生物物理学报　第二十卷　第四期　二〇〇四年八月
ACTA BIOPHYSICA SINICA　Vol.20　No.4　Aug. 2004

四种结构类型的蛋白质设计方法

刘　赟[1]，　王屹华[1]，　王宝翰[2]，　王存新[1]，　陈慰祖[1]
(1. 北京工业大学生命科学与生物工程研究院，北京 100022；
2. 中国科学院生物物理研究所，北京 100101)

摘　要： 给出了以疏水 - 亲水模型为基础的蛋白质设计方法，该方法以物理学原理为基础，以相对熵作为优化的目标函数。对四种不同结构类型的天然结构的真实蛋白质进行了检测，分析了影响检测成功率的主要因素，结果表明，该方法是普适的，可用于对不同结构类型的蛋白质设计序列。

关键词： 蛋白质设计；蛋白质逆折叠；结构类型；非格子模型
中图分类号： Q615

1　引　　言

蛋白质是生命系统中重要的大分子物质。蛋白质特有的结构特征是其功能形成与展现的根本物理基础。因此，探索蛋白质折叠机制，归纳蛋白质组成与结构之间的关系就成为蛋白质研究中的重要问题。理论和计算对于这一领域问题的解决有着不可忽视的作用。研究折叠问题的目的是从蛋白质序列出发来预测结构，这一研究已获得很大进展[1,2]。蛋白质逆折叠问题又称为蛋白质设计，相当于给定结构和温度，在序列空间中搜索一个（或几个）能折叠到此结构并保持结构稳定的序列。逆折叠问题广泛应用于分子生物学的基础研究以及药物设计领域。

以优化哈密顿量为基础的蛋白质设计方法最早由 Shakhnovich 和 Gutin[3–5]提出，他们用蒙特卡罗方法在目标结构中寻找能量最低的序列（SG 方法），在优化过程中需对疏水与亲水残基的比例作出限制。Deutsch 和 Kurosky[6]进一步发展了以自由能差为基础的蛋白质设计方法，求出了自由能的累积（cumulant）展开的一阶近似（DK 方法）。DK 方法与 SG 方法相比，不需要对残基比例作出限制。 在格子模型上用模拟退火对此方法进行的测试表明，DK 方法优于 SG 方法。Seno 等[7]利用对构象空间和序列空间同时搜索的方法作蛋白质设计，发展出双重蒙特卡罗方法。此方法原则上比前述方法精确，但在双重空间上的搜索需要大量计算机 CPU 时间，很难用于较大体系，对于非格子模型的真实蛋白质结构则更加困难。

为了克服双重蒙特卡罗方法很难用于较大分子体系及非格子模型的困难，我们进一步发展了基于相对熵概念提出的算法[8]，用非格子模型研究蛋白质的折叠问题[9]。在此基础上我们发展了基于相对熵的蛋白质设计方法[10]，用优化相对熵来代替优化哈密顿量，从而克服了用 SG 方法直接优化哈密顿量所遇到的困难。与前述方法相比，此方法更为快速有效。

自然界总共存在有 650 种折叠模式[11]，不同折叠子组合成不同的结构类型，我们的方法对这些结构类型是否普适，以及对不同二级结构中的残基种类的预测是否同样有效，还有待进一步研究。本文讨论了这个问题，进一步分析了该方法的特点。

2　理论与方法

首先简单介绍基于相对熵的蛋白质设计方法[10]。假设 $H(r,s)$是蛋白质分子体系的哈密顿量，可写为

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

收稿日期：2003-10-17
基金项目：国家自然科学基金项目（10174005，30170230）和北京市自然科学基金项目（5032002）
通讯作者：王存新，电话：(010)67392724，
E-mail：cxwang@bjut.edu.cn

