Protein design based on the relative entropy

Xiong Jiao,[1] Baohan Wang,[2,†] Jiguo Su,[1] Weizu Chen,[1] and Cunxin Wang[1,*]
[1]*College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China*
[2]*Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*

(Received 2 January 2006; published 2 June 2006)

An approach to protein design is proposed based on the relative entropy and a reduced amino acid alphabet. In this approach, the relative entropy is used as a minimization object function. The method has been tested on a real protein's off-lattice model successfully, and the results are similar to those obtained from other design studies. It can be applied as a uniform frame for both folding and inverse folding of protein. An iterative calculation method of the ensemble average of the contact strength is proposed at the same time.

DOI: 10.1103/PhysRevE.73.061903 PACS number(s): 87.14.Ee

I. INTRODUCTION

One of the most important problems in molecular biology is protein design [1–3]. When a new conformation is preassigned, amino acid sequences that can fold into the given conformation need to be identified; this is called protein design (or inverse folding). It concerns which sequences can fold and how, into the target conformation. This issue has enormous practical and theoretical significance.

Some proteins have been designed *de novo* and tested experimentally, but this is mostly limited to small protein. Theoretically, protein design is mainly based on the matching of sequences and structures. There are two models mostly used in protein design; one is a detailed model based on the atomic level, and the other is a simple model based on the amino acid backbone. In the simplified model of the protein, side chains are either replaced by effective atoms or not represented at all. Therefore, the simplified model of a protein has fewer degrees of freedom than the atomic level representations of proteins. It is possible to get a larger variation in sequence or in backbone conformation. A coarse-grained model is used in this paper to simplify the calculation.

As for the energy function describing the interactions in a protein, it can also be distinguished: one type is an atomic level energy function, based on physical principles, and includes covalent bonds and nonbonded interactions; and the other energy function is based on statistics, at the amino acid level, such as the Miyazawa-Jernigan (MJ) matrix [4,5]. There are two difficulties for the atomic level potential function used for protein design [6]. First, it is not practical to get the free energy difference between native and non-native states since free energies are difficult to calculate computationally through a search of conformation space. Second, these functions often come from fits to small-molecular data; although they are useful, their use is limited in protein folding and protein design. The most common contact potential used for protein design at the amino acid level is the MJ matrix. This set of potential parameters was derived statistically from real proteins, based on the frequencies of residues appearing in the protein. A predigested form of the MJ matrix is used as the energy function in this paper.

The search methods used in protein design can be classed into two categories. One is stochastic methods, such as the Monte Carlo (MC) method and genetic algorithms. In these methods, the sampling of the sequence space is semirandom and moves toward low-energy sequences. Some appropriate algorithms for numerical calculation based on the MC method have been recently proposed [1–3,7], such as the design technique carried out by Shakhnovich and Gutin (SG) [8–10]. A dual MC procedure was devised by Seno *et al.* and tested within the framework of two simple lattice models in two dimensions [7]. It is found that their procedure is more successful than the SG method. Although this design technique leads to a higher design success in principle, it will take considerable CPU time to explore the sequence space and therefore the method is difficult to apply to a larger system. The other category of design methods is deterministic algorithms, such as the dead end elimination algorithm and mean field algorithm [11–13]. These methods perform semiexhaustive searches in the sequence space. In this paper, the search of the sequence space is performed by a quasiparallel arithmetic so that the design calculation is not time consuming.

In terms of the characteristics of the 20 types of amino acids, the number of possible amino acid types can be reduced, and then the search in the sequence space will be simplified in the computation of protein design. In theoretical studies, different simplified models have been considered [14,15]. A simple example is the hydrophobic-hydrophilic (HP) model [16], used in both protein folding and protein design. Despite the success of some algorithms of protein design tested in the HP model, the HP model is problematic when it is used for doing a fold or design study. A design for a longer sequence (48-mer) with the HP model was not successful [17]. The decoys often have a noncompact conformation with a lower energy than the target conformation [17]. Shakhnovich has pointed out that design schemes will fail occasionally for a HP model, but when more types of amino acids are involved in the model, the situation will get better [18]. It becomes easier to get a funneled energy landscape with a more realistic multiple-letter model. So, motivated by

[*] Corresponding author. Fax: +86-10-67392837. Email address: cxwang@bjut.edu.cn

[†] Email address: wangbhz@yahoo.com.cn

these ideas, protein designing is carried out based on a reduced amino acid alphabet including more than two classes. As an attempt, a model with three classes of amino acids, named the hydrophobic-hydrophilic-neutral (HNP) model, is used in this work.

