A., Fedelis,K. and Moult,J. (1999) *Proteins*, **3** (Suppl.), 231–237.
Wang,B.H. *et al*. (1999) *J. Biosci.*, **24** (Suppl. 1), 61.
Wang,Z.H. and Lee,H.C. (2000) *Phys. Rev. Lett.*, **84**, 574–577.
Zhou,Y.Q. and Karplus,M. (1999) *J. Mol. Biol.*, **293**, 917–950.

Received June 25, 2002; revised June 6, 2003; accepted July 29, 2003