其中 N 为残基总数，r_i为第 i 个残基的坐标，$A(r_i-r_j)$表示残基之间相互作用强度，如果 $r_{low}<r<r_{high}$，则 $A(r)>0$，否则 $A(r)=0$，$U(s_i, s_j)$为残基 i 与 j 之间的接触势，$S=(s_1, s_2, \cdots, s_n)$表示蛋白质的残基序列。这里我们采用简单的疏水 - 亲水模型（H/P 模型）来检验我们的算法。把所有氨基酸残基分为两类，即把 Ile(I)、Leu(L)、Val(V)、Phe(F)、Met(M)、Trp(W)、Cys(C)、Tyr(Y)、Pro(P)、Ala(A)归为非极性疏水氨基酸，把其余氨基酸 Gly(G)、Lys(K)、Thr(T)、Ser(S)、Gln(Q)、Asn(N)、Glu(E)、Asp(D)、Arg(R)、His(H)归为极性亲水氨基酸。 定义 $s_i=1$ 时表示疏水残基，$s_i=-1$ 表示亲水残基。

与以前的工作不同，我们用相对熵代替哈密顿量或自由能差作为优化的目标函数，利用最陡下降法搜寻适合目标结构的序列。相对熵定义为[10]:

$$G(s)=\sum_r P_\alpha \ln(P_\alpha/P_0), \tag{2}$$

其中下标 α 表示目标结构。P_0表示在任一构象 $\{r\}$ 的条件下分子具有序列 $\{s\}$ 的几率；对蛋白质外加一个固定势作用于每一个氨基酸残基，使其处于特定构象 $\{r^\alpha\}$，这时出现序列 $\{s\}$ 的几率表示为 P_α。相对熵 $G\geqslant 0$ 且 $P_0=P_\alpha$时 G 有最小值 0。相对熵 $G(s)$ 的极小化就相当于 P_0的极大化。给定目标结构 α，利用最陡下降法搜寻使 P_0最接近于 P_α的序列。用最陡下降法优化相对熵得到的数值迭代公式为[10]:

$$s_i^{k+1}=-sgn\left(b\eta\beta\sum_{j\neq i}[A(r_i^\alpha-r_j^\alpha)-\langle A(r_i-r_j)\rangle_0](1+\omega s_j^k)\right), \tag{3}$$

$\langle A(r_i-r_j)\rangle_0$ 表示对于几率分布 P_0的平均接触强度，r_i^α为目标结构为 α 的蛋白质的第 i 个残基的坐标。sgn 为符号函数，即

$$sgn(x)=\begin{cases}1 & x\geqslant 0\\ -1 & \text{其他情况。}\end{cases} \tag{4}$$

由于$\langle A(r_i-r_j)\rangle_0$ 难以计算，我们把$\langle A(r_i-r_j)\rangle_0$ 粗略地看作是与 i 和 j 无关的常数 A_0, 从而有[10]

$$A_0\approx\sum_i\sum_{j\neq i}A(r_i^\alpha-r_j^\alpha)/[N(N-1)]。 \tag{5}$$

需要指出，$(N-1)A_0$ 表示残基的平均接触数，这里用 N_0 来表示。我们用 A_0 代替（3）式中的 $\langle A(r_i-r_j)\rangle_0$ 进行迭代计算，即

$$s_i^{k+1}=-sgn\left(b\eta\beta\sum_{j\neq i}[A(r_i^\alpha-r_j^\alpha)-A_0](1+\omega s_j^k)\right) \tag{6}$$

在迭代计算中，取 $\eta=0.5$，温度 $T=1$，$b=-0.5$，$\omega=0.01$。从任意初始序列开始进行迭代计算。迭代收敛以后再把收敛的结果与蛋白质真实序列进行比较，可以得到检测的成功率（预测正确的残基数与蛋白质链长的百分比）。在计算中，接触强度函数选为下面的形式[10]：

$$A(r_i-r_j)=\begin{cases}(1+e^{r_{ij}-6.5})^{-1} & j\notin|i-1,\ i,\ i+1|\text{且 } r_{ij}<1\text{ nm}\\ 0 & \text{其他情况。}\end{cases} \tag{7}$$