A self-consistent knowledge-based approach to protein design was developed by Rossi *et al.* [19], in which the 20 types of amino acids are subdivided into three classes, i.e., hydrophobic, neutral, and charged classes. The energy function includes the pairwise interaction and three-body interactions. Through the minimization of the energy function, the results of the design procedure obtained were similar to homologous sequences. The key sites for the folding process also can be identified in agreement with the experiment data.

A method based on the relative entropy to study real protein folding using the off-lattice model has been developed by our group, which has been applied successfully [20]. Then an algorithm for protein design based on the relative entropy and the HP model has been proposed [21,22], in which the relative entropy instead of the Hamiltonian is used as a minimization object function. Better results were obtained compared to other studies with the HP model. However, in the previous work, deducing the ensemble average of the contact strength $[\langle A(r_i-r_j)\rangle_0]$ is under certain conditions limited to the HP model [21,22], and a cursory approximation of $\langle A(r_i-r_j)\rangle_0$ has a relation to the sequence and the type of the object structure. In addition, the approximation calculation of $\langle A(r_i-r_j)\rangle_0$ cannot be used for other models. In this paper, the algorithm is extended from the HP model to the HNP model with an off-lattice model of real proteins, and an iterative calculation of the ensemble average of the contact strength is proposed at the same time. In this case, the shortcoming of the calculation method used in the design based on the HP model will be corrected.

The present paper is organized as follows. First, the algorithm based on the relative entropy and the HNP model is introduced briefly. Then the algorithm is examined on a group of 20 real proteins and the results are discussed. Finally, the conclusion and remarks are presented.

II. THEORY AND METHOD

The total potential energy of a protein takes the form of a sum over all pairwise interactions with a distance-dependent decay term. Assuming $H(r,s)$ is the Hamiltonian of a protein system, it can be expressed as

$$H(r,s)=\frac{1}{2}\sum_{i,j\neq i}^{N}U(s_i,s_j)A(r_i-r_j), \tag{1}$$

where N is the number of residues in the chain, $U(s_i,s_j)$ determines the magnitude of the contact potential between the residues i and j, $S=(s_1,s_2,\ldots,s_n)$ is the sequence of a protein, and $A(r_i-r_j)$ is the contact-strength function which defines the range of the contact potential. Here r_i and r_j are the coordinates of the ith and jth residues. As an object function, the relative entropy G can be written as [21,23]

$$G(s)=\sum_r P_\alpha(r^\alpha,s)\ln\left(\frac{P_\alpha(r^\alpha,s)}{P_0(r,s)}\right), \tag{2}$$

where the subscript α means the object conformation. For a sequence $\{s\}$, $P_0(r,s)$ is the probability that the molecule adopts the conformation $\{r\}$:

$$P_0(r,s)=\frac{1}{Z_0}e^{-\beta H(r,s)}, \tag{3}$$

where $Z_0=\Sigma_{\{r\}}e^{-\beta H(r,s)}$ and the sum is done over all possible conformations. P_α means the probability that the molecule has a given conformation $\{r^\alpha\}$:

$$P_\alpha=\frac{1}{Z_\alpha}e^{-\beta H(r,s)}\prod_i \delta_{r_i,r_i^\alpha}, \tag{4}$$

where $Z_\alpha=\Sigma_r e^{-\beta H(r,s)}\Pi_i\delta_{r_i,r_i^\alpha}\cdot\delta_{r_i,r_i^\alpha}$ is the Kronecker delta function; when $r_i=r_i^\alpha$, $\delta_{r_i,r_i^\alpha}=1$, otherwise, $\delta_{r_i,r_i^\alpha}=0$.

The protein design can be carried out by searching the sequence space through minimizing G to find an optimal sequence $\{s_i\}$ for a given object conformation $\{r_i^\alpha\}$.

