Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 20010B309309) and the National Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KJCX2-W1).

References

1. Nielson, M. A., Chuang, I. L., *Quantum Computation and Quantum* Information. Cambridge: Cambridge University Press, 2000.
2. Bennett, C. H., DiVincenzo, D. P., Quantum information and computation, Nature, 2000, 404(6775): 247—255.
3. Shor, P., Polynomial-time algorithm for prime factorization and discrete logarithms on a quantum computer, SIAMJ J. Comput., 1997, 26(5): 1484—1509.
4. Grover, L. K., Quantum mechanics helps in searching for a needle in a haystack, Phys. Rev. Lett., 1997, 79(2): 325—328.
5. Wei Daxiu, Luo Jun, Yang Xiaodong et al., NMR experimental realization of seventh-order coupling transformations and the seven-qubit modified Deutsch-Jozsa algorithm, LANL e-print, 2003, quant-ph/0301041.
6. Bennett, C. H., Wiesner, S. J., Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states, Phys Rev Lett, 1993, 69(20): 2881—2884.
7. Bennett, C. H., Brassard, G., Crépeau, C. et al., Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels, Phys. Rev. Lett., 1993, 70(13): 1895—1898.
8. Bennett, C. H., Quantum cryptography using any two nonorthogonal states, Phys. Rev. Lett., 1992, 68(21): 3121—3124.
9. Luo, J., Wei, D. X., Xiao, L. et al., Quantum entanglement and information transmission between non-direct-coupled qubits in an array of spatially fixed qubits, Chin. Phys. Lett., 2002, 19(1): 7—9.
10. Liu, X. S., Long, G. L., Tong, D. M. et al., General scheme for superdense coding between multipaties, Phys. Rev. A, 2002, 65(2): 022304.
11. Crudka, A., Wójcik, A., Symmetric scheme for superdense coding between multipaties, Phys. Rev. A, 2002, 66(1): 014301.
12. Mattle, K., Weinfurter, H., Kwiat, P. G. et al., Dense coding in experimental quantum communication, Phys. Rev. Lett., 1996, 76(25): 4656—4659.
13. Fang, X. M., Zhu, X. W., Feng, M. et al., Experimental implementation of dense coding using nuclear magnetic resonance, Phys. Rev. A, 2000, 61(2): 022307.
14. Jones, J. A., NMR quantum computation, Prog. NMR Spectrosc., 2001, 38(4): 325—360.
15. Cory, D. G., Laflamme, R., Knill, E. et al., NMR based quantum information processing: Achievements and prospects, Fortschr. Physik, 2000, 48(9—11): 875.
16. Sharf, Y., Havel, T. F., Cory, D. G., *Spatially encoded pseudopure* states for NMR quantum-information processing, Phys. Rev. A, 2000, 62(5): 052314.
17. Gershenfeld, N. A., Chuang, I. L., Bulk spin-resonance quantum computation, Science, 1997, 275(5298): 350—356.
18. Cory, D. G., Fahmy, A. F., Havel, T. F., Ensemble quantum computing by NMR-spectroscopy, Proc. Natl. Acad. Sci. USA, 1997, 94(5): 1634—1639.
19. Cory, D. G., Price, M. D., Havel, T. F., Nuclear magnetic resonance spectroscopy: An experimentally accessible paradigm for quantum computing, Physica D, 1998, 120(1-2): 82—101.
20. Ernst, R. R., Bodenhausen, G., Wokaun, A., *Principles of Nuclear* Magnetic Resonance in One and Two Dimensions, Oxford: Oxford, 1987.

(Received August 18, 2003; accepted November 10, 2003)

Chinese Science Bulletin 2004 Vol. 49 No. 5 426—431

A generalized approach for protein design based on the relative entropy

WANG Yihua[1], WANG Baohan[2], LIU Yun[1], CHEN Weizu[1] & WANG Cunxin[1]

1. College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China;
2. Institute of Biophysics (IBP), Chinese Academy of Sciences, Beijing 100101, China

Correspondence should be addressed to Wang Cunxin (e-mail: cxwang@bjut.edu.cn)

Abstract **In the present study, we have developed the method brought forward recently for protein design based on the relative entropy. The new approach can be used in more common situation other than the special limits in the anterior method. The results indicate that our generalized method has increased the prediction precision for protein sequence and will be in favor of the study for protein design.**

Keywords: protein design, relative entropy, off-lattice model.