Micheletti 和 Seno 等[12]将在 C_α 与 C_β 连线上、距 C_α 原子 0.3 nm（从 C_α 指向 C_β 方向）的点作为残基坐标，由于 GLY（甘氨酸）残基没有 C_β 原子，就用 C_α 原子的坐标标示位置。本文采用这种方法确定蛋白质残基的坐标。进行氨基酸残基序列预测时，对每一个残基只知道一个特定点的坐标，在此基础上预测残基种类[4,12]。

3 结果与讨论

Micheletti 等[12]直接给出自由能 $F(s)$表达式并用于设计真实蛋白质，他们以 $H(r,s)-F(s)$作为目标函数，用蒙特卡罗方法对 20 个蛋白质做检测，其序列预测的平均成功率为 73.4%。为了便于与他们的结果相比较，我们从蛋白质数据库挑选出相同的 20 个蛋白质对我们的方法进行检验[10]。利用（5）式与（6）式，得到的检测成功率为 72.5%，与 Micheletti 等人的结果接近；以 A_0 作为调节参数作迭代计算可以把成功率提高到 75.2%, 这表明我们的方法是可行的。

为了验证该方法的普适性，我们选取不同结构类型蛋白质进行测试。蛋白质结构类型主要有 4 种[13]：全 α 型、全 β 型、α 与 β 分离型（α+β）、α 与 β 相间型（α/β）。 我们选取了 49 个全 α 型蛋白质（α 螺旋含量都大于 40%、β 折叠含量都小于 5%, 见表 1）、36 个全 β 型蛋白质（ α 螺旋含量小于 5%、β 折叠含量大于 40%，见表 2）、10 个 α+β 型蛋白质（α 螺旋含量大于 15%、β 折叠含量大于 15%且 α 螺旋和 β 折叠相间交叉排布，见表 3）、36 个 α/β 型蛋白质（α 螺旋含量大于 15%、β 折叠含量大于 15%且 α 螺旋和 β 折叠分隔排布，见表 4）进行测试，总共选取了 135 个单链蛋白质，所有蛋白质的同源性小于 50%。这些蛋白质的晶体坐标均来自 PDB (Protein Data Bank)[14]。

Table 1 The test results for structure of α-Helix with two procedures[1)]

Label	PDB code	Residue number	N_0	Success rate1(%)	Success rate2(%)
1	1JF2	190	3.52	70.0	75.7
2	1JLI	112	3.85	66.9	68.7
3	1A1W	83	3.57	78.3	78.3
4	1A56	81	3.44	71.6	72.8
5	1A6M	151	3.61	75.4	74.8
6	1K95	161	3.59	78.8	80.1
7	1A7W	68	2.78	77.9	66.1
8	1LFB	77	3.06	74.0	74.0
9	1AD6	185	3.83	73.5	75.1
10	1LKI	172	3.63	72.6	70.9
11	1AEP	153	4.07	74.5	75.8
12	1AF3	145	3.57	77.2	82.7
13	1MBA	146	3.65	70.5	73.2
14	1AIL	70	2.71	71.4	71.4
15	1MYT	146	3.36	78.7	75.3
16	1ASH	147	4.39	59.1	57.8
17	1AX8	130	3.72	74.6	73.8
18	1B09	206	3.61	74.7	76.6
19	1NKD	59	3.04	77.9	69.4
20	1B0B	141	3.48	76.5	76.5
21	1NKL	78	3.76	79.4	76.9
22	1B0X	72	3.14	75.0	72.2
23	1B1B	140	3.89	65.7	67.1
24	1B5L	152	3.83	73.6	75.6
25	1BD8	156	4.28	72.4	73.7
26	1PBV	195	3.46	75.3	77.9
27	1PRB	53	3.25	73.5	69.8
28	1BGF	124	3.52	71.7	69.3
29	1BKR	108	3.77	74.0	75.9
30	1QSQ	162	3.69	73.4	75.3
31	1R69	63	3.48	74.6	79.3
32	1BUY	166	3.52	72.8	74.0
33	1RZL	91	3.74	73.6	80.2
34	1SRA	151	3.51	73.5	73.5
35	1TOP	162	3.11	73.4	74.0
36	1UXC	50	3.99	82.0	80.0
37	1C75	71	3.11	80.2	78.8
38	1YCC	103	3.04	73.7	79.6
39	1CC5	83	2.92	68.6	67.4
40	1CTJ	89	2.95	71.9	69.6
41	1DLY	121	3.36	74.3	70.2
42	1ED1	114	3.63	78.0	81.5
43	1FK5	93	3.32	77.4	78.4
44	1GJH	164	3.11	77.4	79.2
45	1HE9	131	3.40	74.8	78.6
46	1HLB	157	3.38	71.9	70.7
47	1HQB	80	3.61	73.7	76.2
48	1HYP	75	3.23	80.0	73.3
49	1HZI	129	3.83	73.6	73.6
Average				74.2	74.3

