基于相对熵的蛋白质设计方法的普遍近似

王屹华① 王宝翰② 刘 赟① 陈慰祖① 王存新①*

(①北京工业大学生命科学与生物工程学院, 北京 100022; ②中国科学院生物物理研究所, 北京 100101. * 联系人, E-mail: cxwang@bjut.edu.cn)

摘要 发展了最近提出的一种基于相对熵原理的蛋白质设计方法. 将该方法从必须满足特定的条件才能适用, 推广到普遍适用. 研究表明, 扩展前的方法可被看作扩展后方法的特例, 扩展后的采用普遍近似的蛋白质设计方法更利于蛋白质设计的研究, 使基于相对熵的蛋白质设计方法更具有普遍性, 同时还提高了对蛋白质序列的预测精度.

关键词 蛋白质设计 相对熵 非格子模型

蛋白质设计亦称蛋白质逆折叠, 是分子生物学中一个十分重要且极具挑战性的问题. 它主要研究在已知蛋白质三级结构的情况下, 如何找到适合该三级结构的一级序列. 由于蛋白质只有处在特定空间结构时才能发挥其生物功能, 蛋白质设计的研究在分子设计与药物设计等方面有广泛的应用前景. 蛋白质逆折叠的研究, 是蛋白质全新设计(*de novo* design)的基础.该研究的意义在于, 不仅可用于研究自然界已有的蛋白质, 还能设计出自然界不存在的蛋白质. 在此领域, 已发展出多种不同的预测方法[1~11], 并获得了十分有意义的结果.Shakhnovich 和 Gutin[1~3]采用直接优化体系 Hamilton 量的方法(SG 方法)遇到困难后, Kurosky 和 Deutsch 指出, 体系中的自由能不能忽略, 并发展出求自由能的累积(cumulant)展开的一阶近似方法(KD 方法)[4,5]. 在格子模型上用模拟退火进行的测试表明, KD 方法优于 SG 方法[4,5]. Seno 等人[8]使用双重 Monte Carlo 方法对构象空间和序列空间同时搜索, 在特定格子模型上进行了测试并获得成功. 虽然此方法较前述方法精确, 但该方法的困难在于, 双重 Monte Carlo 搜索需要消耗大量 CPU 时间. 当预测体系较大的蛋白质分子时, 此问题更为突出[12]. 对于非格子模型的真实蛋白质结构, 迄今尚未见到使用该方法的工作报道.

为了克服双重 Monte Carlo 方法较难用于较大分子体系和非格子模型这两个困难, 我们把基于相对熵概念提出的算法应用到蛋白质设计中来, 发展出基于相对熵的蛋白质设计新方法[13~16]. 该方法用蛋白质分子的相对熵来代替 Hamilton 量进行优化, 从而有效地解决了直接优化 Hamilton 量所面临的困难. 用一组真实蛋白质检验后表明, 与前述方法相比, 该方法简单且更为快速有效. 该方法在提出时所做的一些近似[14]原则上是合理的, 但由于这些近似只有在特定条件下才能成立, 从而限制了该方法的适用范围.本工作将提出新的近似方法, 使基于相对熵的蛋白质设计方法更具有普遍性, 同时还提高了对蛋白质序列的预测精度.

1 理论与方法

假设蛋白质分子体系的 Hamilton 量 $H(s, r)$, 可表示为简单的接触势形式

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

其中 N 为残基总数, $S = (s_1, s_2, \cdots, s_n)$表示蛋白质的残基序列, r_i 表示第 i 个残基的坐标. $U(s_i, s_j)$是残基 i 与 j 之间的接触势, 可表示为

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

其中 s_i, s_j 表示蛋白质的残基序列, a_0, a_1, a_2, a_3 为势参数, 具体参数如何设定可见文献[14]. 函数 $A(r_i - r_j)$ 表示残基之间接触强度, 是介于 0 到 1 之间连续变化的值, 可用下公式[11]描述:

$$A(r_i-r_j)=\begin{cases}(1+e^{10r_{ij}-7.5})^{-1}, & \text{如果}\quad j\notin|i-1,i,i+1|;\\ 0, & \text{其他情况,}\end{cases} \tag{3}$$

其中 r_{ij} 表示第 i 个残基与第 j 个残基之间距离(单位: nm). $A(r_i - r_j)$越大, 表明两个残基之间接触越紧密. 一般的能量优化方法是, 对于给定目标构象 $\{r_i^{\alpha}\}$, 从所有可能形成该构象的序列中寻找 Hamilton 量最小的序列作为最佳序列. 而基于相对熵的蛋白质设计方法[14]是, 对于给定的目标构象, 从形成该构象的序列中, 找到相对熵最小的序列, 即对应于该构象的最佳序列. 优化相对熵 $G(s)$的最陡下降法可写为