DOI: 10.1360/03ww0146

Protein design can be also called inverse protein folding which is a very important and challenging topic in molecule biology. The aim of protein design is to find out the amino acid sequence which admits a given target structure as its stable conformation. Since protein function is mainly connected with its three-dimensional structure, there is a broad application prospect for protein design in the field of molecule design and drug design. The study of inverse protein folding is also groundwork of *de novo* design of protein. We can use this method to deal with proteins in nature as well as to create proteins that never appear in nature. For the field of protein design, many appropriate algorithms[1-11] have been proposed and gain great achievements. When Shakhnovich and Gutin[1-3] met difficulty using the method to the Hamiltonian of the molecule system of protein as a minimization function (SG method), Kurosky and Deutsch pointed out that the free energy of the system could not be neglected and thus introduced a free-energy-based design method (KD method)[4,5], which expanded out the free energy by keeping the lowest cumulant with itself. The KD method has been tested on the lattice model using the simulated annealing and proved to be better than the SG method[4,5]. Seno et al.[8] devised a dual Monte Carlo (MC) procedure for searching in both the conformation space and the sequence space and tested it within the framework of two simple lattice models. It is found that their procedure is more successful than the SG and KD methods mentioned above. Unfortunately, the dual MC method will consume

CPU time and this problem becomes apparent when the method is applied to the larger system. There is no report about the dual MC procedure on the off-lattice model of real proteins up to now.

In order to overcome the difficulty in use of the dual MC procedure for the large molecules and the off-lattice model, we have proposed a new algorithm[13—16] for protein design based on the relative entropy. Our method can overcome the problem in the SG method because the Hamiltonian minimization is replaced by the relative entropy minimization in the method. It is found that the method is more effective and faster than the previous ones in the test for a group of protein. When we established the method for protein design based on the relative entropy, some approximations are correct only under the special condition. Actually, these approximations restrict the use of the method although they are logical in physics principle. In this paper, some new approximations were put forward to make the general use of the method as well as to improve the prediction accuracy of the protein sequence.

1 Theory and method

Assuming $H(S, r)$ is Hamiltonian of a protein system, it can be expressed as a type of the contact potential:

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

where N is the total residue number, $S=(s_1, s_2, \cdots, s_n)$ is the sequence of a protein and r_i is the coordinate of the residue i. $U(s_i, s_j)$ is the contact potential between the residues i and j, which can be written as

$$U(s_i,s_j)=a_0+a_1s_i+a_2s_j+a_3s_is_j, \tag{2}$$

where s_i, s_j are the residue sequences of a protein, and a_0, a_1, a_2, a_3 are the potential parameters[14]. The function $A(r_i - r_j)$ is the contact intensity between the residues i and j, which has a continuous value between 0 and 1. It is defined as[11]

$$A(r_i-r_j)=\begin{cases}(1+e^{10r_{ij}-7.5})^{-1} & \text{if } j\notin\{i-1,i,i+1\}\\ 0, & \text{otherwise,}\end{cases} \tag{3}$$

where r_{ij} is the distance between the residues i and j, the unit of r_{ij} is in nm. The bigger the value of $A(r_i-r_j)$, the more compact the contact between the residues i and j. The normal energy minimization is to optimize the system's Hamiltonian and find the best sequences with the Hamiltonian minimum from all possible sequences for the given target structure $\{r_i^{\alpha}\}$. The method for protein design based on the relative entropy is that the relative entropy $G(S)$ is minimized as an object function instead of directly minimizing the Hamiltonian of the system to find the best sequence with the minimum entropy. The deepest descent algorithm for minimization of $G(S)$ can be expressed as