The numerical iteration formula can be written as [21]

$$s_i^{k+1}-s_i^k=-\eta\beta\sum_{j\neq i}[A(r_i^\alpha-r_j^\alpha)-\langle A(r_i-r_j)\rangle_0]\left(\frac{\partial U(s_i^k,s_j^k)}{\partial s_i^k}\right), \tag{5}$$

where the superscript k represents the kth iteration, $\beta=1/RT$, T is the absolute temperature, R is the general gas constant, and $\langle A(r_i-r_j)\rangle_0$ is the ensemble average of the contact-strength function over the probability distribution $P_0(r,s)$. The parameter η is an adjustable parameter with a value between 0 and 1 for controlling the iterative convergence speed. r_i^α is the coordinate of the ith residue in the object conformation and r_j corresponds to the coordinate of the jth residue of a protein with any conformation.

A simple HNP model, in which amino acids are subdivided into three classes, is chosen for this algorithm. In the HNP model, hydrophobic residues (H) include Cys, Phe, Tyr, Trp, Met, Leu, Ile, and Val; the hydrophilic ones (P) include Asn, His, Gln, Glu, Asp, Arg, and Lys; and the remaining residues, such as Gly, Pro, Ala, Thr and Ser, are considered the neutral ones (N) [24]. According to the frequencies of the 20 kinds of amino acids in nature, the existing proportion of hydrophobic amino acids can been deduced as 33.56%, 33.39% for hydrophilic ones, and 33.01% for neutral ones. It shows that these three classes have the same frequency in the HNP model. So the calculation of the design can start with a random sequence including three classes of amino acids with the same probability. In addition, if the success rate is calculated based on a random sequence without optimization with this design method, its value is about 33%.

As for the protein, a coarse-grained model is used in this paper. Every amino acid is simplified as a node, 0.3 nm from a C_α atom, located on the line linking C_α and C_β atoms (in other methods, it also can be the geometrical center of the sidechain or a coordinate of its C_β or C_β atom). For Gly without a C_β atom, the coordinate of the C_α atom is used to locate the amino acid.

Define $s_i=1$ for the hydrophobic residue, $s_i=-1$ for the hydrophilic residue, and $s_i=0$ for the neutral one. For the HNP model, the contact potential $U(s_i,s_j)$ between residues can be expanded as

$$U(s_i,s_j) = (1\ s_i\ s_i^2)\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}\begin{pmatrix} 1 \\ s_j \\ s_j^2 \end{pmatrix}, \qquad (6)$$

where a_{ij} is the parameter determined by the certain potential function for the algorithm.

The statistical potential of the interactions between amino acids, the MJ matrix [4,5,25], is divided into nine submatrices according to the HNP model. For every submatrix, the arithmetic average of the matrix elements is computed, which means the interactions between three classes. The results are as follows:

$$\begin{array}{cccc} & H & N & P \\ H & -5.733 & -3.639 & -3.218 \\ N & -3.639 & -2.010 & -1.669 \\ P & -3.218 & -1.669 & -1.651 \end{array}. \qquad (7)$$

According to Eq. (6), we can get

$$U_{HH} = U(1,\ 1) = (1\ 1\ 1)\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -5.733. \qquad (8)$$

With other expressions of U_{HN}, U_{HP}, U_{NP}, U_{NN}, and U_{PP}, the values of a_{ij} in Eq. (6) can be calculated and shown as

$$U(s_i,s_j) = (1\ s_i\ s_i^2)\begin{pmatrix} -2.010 & -0.985 & -0.644 \\ -0.985 & -0.237 & -0.0355 \\ -0.644 & -0.0355 & -0.157 \end{pmatrix}\begin{pmatrix} 1 \\ s_j \\ s_j^2 \end{pmatrix}. \qquad (9)$$

The potential function obtained from the arithmetic average satisfies the following condition [4,25]:

$$U(i,i) + U(j,j) < 2U(i,j). \qquad (10)$$

In the present work, the range of the contact potential is defined as [26]

$$A(r_i - r_j) = \begin{cases} \dfrac{1}{1+e^{r_{ij}-6.5}} & r_{ij} \leqslant 1\text{nm}, \quad \text{and abs}(i-j) > 1 \\ 0 & \text{otherwise} \end{cases}. \qquad (11)$$

As to the calculation method of $\langle A(r_i-r_j)\rangle_0$, which is the ensemble average of the contact strength over the probability distribution P_0 for all the conformations [21], it is very difficult to give a rigorous solution or even a precise estimation. A cursory approximation of $\langle A(r_i-r_j)\rangle_0$ is given by Liu *et al.* [21]. Nevertheless, this approximation is deduced a certain condition: the interaction between the ith and the jth residues is ignored [21]. The deduction of the cursory approximation of $\langle A(r_i-r_j)\rangle_0$ also neglects some long-rang interactions between residues and eliminates the dependence on the object sequence [22]. The approximate calculation of $\langle A(r_i-r_j)\rangle_0$ cannot be extended to other models.