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

其中 s_i 表示蛋白质的残基序列, η是介于 0 和 1 之间的调节参数, 用来控制迭代的收敛速度. 进而可得优化算法的数值迭代公式[14]为

$$s_i^{k+1}-s_i^k=-\eta\beta\sum_{j\neq i}[A(r_i^\alpha-r_j^\alpha)-\langle A(r_i-r_j)\rangle_0](a_1+a_3s_j^k), \tag{5}$$

其中上标 k 表示第 k 步迭代, $\beta = 1/RT$. $A(r_i^\alpha-r_j^\alpha)$ 为特定构象α下, 第 i 个残基与第 j 个残基间的接触强度. r_i^α 是目标结构为α的蛋白质第 i 个残基的坐标, r_i 为任意结构的蛋白质第 i 个残基的坐标.$\langle A(r_i-r_j)\rangle_0$ 表示接触强度 $A(r_i-r_j)$ 对于几率分布为 P_0 的系综平均值.由于该方法是在平衡位置附近进行优化, 所以具有迭代速度快的优点. 如果使用经典的疏水亲水模型(HP 模型)[17~19]来检验算法, 即把组成蛋白质的 20 个氨基酸分为疏水残基和亲水残基两类. 那么, 对于疏水残基, 定义 $S_i = 1$, 对亲水残基, $S_i = -1$. 在以上基础上, 公式(5)可进一步简化为以下符号函数形式:

$$s_i^{k+1}=-\mathrm{sgn}\left(\eta\beta\sum_{j\neq i}\left[A(r_i^\alpha-r_j^\alpha)-\langle A(r_i-r_j)\rangle_0\right](a_1+a_3s_j^k)\right), \tag{6}$$

其中 sgn 为符号函数, 即

$$\mathrm{sgn}(x)=\begin{cases}1, & \text{如果}\quad x\geqslant 0,\\ -1, & \text{其他情况.}\end{cases} \tag{7}$$

(6)式的关键是求出 $\langle A(r_i-r_j)\rangle_0$. 然而对非格子模型来说, 求出系综平均 $\langle A(r_i-r_j)\rangle_0$ 的值或仅仅给出较精确的估计值都非常困难. 文献[14]求 $\langle A(r_i-r_j)\rangle_0$ 的近似值的方法如下: 在使用最陡下降法迭代收敛时会有 $\frac{\mathrm{d}s_i}{\mathrm{d}t}=0$, 将此条件应用于公式(5)会得到

$$\sum_{j\neq i}^{N}A(r_i^\alpha-r_j^\alpha)(a_1+a_3\overline{S}_j)=\sum_{j\neq i}^{N}\langle A(r_i-r_j)\rangle_0(a_1+a_3\overline{S}_j), \tag{8}$$

其中 $\overline{S}_j$ 表示收敛时, 对应于特定构象α的蛋白质序列. 然而, 在蛋白质设计时, $\overline{S}_j$ 的具体取值是未知的. 为此文献[14]作了以下近似: 假定 $|a_1|>>|a_3|$, 那么(8)式中的 $\overline{S}_j$ 可以忽略, 即有 $a_1+a_3\overline{S}_j\approx a_1$. 对(8)式等号两边关于 i 求和, 并把 $\langle A(r_i-r_j)\rangle_0$ 粗略地看作是与 i 和 j 无关的常数 $\overline{A^1}$, 最终有

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

该近似会产生以下后果: 由于假定 $|a_1|>>|a_3|$, 这相当于残基间接触势 $U(s_i,s_j)=a_0+a_1s_i+a_2s_j+a_3s_is_j$ 中的第 4 项 $a_3S_iS_j$ 被忽略掉了. 在蛋白质设计时, 残基间接触势的选取非常重要. 虽然在此限制条件下得到的 $U(s_i,s_j)$ 也能满足 Miyazawa 与 Jernigan[20]或 Maiorov 和 Crippen[21]提出的关于接触势函数的条件. 但是如果只有忽略了该项, 文献[14]提出蛋白质设计方法才能适用, 这在应用上就缺乏普遍性. 因为 $a_3S_iS_j$ 实际上体现了残基 i 与残基 j 之间的关联, 它在蛋白质的形成过程中起着极为关键的作用. 一般的情况应该是, 只要 a_1, a_3 的选取能使 $U(s_i,s_j)$ 满足 $U(1,1)< U(1,-1)<U(-1,-1)\leqslant 0$ 这一基本条件 (即两疏水残基之间接触势最强, 两亲水残基之间接触势最弱), 不应额外增加条件.本工作的主要目的是给出系综平均 $\langle A(r_i-r_j)\rangle_0$ 新的估计值, 使得基于相对熵的蛋白质设计方法具有普遍性. 现将本工作的基本思想陈述如下: 先对(8)式中等号右边项作如下变化, 即