$$\frac{\mathrm{d}s_i}{\mathrm{d}t}=-\eta\frac{\partial}{\partial s_i}G(S), \tag{4}$$

where s_i is the ith residue sequence, and η is an adjustable parameter with a value between 0 and 1 for controlling the iterative convergence speed. The numerical iteration formula[14] is obtained as

$$s_i^{k+1}-s_i^k=-\eta\beta\sum_{j\neq i}[A(r_i^{\alpha}-r_j^{\alpha})-<A(r_i-r_j)>_0](a_1+a_3s_j^k), \tag{5}$$

where the superscript k represents the kth iteration, $\beta= 1/RT$. $A(r_i^{\alpha}-r_j^{\alpha})$ is the contact strength function between the ith and the jth residues for the given structure α. r_i^{α} is the coordinate of the ith residue i for the target structure α, r_i corresponds to the coordinate of ith residue of protein with any structure. $<A(r_i-r_j)>_0$ is the ensemble mean value of contact strength function $A(r_i-r_j)$ over the probability distribution P_o. Because the function of $G(S)$ is minimized with restrict to sequence s, this algorithm has a fast iteration pace. The classical hydrophobic and polar model (HP model)[17—19] was chosen to test the algorithm, in which twenty amino acid residues were sorted into the hydrophobic (H) and polar (P) ones. Defining $s_i = 1$ for the hydrophobic residue and $s_i = -1$ for the polar residue, eq. (5) can be expressed as

$$s_i^{k+1}=-\mathrm{sgn}\left(\eta\beta\sum_{j\neq i}[A(r_i^{\alpha}-r_j^{\alpha})-<A(r_i-r_j)>_0](a_1+a_3s_j^k)\right), \tag{6}$$

where sgn ( ) is the sign function, i.e.

$$\mathrm{sgn}(x)=\begin{cases}1 & \text{if } x\geqslant 0,\\ -1 & \text{otherwise.}\end{cases} \tag{7}$$

The key to eq. (6) is to calculate $<A(r_i-r_j)>_0$. However, it is very difficult to calculate the value of $<A(r_i-r_j)>_0$ or only give its precise more estimate value for the off-lattice model. In Ref. [14], an approximation to $<A(r_i-r_j)>_0$ for the off-lattice model was given as follows. When the iteration converges with the deepest descent algorithm, we can get $\frac{\mathrm{d}s_i}{\mathrm{d}t}=0$. Substituting it into eq. (5), we obtain

$$\sum_{j\neq i}^{N}A(r_i^{\alpha}-r_j^{\alpha})(a_1+a_3\bar{s}_j)=\sum_{j\neq i}^{N}<A(r_i-r_j)>_0(a_1+a_3\bar{s}_j), \tag{8}$$

where $\bar{s}_j$ denotes the residue sequence for the given structure α after the iteration is convergent. However, the value of $\bar{s}_j$ is unknown for protein design. Therefore, an approximation to $< A(r_i - r_j) >_0$ was used to solve this problem in Ref. [14] as follows. The term $\bar{s}_j$ in eq. (8) can be neglected in the case of $|a_1| >> |a_3|$, then we can get $a_1 + a_3\bar{S}_j \approx a_1$. If the sum in eq. (8) takes over all subscripts i and $< A(r_i - r_j) >_0$ is regarded as a constant $\overline{A^1}$ without relation with both i and j, from eq. (8), we finally have

$$\overline{A^1} = \sum_i \sum_{j \neq i} A(r_i^\alpha - r_j^\alpha)/[N(N-1)]. \quad (9)$$