[1)]Success rate1 and success rate2 denote two kinds of success rates. 1: Substituting into A_0 iteration Equation (6); 2: Substituting $4.49/(N-1)$ into A_0 of Equation (6)

把由（5）式确定的A_0代入（6）式进行迭代计算，相对全α型、全β型、α与β分离型（α+β）和α与β相间型（α/β）这4种结构类型的蛋白质所得平均成功率（success rate1）依次为：74.2%、72.8%、73.2%、73.8%；对每一种结构类型都以N_0或$(N-1)A_0$作为调节参数（其值依次为4.49、4.88、4.17和4.39），利用（6）式进行迭代计算，可以把平均成功率(success rate2)依次提高到74.3%、75.5%、74.7%和75.2%（见表1~4），这些平均成功率数值接近或高于文献[12]所得的结

Table 2 The test results for structure of β-Sheet with two procedures[1)]

Label	PDB code	Residue number	N_0	Success rate1(%)	Success rate2(%)
1	1A1X	106	3.34	66.0	66.0
2	1A3K	137	3.60	78.8	81.7
3	1AAC	105	3.58	70.4	70.4
4	1ALY	146	3.77	69.8	73.2
5	1AT0	142	3.97	69.0	66.9
6	1BKB	132	3.25	78.7	75.7
7	1BWB	198	3.70	77.2	79.7
8	1BXV	91	3.49	73.6	71.4
9	1C1L	135	3.42	69.6	68.8
10	1CD8	114	3.43	71.0	75.4
11	1CDY	178	3.74	78.6	83.1
12	1D2S	170	3.75	77.0	81.7
13	1DFX	125	3.48	64.8	72.0
14	1EJ8	140	3.51	65.0	73.5
15	1F53	84	3.84	71.4	78.5
16	1F94	63	2.95	74.6	65.0
17	1FNA	91	2.99	74.7	80.2
18	1G0X	192	3.55	78.6	82.8
19	1G1C	196	3.59	72.9	72.9
20	1G43	160	3.75	70.6	77.5
21	1IFG	140	3.25	73.5	77.1
22	1J8S	193	3.63	68.3	71.5
23	1NEU	115	3.15	63.4	74.7
24	1NOA	113	3.19	70.7	73.4
25	1PFS	156	3.19	69.2	79.4
26	1QFP	118	3.40	80.5	76.2
27	1QMC	104	3.37	82.6	84.6
28	1SFP	111	3.40	70.2	72.9
29	1TEN	89	3.53	79.7	83.1
30	1TNM	91	3.70	73.6	81.3
31	1TUL	102	3.55	66.6	72.5
32	1WBA	171	3.77	70.7	74.2
33	1WIT	93	3.90	73.1	72.0
34	2EIF	133	2.95	72.1	66.9
35	2FCB	173	3.59	75.1	79.1
36	2FNB	95	3.12	75.7	73.6
37	2I1B	153	3.56	69.2	75.8
38	2MCM	112	3.29	73.2	76.7
39	2RHE	114	3.19	78.9	80.7
Average				72.8	75.5

[1)]Success rate1 and success rate2 denote two kinds of success rates. 1: Substituting into A_0 iteration Equation (6); 2: Substituting $4.88/(N-1)$ into A_0 of Equation (6)

Table 3 The test results for structure of α+β with two procedures[1)]