$$\sum_{j\neq i}^{N}\langle A(r_i-r_j)\rangle_0(a_1+a_3\overline{S}_j)=A_i\sum_{j\neq i}^{N}(a_1+a_3\overline{S}_j). \tag{10}$$

能做上述变化, 是借鉴了求多粒子系统质心的方法. 对于多粒子系统有

$$\sum_i m_ir_i=r_c\sum_i m_i. \tag{11}$$

其中 m_i, r_i 分别为为第 i 个粒子的质量和坐标, r_c 为系统的质心坐标. (10)式中 $\langle A(r_i-r_j)\rangle_0$, $a_1+a_3\overline{S}_j$ 和 A_i 分别对应于(11)式中的 m_i, r_i 和 r_c. 注意 A_i 是和 i 相关的. 将(10)式代入(8)式有

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

要求出上式的精确值, 必须知道 $\overline{S}_j$ 的具体值. 但是此信息在蛋白质设计时是未知的. 为此可作如下近似: (12)式中的分子项是对所有不等于 i 的 j 求和, 而 $A(r_i-r_j)$ 介于 0 到 1 之间. 因此, 所有 $A(r_i-r_j)$ 接近于 0 的项对(12)式分子项的贡献很小, 可近似为

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

其中 $\sum'$ 表示对所有 $A(r_i^a - r_j^a) \approx 1$ 的项求和. 由于疏水残基之间接触最紧密, 所以两疏水残基之间的接触强度 $A(r_i - r_j)$ 最大. 可以近似认为 $A(r_i^a - r_j^a) \approx 1$ 时, $\overline{S}_j = 1$. 当然, 也可能存在 $A(r_i - r_j)$ 接近 1 且 $\overline{S}_j = -1$ 的情况, 但是这种情况很少. 根据(2)式, 此时残基间接触势 $U(s_i, s_j)$ 很小, 所以 $A(r_i - r_j)$ 接近 1 且 $\overline{S}_j = -1$ 的情况可以忽略. 所以上述近似是合理的. 因此, (12)式的分子项可近似为

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

分析(3)式发现, 当 $r_{ij} \leqslant 0.75$ nm 时, $A(r_i - r_j) \approx 1$. 所以, 式中 $\sum'$ 就由对所有 $A(r_i^a - r_j^a) \approx 1$ 的 j 求和, 化简为对所有 $r_{ij} \leqslant 0.75$ nm 的项求和. $A(r_i - r_j)$ 的函数图象见图 1.

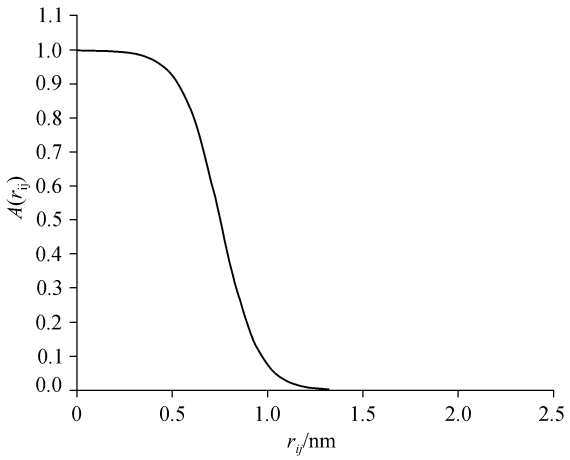

图 1　残基间接触强度函数 $A(r_{ij})$的分布
r_{ij} 表示两残基之间距离

对(12)式的分母项, 可作以下近似. 定义 $\lambda = m/n$, 其中 m 为蛋白质中疏水残基的数目, n 为亲水残基的数目. 如果 N 是蛋白质残基总数, 应有 $N = m + n$. 将 λ 代入(12)式中的分母项, 得到

$$\sum_{j\neq i}(a_1 + a_3\overline{S}_j) = \sum_{j\neq i} a_1 + a_3\sum_{j\neq i}\overline{S}_j \approx (N-1)a_1 + a_3(m-n)$$
$$= (N-1)\left(a_1 + a_3\frac{\lambda-1}{\lambda+1}\right). \quad (15)$$