The approximation to $< A(r_i - r_j) >_0$ is better but it is still fairly crude. It is clear that the assume of $|a_1| >> |a_3|$ means that the 4th term of eq. (2) can be neglected. Nevertheless, it is very important to select a proper contact potential in protein design. Although in the assume the property of $U(s_i, s_j)$ can satisfy the condition on the contact potential function suggested by Miyazawa et al.[20] or Maiorov et al.[21], the method given by Ref. [14] is not general enough. In fact, the term of $a_3 s_i s_j$ embodies the relationship between residues i and j, which plays a pivotal role for the structure of forming protein. More general situation is that, once the selection of a_1 and a_3 makes $U(s_i, s_j)$ satisfy the essential condition that the contact potential between two hydrophobic residues is the strongest while the contact potential between two polar residues is the feeblest (i. e. $U(1,1) < U(1,-1) < U(-1,-1) \leqslant 0$), the method is applicable without any other conditions. The goal of this work is to give a new approximation to $< A(r_i - r_j) >_0$, which can make the method for protein design based on the relative entropy general. The basic ideal is as follows. First, the right side of eq. (8) can be transformed as

$$\sum_{j \neq i}^{N} < A(r_i - r_j) >_0 (a_1 + a_3\bar{s}_j) = A_i \sum_{j \neq i}^{N} (a_1 + a_3\bar{s}_j). \quad (10)$$

The basic idea for the change from eq. (8) to eq. (10) is according to the way to calculate the center of mass of the multiparticle system. In the multiparticle system, we can get

$$\sum_i m_i r_i = r_c \sum_i m_i, \quad (11)$$

where m_i and r_i are the mass and coordinate of the ith particle, respectively, and r_c is the center of mass of the system. In eq. (10), $< A(r_i - r_j) >_0$, $a_1 + a_3\bar{S}_j$ and A_i correspond with m_i, r_i and r_c in eq. (11), respectively. Note that A_i is relative with i. Substituting eq. (10) into eq. (8), we have

$$A_i = \frac{\sum_{j \neq i} A(r_i^\alpha - r_j^\alpha)(a_1 + a_3\bar{s}_j)}{\sum_{j \neq i} (a_1 + a_3\bar{s}_j)}. \quad (12)$$

In order to calculate the precise value of A_i in eq. (12), the value of $\bar{s}_j$ should be given. However, $\bar{s}_j$ is unknown at beginning of protein design. To solve this problem, we can make the following approximation. It is noticed that the numerator of eq. (12) is the sum taken over subscript j which are unequal to i, and the value of $A(r_i - r_j)$ is between 0 and 1. Consequently, the terms whose values of $A(r_i - r_j)$ approach 0 can be neglected due to their little contribution to the numerator of eq. (12). Hence, we get

$$\sum_{j \neq i} A(r_i^\alpha - r_j^\alpha)(a_1 + a_3\bar{s}_j) \approx \sum{}' A(r_i^\alpha - r_j^\alpha)(a_1 + a_3\bar{s}_j), \quad (13)$$

where Σ' is defined as the sum taken over all the terms $A(r_i^\alpha - r_j^\alpha) \approx 1$. As mentioned above, the contact intensity $A(r_i - r_j)$ between two hydrophobic residues has the maximum value because the contact between two hydrophobic residues is the most compact. We consider approximately $\bar{s}_j = 1$ when $A(r_i^\alpha - r_j^\alpha) \approx 1$. Of course, there is a possible case of $\bar{s}_j = -1$ when $A(r_i^\alpha - r_j^\alpha) \approx 1$, but it is insignificant. According to eq. (2), the terms that have $A(r_i - r_j)$ with a value close to 1 and $\bar{s}_j = -1$ can be neglected because their contact potential $U(s_i, s_j)$ is quite small. Taking all factors above into consideration, the above approximation we get is reasonable. The numerator of eq. (12) can be expressed as

$$\sum_{j \neq i} A(r_i^\alpha - r_j^\alpha)(a_1 + a_3\bar{s}_j) \approx (a_1 + a_3)\sum{}' A(r_i^\alpha - r_j^\alpha). \quad (14)$$

It is found from eq. (3) that $A(r_i - r_j) \approx 1$ when $r_{ij} \leqslant 0.75$ nm. Therefore, Σ' becomes the sum taken over all terms $r_{ij} \leqslant 0.75$ nm instead of the sum taken over all possible j for $A(r_i^\alpha - r_j^\alpha) \approx 1$. The function image of $A(r_i - r_j)$ is shown in Fig. 1.