To overcome the shortcoming of the approximation calculation and get a more precise value of $\langle A(r_i-r_j)\rangle_0$, an iterative calculation method is proposed in this paper.

From Eq. (5), we can get

$$\frac{ds_i}{dt} = -\eta\beta\left\{\sum_{j\neq i}[A(r_i^\alpha r_j^\alpha) - \langle A(r_{ij})\rangle_0]\frac{\partial U(s_i s_j)}{\partial s_i}\right\}. \qquad (12)$$

When the iterative calculation is convergent, the value of Eq. (12) should be zero, i.e.,

$$\left\{\sum_{j\neq i}[A(r_i^\alpha r_j^\alpha) - \langle A(r_{ij})\rangle_0]\frac{\partial U(s_i s_j)}{\partial s_i}\right\} = 0. \qquad (13)$$

Since

$$\sum_{j\neq i}\langle A(r_{ij})\rangle_0\frac{\partial U}{\partial s_i} = \langle A(r_{ij})\rangle_{0s}\sum_{j\neq i}\frac{\partial U}{\partial s_i} = \sum_{j\neq i}A(r_{ij}^\alpha)\frac{\partial U}{\partial s_i}, \qquad (14)$$

we get

$$\langle A(r_{ij})\rangle_{0s} = \sum_{j\neq i}A(r_{ij}^\alpha)\frac{\partial U/\partial s_i}{\sum_{j\neq i}\partial U/\partial s_i}. \qquad (15)$$

In Eq. (14), the subscript s means that the value of $\langle A(r_i - r_j)\rangle_0$ is correlative with the designed sequence. To eliminate the dependence on i, Eq. (15) can be written as

$$\langle A(r_{ij})\rangle_{0s} = \frac{1}{N}\sum_i\sum_{j\neq i}A(r_{ij}^\alpha)\frac{\partial U/\partial s_i}{\sum_{j\neq i}\partial U/\partial s_i}. \qquad (16)$$

For making the computing program, Eqs. (11) and (16) are substituted into Eq. (5), which is iterated from a random sequence $\{s_i^{(0)}\}$. The detailed program flow is as follows. First, $\partial U/\partial s_i/\Sigma_{j\neq i}\partial U/\partial s_i$ can be deduced with the analytical form of $U(s_i,s_j)$ and the stochastic sequence $\{s_i^{(0)}\}$. Therefore, the value of $\langle A(r_{ij})\rangle_{os}$ can be calculated from Eq. (16). Second, substituting the value of $\langle A(r_{ij})\rangle_{os}$ into the iterative computation expression Eq. (5), we get a new sequence and one iterative round of calculation is finished. Third, if the sequence is not convergent, a new iterative round will be started with the calculation of the new value of $\langle A(r_{ij})\rangle_{os}$ from the sequence in the previous round. The calculation will end when the iterative round converges at a unique sequence that has no change in two proximate iterative rounds.

In the kth round, the value of S_i^{k+1} computed from Eq. (5) will not be equal to 1 or 0 or −1. In order to let S_i^{k+1} convert to 1, 0, or −1, we give a rule in the next equation as

$$S_i^{k+1,} = \begin{cases} +1 & S_i^{k+1} > 0.5, \\ 0, & \text{abs}(S_i^{k+1}) \leqslant 0.5 \\ -1, & S_i^{k+1} < -0.5. \end{cases} \qquad (17)$$

TABLE I. The calculated results for 20 proteins.

Protein ID	Residue number	Success rate (%)
1aaj	105	44
1aba	87	40
1aps	98	44
1bba	36	50
1ecd	136	47
2lzm	164	46
1erv	105	56
1hel	129	53
1ifb	131	46
1l92	162	47
1mbd	153	45
1osa	148	51
1ra8	159	45
1ycc	108	44
1bbl	37	43
2hpr	87	45
3rn3	124	49
3ebx	62	50
5cpv	108	54
9pap	212	42

The converged predicted sequences of proteins are compared with the experimental sequences of proteins. Then, the success rate can be calculated, is defined as the percentage of the correct classes of amino acid residues predicted with our protein design procedure compared to the total residues.