Label	PDB code	Residue number	N_0	Success rate1(%)	Success rate2(%)
1	1A62	122	3.07	71.3	77.0
2	1B1I	122	3.54	74.5	75.4
3	1B2V	173	3.19	73.4	78.6
4	1BM8	99	3.89	78.7	78.7
5	1BT0	73	3.77	68.4	72.6
6	1BXE	108	3.59	73.1	75.0
7	1BYW	110	4.44	64.5	65.4
8	1DDW	109	3.39	77.0	74.3
9	1G2R	94	3.50	76.5	74.4
10	1GD3	98	3.46	74.4	75.5
Average				73.2	74.7

[1)]Success rate1 and success rate2 denote two kinds of success rates. 1: Substituting into A_0 iteration Equation (6); 2: Substituting 4.17/(N−1) into A_0 of Equation (6)

Table 4 The test results for structure of α/β with two procedures[1)]

Label	PDB code	Residue number	N_0	Success rate1(%)	Success rate2(%)
1	1A3C	166	4.00	77.7	75.3
2	1ACF	125	3.61	78.4	82.4
3	1AHO	315	4.47	69.5	69.8
4	1AHN	169	3.98	73.9	75.7
5	1AIU	105	3.65	82.8	81.9
6	1B24	169	3.49	73.9	77.5
7	1B87	181	3.49	73.4	74.5
8	1BFE	110	3.20	76.3	77.2
9	1BTN	106	3.40	75.4	78.3
10	1BUD	197	4.20	71.0	71.0
11	1BYR	152	4.26	71.7	71.7
12	1C44	123	3.45	78.0	82.9
13	1CC8	72	3.44	77.7	76.3
14	1CJW	166	3.69	72.8	72.8
15	1CNU	133	3.79	69.1	78.1
16	1CTQ	166	3.99	74.6	80.1
17	1DOI	150	3.99	72.0	74.6
18	1D8Z	89	2.99	78.6	78.6
19	1DI6	183	3.94	69.3	69.9
20	1DIV	149	3.26	76.5	77.1
21	1DTP	190	3.42	68.4	73.6
22	1DUS	192	3.89	75.0	76.5
23	1DV8	128	3.75	64.0	70.3
24	1DZ3	123	3.51	72.3	73.9
25	1F2H	169	2.97	62.7	65.0
26	1F32	127	3.25	73.2	74.8
27	1F4P	147	3.97	77.5	79.5
28	1F7W	144	3.39	72.9	73.6
29	1FAA	121	3.66	76.0	78.5
30	1FUE	163	3.81	76.6	76.6
31	1FUK	157	3.96	80.2	81.5
32	1FZQ	176	3.93	75.5	77.8
33	1GH2	107	3.68	75.7	78.5
34	1GHH	81	2.84	76.5	69.1
35	1GMX	107	3.62	68.2	71.9
36	1GRQ	178	3.44	74.1	68.5
Average				73.8	75.2

[1)]Success rate1 and success rate2 denote two kinds of success rates. 1: Substituting into A_0 iteration Equation (6); 2: Substituting 4.39/(N−1) into A_0 of Equation (6)

果，证明我们的方法对不同结构类型的蛋白质都是适用的。对于全 α 型蛋白质，把$(N-1)A_0$作为调节参数，成功率没有明显提高（见表 1）。对于其它三种结构类型蛋白质，把$(N-1)A_0$作为调节参数，成功率提高的幅度在 1%到 3%之间（见表 2）。可见，对于全 α 型蛋白质，公式（5）的估计较为精确。但总的来说，利用A_0估计值（见（5）式）检测四种结构类型的蛋白质，其结果是相近的，该方法对四种结构类型的蛋白质是普遍适用的。

我们注意到设计的序列与天然蛋白质序列并非完全符合，存在着一定的误差，原因可能是把 20 种蛋白质残基粗略地分为两类（疏水 - 亲水模型）导致的[12]，为了比较不同残基对检测成功率的影响，我们利用 dssp 软件对不同结构类型的蛋白质进行了分析（dssp 软件来自 www.dssp.com）。