The denominator of eq. (12) can be treated as the following approximation. Defining $\lambda = m / n$, here m is the number of hydrophobic residues, and n is the number of polar residues. If N is the total number of residues, we can get $N = m + n$. Substituting λ into the denominator of eq. (12), we have

$$\sum_{j \neq i} (a_1 + a_3\bar{s}_j) = \sum_{j \neq i} a_1 + a_3 \sum_{j \neq i} \bar{s}_j \approx (N-1)a_1 + a_3(m-n)$$

$$= (N-1)\left(a_1 + a_3\frac{\lambda-1}{\lambda+1}\right). \quad (15)$$

Substituting eq. (14) and eq. (15) into eq. (12), we obtain

$$A_i \approx \frac{(a_1+a_3)\sum' A(r_i^{\alpha}-r_j^{\alpha})}{(N-1)\left(a_1+a_3\frac{\lambda-1}{\lambda+1}\right)}. \quad (16)$$

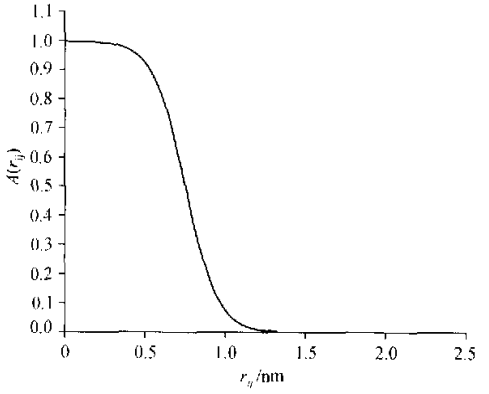

Fig. 1. The image of the contact intensity $A(r_{ij})$, r_{ij} is the distance between two residues, and unit of r_{ij} is in nm.

Note that there are different values of A_i in eq. (16) for the different residues i. In order to cancel the dependence of A_i on i, we let $\overline{A} = \frac{1}{N}\sum_i A_i$ and substitute it into eq. (16). Thus, we get

$$\overline{A} \approx \frac{(a_1+a_3)\sum_i\sum' A(r_i^{\alpha}-r_j^{\alpha})}{N(N-1)\left(a_1+a_3\frac{\lambda-1}{\lambda+1}\right)}. \quad (17)$$

Replacing $< A(r_i - r_j) >_0$ with $\overline{A}$ in eq. (6), we have the numerical iteration formula as

$$s_i^{k+1} = -\mathrm{sgn}\left(\eta\beta\sum_{j\neq i}[A(r_i^{\alpha}-r_j^{\alpha})-\overline{A}](a_1+a_3 s_j^k)\right). \quad (18)$$

Because eq. (18) is obtained without any restriction on the selection of a_1, a_3, the modified method in the present paper is independent of the choice of a_1 and a_3. This means that the contact potential between residues can be chosen without any restrict using the method developed in this paper for sequence prediction. Of course, the fundamental physical principle must be satisfied. For any protein molecule, $\overline{A}$ can be calculated with eq. (17) directly as long as its crystal structure is known. Therefore, the present method improves applicability of protein design. Comparing eq. (17) with eq. (9), we find $\overline{A} \approx \varepsilon\overline{A^1}$, here ε is expressed as

$$\varepsilon = \frac{a_1+a_3}{a_1+a_3\frac{\lambda-1}{\lambda+1}}. \quad (19)$$

Because of $\lambda < 1$ and $U(1,1) < U(1,-1) < U(-1,-1) \leqslant 0$, we can obtain $a_1 < 0$ and $(a_1 + a_3) < 0$. Consequently, we have $\varepsilon > 1$. That means that $\overline{A^1}$, as the approximation of $< A(r_i - r_j) >_0$ in Ref. [14], is smaller than the real value. Obviously, when $|a_1| >> |a_3|$, we can get $\varepsilon \approx 1$ and $\overline{A} \approx \overline{A^1}$ further. Obviously, the approximation of $< A(r_i - r_j) >_0$ in this paper is consistent with that obtained from Ref. [14]. In other words, the method reported in Ref. [14] can be regarded as the special case of the method given in this paper.