将(14)、(15)式代入(12)式, 得到

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

注意对于不同残基 i, 上式中 A_i 的取值不一样. 为了取消 A_i 对 i 的依赖, 令 $\overline{A} = \frac{1}{N}\sum_i A_i$, 代入(16)式有

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

用 $\overline{A}$ 取代(6)式中的 $\langle A(r_i - r_j)\rangle_0$, 可得到以下迭代公式:

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

在进行以上推导时, 并未对 a_1, a_3 作任何额外的限制, 所以改进后的方法不依赖于 a_1, a_3 的选择. 这就意味着使用现在的方法进行序列预测时, 蛋白质分子残基间接触势只要满足基本物理规律, 可自由选取. 对于任意一个蛋白质分子, 只要知道其晶体结构坐标, $\overline{A}$ 可由(17)式直接求出. 这有利于增强该设计方法的可用性. 比较(17)和(9)式发现, $\overline{A} \approx \varepsilon\overline{A^1}$, 其中 ε 为

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

由于 $\lambda < 1$, 根据 U(1, 1) < U(1, −1) < U(−1, −1) ≤ 0, 可知 $a_1 < 0$ 且 $(a_1 + a_3) < 0$, 所以 $\varepsilon > 1$. 也就是说, 文献[14]用 $\overline{A^1}$ 作为 $\langle A(r_i - r_j)\rangle_0$ 的近似值, 而此值偏小. 当 $|a_1| >> |a_3|$ 时, $\varepsilon \approx 1$, 此时有 $\overline{A} \approx \overline{A^1}$. 显然, 当 $|a_1| >> |a_3|$ 时, 本工作求出的 $\langle A(r_i - r_j)\rangle_0$ 的近似值与文献[14]所得结果一致. 所以文献[14]的方法可看作为本工作的特例.

2　结果与讨论

根据以上算法, 我们编制了新的蛋白质序列预测程序. 在所有的测试运行中, 具体参数作如下选取: η = 0.2, T = 1, $a_1 = a_2 = -5$, $a_3 = -2.2$. 我们沿用了 Sun 等人[7]对 20 种氨基酸的分类方法, 即把 A, V, L, I, C, M, F, Y, W 归为疏水残基, 把余下的氨基酸 G, P, H, S, T, K, R, D, N, E, Q 归为亲水残基. 测试时随机选取初始序列, 迭代收敛后与真实数据比较. 定义预测成功率为预测正确的残基总数与蛋白质残基总数的比值.

我们把蛋白质分为 4 种结构类, 具体分类方法[22]如下: 全α 蛋白, 即α 螺旋含量大于 40%、β 折叠含量

小于 5%的蛋白; 全β蛋白, 即α螺旋含量小于 5%、β折叠含量大于 40%的蛋白; α + β蛋白, 即α螺旋含量大于 15%、β折叠含量大于 15%且α螺旋和β折叠相间交叉排布的蛋白; α/β蛋白, 即α螺旋含量大于 15%、β折叠含量大于 15%且α螺旋和β折叠分隔排布的蛋白. 选取了 60 个蛋白质进行测试, 所有蛋白质的同源性小于50%. 其中α类和β类各 20 个, 其余两类各 10 个. 所有蛋白质的晶体坐标均来自 PDB (Protein Data Bank)[23]. 我们统计出四类蛋白质的疏水残基(H)与亲水残基(P)数目的比值λ依次为: 0.727, 0.646, 0.646, 0.733, λ的平均值为 0.688.

表 1 所选四种结构类蛋白质的 PDB 代码、残基数目及预测成功率

编号	α			β		
	PDB	残基数	成功率/%	PDB	残基数	成功率/%
1	1a1w	83	75.9	1a3k	137	83.2
2	1a32	85	75.3	1aly	146	71.9
3	1ad6	185	73	1cd8	114	72.8
4	1aep	153	78.4	1cdy	178	83.7
5	1bd8	156	70.5	1czs	160	70.6
6	1buy	166	72.3	1dfx	125	68.8
7	1ddf	127	74.8	1f53	84	79.8
8	1ed1	114	81.6	1fna	91	79.1
9	1uxc	50	82	1fnl	173	74
10	1fk5	93	78.5	1fsc	61	63.9
11	1neq	74	75.7	1g43	160	77.5
12	1hqb	80	78.8	1noa	113	69
13	1hyp	75	76	1tnm	91	78
14	1jf2	190	75.3	1tul	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
平均成功率/%			75.3			74.3