对所有的蛋白质来说，残基预测错误率最高的依次是：W、I、C、P、F、V、M、L。 其中疏水残基的错误最多（见图 1A）。 考虑到不同残基在蛋白质中的含量不同，含量较高的残基的预测错误对平均成功率影响较大。把不同残基的错误率乘以相应残基的含量百分比，可以看出，L、A、V、G、E、S 的错误对预测的成功率影响较大（见图 1 B）。

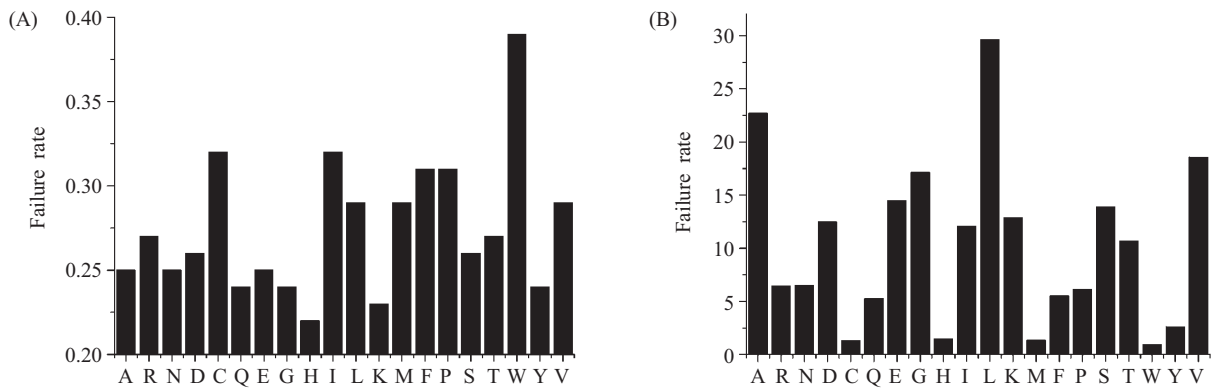

Fig.1 Histogram of the failure rate. (A) Failure rate in identifying the correct H/P class of the 20 amino acids; (B) Failure rate obtained by multiplying failure rates of (A) by the fraction of each residue type in all proteins

我们进一步分析了二级结构中残基的预测错误率。这些二级结构包括 H（α-Helix）、B（β-Bridge）、T（Turn）、E（β-sheet）、G（3_{10}-Helix）、S（Bend），出现错误最多的二级结构依次是 B、T、H、E、G、S，其中 H（α-Helix）中残基的预测错误率达到了 28%，而 E（β-sheet）中残基的预测错误率达 27%。考虑到各个二级结构在所有蛋白质中所占的比例，H（α-Helix）和 E（β-sheet）对预测的成功率影响最大（见图 2A 和 2B），所以我们重点分析 H（α-Helix）和 E（β-sheet）中的残基预测错误率对预测的影响。

在所有的 H（α-Helix）中（总共 4 259 个残基），出现错误最多的残基依次是 W、I、T、G、M、F、V、L（见图 3A）。考虑各个残基在 H（α-Helix）中的含量，L、A、E、K、V、I 的错误对预测的成功率影响较大（见图 3B）。在所有的 E（β-sheet）中（总共 2 946 个残基），出现错误最多的残基依次是 W、C、M、F、L（见图 3C）。考虑各个残基在 E（β-sheet）中的含量，V、L、I 的错误对预测的成功率影响较大（见图 3D）。在 20 种残基中，预测 W（色氨酸）时出现的错误最多，超过 40%。

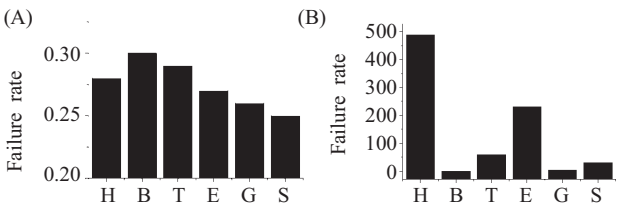

Fig.2 Histogram of the failure rate in the second structures. (A) Failure rates in identifying the correct H/P class of the 20 amino acids in H(α-Helix), B(β-Bridge), T (Turn), E(β-sheet), G(310-Helix) and S(Bend); (B) Failure rates obtained by multiplying failure rates of (A) by the fraction of each second structure in all proteins