2 Results and discussion

A new computational program was generated for predicting protein sequence based on the method given above. The parameters were taken as $\eta = 0.2$, $T = 1$, $a_1 = a_2 = -5$, $a_3 = -2.2$. The classification of residues suggested by Sun et al.[17] was used, i. e. residues of A, V, L, I, C, M, F, Y and W were classified as the hydrophobic residues, and the remaining residues, such as G, P, H, S, T, K, R, D, N, E and Q are considered as the polar ones. The initial sequence was chosen randomly. The convergent results were compared with the real protein sequences, and the success rate was defined as a percentage of the correct class of amino acid residues predicted with our procedure to the total residue number of protein.

The types of protein are classified as four classes[122] according to their secondary structure: the whole α protein (content of α-helix is more than 40% and content of β-sheet is lower than 5%), the whole β protein (content of β-sheet is more than 40% and content of α-helix is less than 5%), the α + β protein (both contents of α-helix and β-sheet are more than 15%, but in separate parts of the structure), and the α /β protein (both contents of α-helix and β-sheet are more than 15%, where both helices and sheets interact and often alternate along the polypeptide chain). The 60 proteins, which were taken from the Protein Data Bank (PDB)[123], were selected to test our method. The homology of all the proteins is lower than 50%. All selected protein included 20 whole α proteins, 20 whole β proteins, 10 α+β proteins and 10 α /β proteins. The analysis results of λ (the ratio of the number of hydrophobic residues to the number of polar residues) for four classes proteins (α, β, α + β, α /β) are 0.727, 0.646, 0.646, 0.733, respectively, and the mean value of λ is 0.688.

First, we have predicted the sequences of four classes proteins using their respective λ. The tested results are listed in Table 1. The arithmetical mean values of success rates of whole α proteins, whole β proteins, α + β protein

Table 1 The test result of four classes proteins

No.	α			β			α+β			α/β		
	PDB ID	Residues number	Success rate (%)	PDB ID	Residues number	Success rate (%)	PDB ID	Residues number	Success rate (%)	PDB ID	Residues number	Success rate (%)
1	1a1w	83	75.9	1a3k	137	83.2	1bm8	99	78.8	1acf	125	74.4
2	1a32	85	75.3	1aly	146	71.9	1bta	89	76.4	1ahn	169	71
3	1ad6	185	73	1cd8	114	72.8	1c9x	124	71	1aiu	105	81.9
4	1aep	153	78.4	1cdy	178	83.7	1d8z	89	74.2	1b1i	122	75.4
5	1bd8	156	70.5	1czs	160	70.6	1ddw	63	63.5	1bfe	110	75.5
6	1buy	166	72.3	1dfx	125	68.8	1ek8	185	82.2	1byr	152	71.1
7	1ddf	127	74.8	1f53	84	79.8	1ekg	119	77.3	1cjw	166	72.3
8	1ed1	114	81.6	1fna	91	79.1	1ew4	106	81.1	1e0s	173	78.6
9	1uxc	50	82	1fnl	173	74	1f7w	144	72.2	1eq6	189	70.4
10	1fk5	93	78.5	1fsc	61	63.9	1gd3	98	75.5	1czk	149	78.5
11	1neq	74	75.7	1g43	160	77.5	–	–	–	–	–	–
12	1hqb	80	78.8	1noa	113	69	–	–	–	–	–	–
13	1hyp	75	76	1nm	91	78	–	–	–	–	–	–
14	1jf2	190	75.3	1nl	102	71.6	–	–	–	–	–	–
15	1mof	53	60.4	1wba	171	74.3	–	–	–	–	–	–
16	1qsq	162	73.5	1wit	93	69.9	–	–	–	–	–	–
17	1r69	63	79.4	2eif	133	68.4	–	–	–	–	–	–
18	1rzl	91	81.3	2fcb	173	76.9	–	–	–	–	–	–
19	1sra	151	72.8	2i1b	153	75.2	–	–	–	–	–	–
20	1bgf	124	70.2	2rhe	114	77.2	–	–	–	–	–	–
Average success rate (%)			75.3			74.3			75.2			74.9