编号	α+β			α/β		
	PDB	残基数	成功率/%	PDB	残基数	成功率/%
1	1bm8	99	78.8	1acf	125	74.4
2	1bta	89	76.4	1ahn	169	71
3	1c9x	124	71	1aiu	105	81.9
4	1d8z	89	74.2	1b1i	122	75.4
5	1ddw	63	63.5	1bfe	110	75.5
6	1ek8	185	82.2	1byr	152	71.1
7	1ekg	119	77.3	1cjw	166	72.3
8	1ew4	106	81.1	1e0s	173	78.6
9	1f7w	144	72.2	1eq6	189	70.4
10	1gd3	98	75.5	1ezk	149	78.5
平均成功率/%			75.2			74.9

首先, 对不同结构类的蛋白质采用各自对应的λ进行预测. 测试结果见表 1. 分析结果发现, 四类蛋白质的预测成功率的算术平均值按照α, β, α + β, α/β的顺序依次为: 75.3%, 74.3%, 75.2%, 74.9%, 取得了比较好的结果, 证明该方法是可行的. 对于四类蛋白质, 我们尚有 25% 的残基识别失败. Micheletti 等人[11]的研究表明, 一个重要原因是 HP 模型过于简单. 即把 20 种氨基酸仅仅分成两类过于粗糙. 另外, 平均接触强度$\langle A(r_i - r_j)\rangle_0$的取值对预测精度有较大影响. 我们作近似求解时, 曾假设$A(r_i^a - r_j^a) \approx 1$时$\overline{S}_j = 1$, 即认为只有疏水残基之间的接触强度才接近 1. 实际情况中, 也有少数非疏水残基之间的接触强度较大. 所以这一近似也会引入误差.

随后, 对所有蛋白质使用同一个λ进行预测. 进行这个测试的目的, 是想知道对所有蛋白质, 能否给出同一个λ ($\lambda = 0.688$). 四类蛋白质的预测成功率的算术平均值按照α, β, α + β, α/β的顺序依次为: 75.2%, 74.3%, 75.0%, 74.8%.成功率无明显的降低. 所以在使用 HP 模型时, 对所有蛋白质也可以使用同一个λ进行预测.

为了与文献[14]的结果比较, 我们又选用了该文曾测试用的 20 个蛋白质, 预测结果见表 2. 我们的预

表 2 对文献[14]所用 20 个蛋白质的预测结果

编号	PDB 代码	残基数	成功率/%	成功率/%
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
平均成功率/%			73.1	72.5

测成功率为 73.1%, 优于文献[14]的预测成功率 72.5%. 这说明本工作不仅使该方法可以普遍使用, 而且使预测精度也有所提高.

在选用了几组 a_1, a_3 测试后发现, 当 a_1, a_3 取不同值时, 预测成功率无明显的变化. 但是有些 a_1, a_3 会使迭代收敛速度变慢.

3 结论

本工作对文献[14]提出的基于相对熵的蛋白质设计方法做了扩展, 提出了具有普遍性的算法. 只有当蛋白质的残基间接触势符合特殊要求以后, 文献[14]提出的方法才能适用. 本工作把该方法发展成为残基间接触势可自由选取. 从而拓宽了该方法的适用范围, 同时也提高了预测精度. 文献[14]提出的基于相对熵的蛋白质设计方法可被看作本工作的特例.

在研究中我们还发现, 平均接触强度 $\langle A(r_i - r_j)\rangle_0$ 的具体取值对预测精度有较大影响. 更精确的估计值会提高预测成功率. 我们正在尝试新的 $\langle A(r_i - r_j)\rangle_0$ 近似算法. 我们选取了简单的 HP 模型来检验方法, 如何把该方法发展到能预测 20 种氨基酸残基的非格子模型还有待进一步研究, 这方面的工作将在后续工作中进行.

致谢 感谢刘春莉同学在检测程序编制和结果分析上所给予的帮助. 本工作为国家自然科学基金(批准号: 19774501, 30170230)、北京市自然科学基金(批准号: 5032002)资助项目.

参 考 文 献

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 刘赟, 王宝翰, 王存新, 等. 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
15 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
16 Lu B Z, Wang B H, Chen W Z, et al. A New Computational Approach for Real Protein Folding Prediction. Protein Eng, 2003, 16(9): 659~663
17 Lau K F and 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 and Dill K A. Origins of structure in globular proteins. Proc Natl Acad Sci USA, 1990, 87: 6388~6392
19 Chan H S and 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

(2003-08-11 收稿, 2003-12-02 收修改稿)