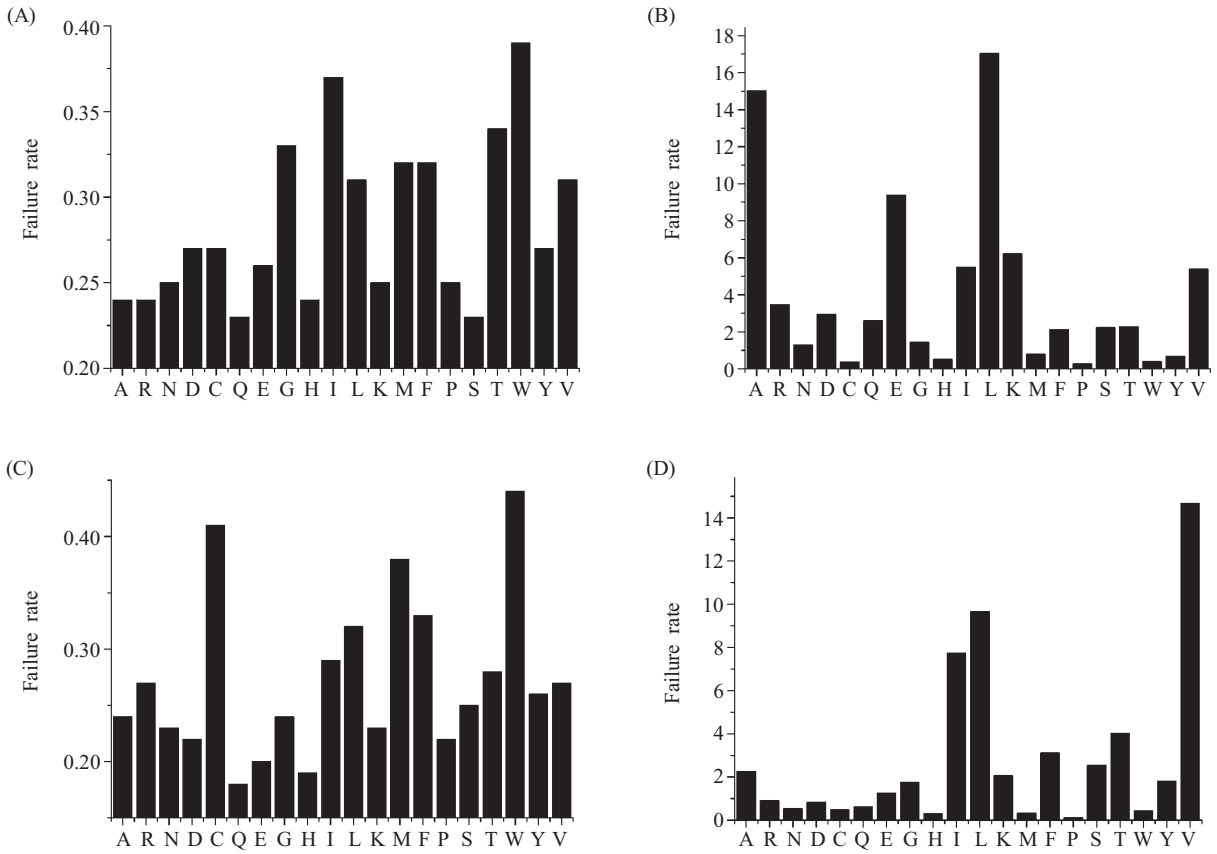

Fig.3 Histogram of the failure rates in α-Helix and β-sheet. (A) Failure rates in identifying the correct H/P class of the 20 amino acids in α-Helix; (B) Failure rates obtained by multiplying failure rates of (A) by the fraction of each residue type in α-Helix of all proteins; (C) Failure rates in identifying the correct H/P class of the 20 amino acids in β-sheet; (D) Failure rates obtained by multiplying failure rates of (C) by the fraction of each residue type in β-sheet of all proteins