III. RESULTS AND DISCUSSION

We selected the same 20 proteins used by Micheletti *et al.* [26] to test our method. The deepest descent algorithm for minimization has a common problem that the predicted result is dependent on the initial sequence. The results shown in this paper are arithmetic averages of 50 000 independent design calculations. Table I reports our calculated results for the 20 proteins. It is found that the average successful rate obtained from the method based on the relative entropy and the HNP model is similar to that obtained by Rossi *et al.* (40–55%) [19]. Therefore, the design method presented in this paper is feasible.

For the HNP model, the success rate of a random sequence is about 33%. However, the success rate is improved to 40–55% through our design approach based on the relative entropy and the HNP model. It is found that the success rate of the HNP model is decreased when the third neutral class is added into the HP model in which the success rate is about 75%. However, the success rate of the HNP model is about 20% higher than the rate for a random sequence.

Naturally occurring homologous sequences, with almost the same steric conformation, have a very low degree of similarity, about 30% [27]. When a predigested alphabet of amino acids is used, the homology threshold will changed to about 55% [19]. This value is very close to the best result obtained with our method.

In this paper, a design method based on the relative entropy is performed on the HNP model. It should be noted that this method also can be used with other models, including a model with more than three classes of amino acids. The iterative calculation method of $\langle A(r_i - r_j)\rangle_0$ proposed in our work can overcome the shortcomings of the approximate calculation. There is no limitation for this iterative calculation method; it can be used as a general method for the calculation of $\langle A(r_i - r_j)\rangle_0$ in protein design.

IV. CONCLUSION

A protein design approach based on the relative entropy and the HNP model is proposed in this paper. The results obtained from this method are similar to those of a recent design study. An iterative calculation method of $\langle A(r_i - r_j)\rangle_0$ is used in the present work. In this method, when the sequence space is searched, the conformation space is searched implicitly at the same time through the calculation of $\langle A(r_i - r_j)\rangle_0$. But the design calculation is not time consuming. This method can be applied to large molecules and an off-lattice model of real proteins.

In addition, the potential function used in this work is calculated with a coarse approximation and is not precise enough. It is expected that a better success rate will be obtained after finding a more precise potential function. Such work for providing a potential function is currently under way.

ACKNOWLEDGMENTS

This work was supported by the Chinese Natural Science Foundation (Grant No. 10574009) and the Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20040005013).

[1] J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).

[2] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **89**, 4163 (1992).

[3] M. P. Morrissey and E. I. Shakhnovich, Folding Des. **1**, 391 (1996).

[4] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).

[5] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[6] J. G. Saven, Chem. Rev. (Washington, D.C.) **101**, 3113 (2001).

[7] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).

[8] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).

[9] E. I. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993).

[10] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[11] J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters, Nature (London) **356**, 539 (1992).

[12] B. Dahiyat and S. L. Mayo, Protein Sci. **5**, 895 (1996).

[13] B. Dahiyat, C. A. Sarisky, and S. L. Mayo, J. Mol. Biol. **273**, 789 (1997).

[14] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).

[15] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).

[16] K. A. Dill, Biochemistry **24**, 1501 (1985); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[17] E. I. Shakhnovich, Folding Des. **3**, 45 (1998).

[18] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[19] A. Rossi, C. Micheletti, F. Seno, and A. Maritin, Biophys. J. **80**, 480 (2001).

[20] B. Z. Lu, B. H. Wang, W. Z. Chen, and C. X. Wang, Protein Eng. **16**, 695 (2003).

[21] Y. Liu, B. H. Wang, C. X. Wang, and W. Z. Chen, Sci. China, Ser. G **46**, 659 (2003).

[22] Y. H. Wang, B. H. Wang, Y. Liu, W. Z. Chen, and C. X. Wang, Chin. Sci. Bull. **49**, 5 (2004).

[23] B. Kull and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).

[24] T. P. Li, K. Fan, J. Wang, and W. Wang, Protein Eng. **16**, 323 (2003).

[25] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).

[26] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Proteins **32**, 80 (1998).

[27] C. Chothia and A. M. Lesk, EMBO J. **5**, 823 (1986).