Table 2 The prediction results for 20 proteins

No.	PDB ID	Residues number	Success rate(%)	Success rate (%)[14]
1	1bba	36	52.8	47.2
2	1bbl	37	70.3	73.0
3	3ebx	62	69.4	74.2
4	1aba	87	73.6	73.6
5	2hpr	87	83.9	82.8
6	1aps	98	78.6	76.5
7	1aaj	105	71.4	68.6
8	1erv	105	81.9	84.8
9	1ycc	103	75.7	73.8
10	5cpv	108	71.3	67.6
11	3rn3	124	72.6	67.7
12	1hel	129	78.3	79.8
13	1ifb	131	74.8	69.5
14	1ecd	136	74.3	70.6
15	1osa	148	72.3	72.3
16	1mbd	153	75.8	75.2
17	1ra8	159	74.8	76.7
18	1l92	162	72.2	73.5
19	2lzm	164	72.6	73.8
20	9pap	212	66.5	69.3
Average success rate (%)		–	73.1	72.5

and α /β protein are 75.3%, 74.3%, 75.2% and 74.9%, respectively. This result proves that our method is reasonable. For the 4 types of proteins, there are 25% residues which can not be recognized correctly. An important reason is that the HP model is too simple. As mentioned by Micheletti et al.[11], the division of 20 types of residues just into two classes, H and P, is too coarse. In addition, the value of the average contact intensity, $< A(r_i - r_j) >_0$, has a stronger influence on the prediction precision. To get the approximation for $< A(r_i - r_j) >_0$, we assume $\bar{s}_j = 1$ when $A(r_i^a - r_j^a) \approx 1$. This means that only the contact intensity between two hydrophobic residues is close to 1.

In fact, there are a few cases that the contact intensity between polar residues is also of a stronger value. Therefore, this approximation also brings error. But these instances are infrequent.

Next, the same λ is used for all proteins to predict their sequences. The purpose of this test is to know whether a union λ ($\lambda = 0.688$) can be used for all proteins. The arithmetical mean values of success rates of α, β, $\alpha+\beta$ and α/β proteins are 75.2%, 74.3%, 75.0% and 74.8%, respectively. It is found that the success rates are very close to that mentioned above using the different λ. Therefore, a union λ used for sequence prediction of all proteins is rational in the HP model.

In order to compare our results with those obtained in Ref. [14], we selected the same 20 proteins used in Ref. [14] from PDB. Table 2 lists the PDB codes, the residue numbers and the success rates for 20 proteins with two methods. Our prediction success rate (73.1%) is better than that given in Ref. [14] (72.5%). This indicates that our method for protein design based on relative entropy can be generalized and the precision of our method is also improved.

Through testing the method with several groups of a_1 and a_3, we have found that there is no obvious change of the success rates for using different a_1 and a_3 while some a_1 and a_3 make iteration converge slowly.

3 Conclusion

In this paper we have generalized the method for protein design based on the relative entropy given in Ref. [14], in which the method takes effect only if the contact potential between residues is useful in the special cases. With our work, the contact potential between residues can be selected freely. That makes the method more powerful and precise. The method in Ref. [14] can be regarded as the special case of the present work.

In addition, we also note that the average contact intensity $< A(r_i - r_j) >_0$ has an obvious influence on the predict precision. The more accurate approximation to $< A(r_i - r_j) >_0$ can make the better success rate. We are attempting to find a new approximation to $< A(r_i - r_j) >_0$. Additionally, our model is only limited to the simple HP model. The work for predicting sequence of 20 kinds of amino acid residue using the off-lattice model is currently under way.