4 结　论

本工作的目的主要是发展一种新的蛋白质设计方法，用来克服单纯优化哈密顿量带来的困难，解决同时在序列空间和构象空间搜索带来的计算量大的难题。首先通过预测残基的疏水性证实了理论的有效性[10]；在已有工作的基础上[10]，重点研究了基于相对熵的蛋白质设计方法是否适用于不同结构类型的蛋白质，分析了 20 种残基的预测错误率以及它们对整体预测成功率的影响，考察了二级结构中各残基的预测错误率以及它们对整体预测成功率的影响。证实了我们的方法普遍适用于 4 种结构类型的蛋白质设计。

该方法完全基于物理学原理，以相对熵概念为出发点[8], 搜寻最适合目标结构的序列。对格子模型[8]和非格子所作的检测表明该方法是快速有效的，与同类工作相比，得到了较好的结果。

需要指出，用该方法预测序列时，成功率主要依赖于对$<A(r_i-r_j)>_0$的估计，原则上，对$<A(r_i-r_j)>_0$更精确的估计有可能提高预测的成功率，有关工作还有待进一步改进。

另外，本文仅局限于疏水 - 亲水模型，只能预测残基的极性，由于该方法对不同结构类型的蛋白质都适用，我们将进一步把这一理论从 2 态（疏水 - 亲水模型）发展到 3 态或 5 态的情形。对于各种残基之间相互作用势的研究将有助于把本文的方法最终扩展到预测 20 种残基。

参考文献：

[1] Bryngelson J, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding. *Proteins*, 1995,21:167~195

[2] Fersht A. Nucleation mechanism of protein folding. *Curr Opin*

Struct Biol, 1997,7:10~14

[3] Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA*, 1993,90:7195~7199

[4] Shakhnovich EI, Gutin AM. A new approach to the design of stable proteins. *Protein Eng*, 1993,6:793~800

[5] Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett*, 1994,72: 3907~3910

[6] Deutsch JM, Kurosky T. New algorithm for protein design. *Phys Rev Lett*, 1996,76:323~326

[7] Seno F, Vendruscolo M, Maritan A, Banavar JR. Optimal protein design procedure. *Phys Rev Lett*, 1996,77:1901~1904

[8] Wang BH, Yun ZX, Wang ZX. A unified design approach for the inverse folding and direct folding of protein. *J Bio-science*, 1999,24(suppl 1):61

[9] 卢本卓, 王存新, 王宝翰. 用于真实蛋白质结构预测的一种新的优化方法. 化学物理学报，2003,16(2):117~121

[10] 刘赟，王宝翰，王存新, 陈慰祖. 基于相对熵的蛋白质设计新方法. 中国科学（G辑），2003,33(4):348~356

[11] Wang ZX. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng*, 1998,11(8):621~626

[12] Micheletti C, Seno F, Maritan A, Banavar JR. Design of proteins with hydrophobic and polar amino acids. *Proteins*, 1998,32:80~87

[13] Chou KC. A novel approach to predicting protein structural classes in a (20-1) d amino acid composition apace. *Proteins*, 1995,21:319~344

[14] Berman HM, Westbrook J, Feng Z, Gilliland G. The protein data bank. *Nucleic Acids Research*, 2000,28:235~242

AN APPROACH TO DESIGN PROTEINS OF FOUR STRUCTURAL CLASSES

LIU Yun[1], WANG Yi-hua[1], WANG Bao-han[2], WANG Cun-xin[1], CHEN Wei-zu[1]

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

Abstract: A design method for identifying the correct hydrophobic or polar class of residues was presented in previous work. This approach is based entirely on the physical principle and the relative entropy is used as a minimization object function. The algorithm is used to perform the design on target conformations corresponding to the native states of four structural classes of real proteins. The failure to increase the design success score are also analyzed and discussed. The method is general and can be extended to design proteins of different structural classes.

Key Words: Protein design; Inverse protein folding; Structural classes; Off-lattice model