Acknowledgements We thank Liu Chunli for the result analysis and making a part of the test program. This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 10174005 and 30170230) and the Beijing Natural Science Foundation (Grant No. 5032002).

References

1. Shakhnovich, E. I., Gutin, A. M., Engineering of stable and fast-folding sequences of model proteins, Proc. Natl. Acad. Sci. USA, 1993, 90: 7195—7199.
2. Shakhnovich, E. I., Gutin, A. M., A new approach to the design of stable proteins, Protein Eng., 1993, 6: 793—800.
3. Shakhnovich, E. I., Proteins with selected sequences fold into unique native conformation, Phys. Rev. Lett., 1994, 72: 3907—3910.
4. Kurosky, T., Deutsch, J. M., Design of copolymeric materials, J. Phys. A: Math. Gen., 1995, 27: L387—L393.
5. Deutsch, J. M., Kurosky, T., New algorithm for protein design, Phys. Rev. Lett., 1996, 76: 323—326.
6. Morrissey, M. P., Shakhnovich, E. I., Design of proteins with selected thermal properties, Fold Des., 1996, 1: 391—405.
7. Sun, S. J., Brem, R., Chan, H. S., et al., Designing amino acid sequences to fold with good hydrophobic cores, Protein Eng., 1995, 8: 1205—1213.
8. Seno, F., Vendruscolo, M., Maritan, A. et al., Optimal protein design procedure, Phys. Rev. Lett., 1996, 77: 1901—1904.
9. Micheletti, C., Seno, F., Maritan, A. et al., Protein design in a lattice model of hydrophobic and polar amino acids, Phys. Rev. Lett., 1998, 80: 2237—2240.
10. Seno, F., Micheletti, C., Maritan, A. et al., Variational approach to protein design and extraction of interaction potentials, Phys. Rev. Lett., 1998, 81: 2172—2175.
11. Micheletti, C., Seno, F., Maritan, A. et al., Design of proteins with hydrophobic and polar amino acids, Proteins, 1998, 32: 80—87.
12. Gutin, A. M., Abkevich, V. I., Shakhnovich, E. I., Chain length scaling of protein folding time, Phys. Rev. Lett., 1996, 77: 5433—5436.
13. Wang, B. H., Yun, Z. X., Wang, Z. X. et al., A unified design approach for the inverse folding and direct folding of protein, J. Bioscience, 1999, 24 (suppl. 1): 61.
14. Liu, Y., Wang, B. H., Wang, C. X. et al., A new approach for protein design based on the relative entropy, Science in China, Ser. G, 2003, 33(4): 348—356.
15. Lu, B. Z., Wang, C. X., Wang, B. H., A new minimization method for real protein folding prediction, Chinese J. Chem. Phys., 2003, 16(2): 117—121.
16. Lu, B. Z., Wang, B. H., Chen, W. Z. et al., A new computational approach for real protein folding prediction, Protein Eng., 2003: 659—663.
17. Lau, K. F., Dill, K. A., A lattice statistical mechanics model of the conformational and sequence spaces of proteins, Macromolecules, 1989, 22: 3986—3997.
18. Chan, H. S., Dill, K. A., Origins of structure in globular proteins, Proc. Natl. Acad. Sci. USA., 1990, 87: 6388—6392.
19. Chan, H. S., Dill, K. A., "Sequence Space Soup" of proteins and copolymers, J. Chem. Phys., 1991, 95: 3775—3787.
20. Miyazawa, S., Jernigan, R. L., Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation, Macromolecules, 1985, 18: 534—552.
21. Maiorov, V. N., Crippen, G. M., Contact potential that recognizes the correct folding of globular proteins, J. Mol. Biol., 1992, 227: 876—888.
22. Chou, K. C., A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins, 1995, 21(4): 319—344.
23. Berman, H. M., Westbrook, J., Feng, Z. et al., The protein data bank, Nucleic Acids Research, 2000, 28: 235—242.

(Received December 2, 2003)