PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(9): 1070~1076
www.pibb.ac.cn

基于 HNP 模型及相对熵的蛋白质设计方法在不同蛋白质体系中的应用 *

齐立省[1)] 苏计国[1,2)] 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 100124；[2]燕山大学理学院，秦皇岛 066004)

摘要 详细考察了基于 HNP(H: hydrophobic, N: neutral , P: hydrophilic)模型及相对熵的蛋白质设计方法对于不同结构类型蛋白质的适用性，并与基于 HP 模型的结果进行了比较．通过对 190 个 4 种不同结构类型的蛋白质进行预测，结果表明，基于 HNP 模型及相对熵的设计方法对于不同结构类型的蛋白质具有普适性．进一步的研究发现，对于 α 螺旋、β 折叠等规则的二级结构，该方法的预测成功率高于无规卷曲结构预测成功率．另外，还比较了对不同氨基酸的预测差异，结果显示亲水残基的预测成功率较高．此外，研究表明该方法对于蛋白质保守残基的预测成功率高于非保守残基．在以上分析的基础上，进一步讨论了导致这些差异的原因．这些研究为基于相对熵的蛋白质设计方法的实际应用和进一步的发展打下了良好基础．

关键词 蛋白质设计，相对熵，结构类型，氨基酸简化模型
学科分类号 Q6

在已知蛋白质三级结构的条件下，如何找到满足该三级结构的一级序列，是分子生物学的一个重要问题，被称为蛋白质设计亦称蛋白质逆折叠．蛋白质设计在分子设计与蛋白质工程等领域有广泛的应用前景．目前从理论上已提出了一些蛋白质设计方法，Shakhnovich 和 Gutin(SG)[1~3]以体系的哈密顿量作为优化目标函数，在格子和非格子模型上通过蒙特卡罗方法(MC)搜寻适用于目标结构的能量最低的序列．但 SG 方法在优化过程中需要对疏水与亲水残基的比例进行限制，否则找到的序列很快收敛为非极性的均聚体．进一步的研究表明，即使做出了这种限制，找到的序列仍有可能在其他非目标结构中拥有更低的能量[4]．Deutsch 和 Kurosky (DK)[5]考虑了体系的自由能，发展了求自由能的累积(cumulant)展开的一阶近似的设计方法，并在格子模型上用模拟退火方法进行了测试．DK 方法不需要对残基比例做出限制，是优于 SG 方法的，但由于体系的熵很难求出，该方法也遇到很大困难．Seno 等[6]利用双重 MC 方法对构象空间和序列空间同时进行搜索，并成功应用于格子模型上．虽然该方法原则上比上述方法精确，但在双重空间上的搜索需要耗费大量的 CPU 时间，很难应用于较大体系，也难应用于非格子模型的真实蛋白质结构．考虑到上述方法在计算体系自由能时存在的困难，我们提出了基于相对熵的蛋白质设计方法[7~11]，用相对熵代替传统的体系能量作为优化目标函数，从而克服了 SG 方法直接优化哈密顿量所遇到的困难．该方法完全基于物理学原理，其实质是在 Boltzmann 分布的序列空间中搜寻一条使给定结构具有最大占有率的序列．这一方法从理论上可以作为研究蛋白质折叠与设计的统一框架，已经成功地应用于真实蛋白质折叠与设计的研究中，并取得了较好的结果[12~14]．基于相对熵的方法是对能量优化的改进，更接近于从自由能的角度考虑体系的优化，并具有势函数简单且收敛速度快的优点．

用简化的氨基酸组合来表示蛋白质序列，是简化表述蛋白质序列和缩小序列复杂度的一种有效方

* 国家自然科学基金(10574009, 30670497)和北京市自然科学基金(5072002)资助项目．
** 通讯联系人．
Tel: 010-67392724, E-mail: cxwang@bjut.edu.cn
收稿日期：2008-01-21，接受日期：2008-04-15

法．为了简化对序列空间的搜索，首先把基于相对熵的蛋白质设计方法应用于最简单的氨基酸两态分类的 HP 模型上[7]，取得了成功，预测平均成功率达到 73%，与国际同类工作的结果相近．进一步的研究表明，基于 HP 模型和相对熵的设计方法普遍适用于不同结构类型的蛋白质序列设计[10]．但蛋白质设计的研究表明，在很多情况下简单的 HP 模型存在一些问题，Shakhnovich[15]指出，当蛋白质设计中考虑更多的氨基酸类型时，这种情形会得到改善．为此我们把基于相对熵的蛋白质设计方法推广到氨基酸三态分类的 HNP 模型上[11]，对几个真实蛋白质进行了测试，预测的平均成功率(40%～55%)与 Rossi 等[16]的结果相当，达到了国际同类工作的水平，表明该方法是切实可行的．本文在此工作的基础上，进一步考查了基于 HNP 模型及相对熵的设计方法是否对不同结构类型的蛋白质也具有普适性．该方法的目标是在氨基酸简化分类工作的基础上，最终发展为特定的目标结构设计出含 20 种残基的序列．氨基酸分类的细化对预测精度有多大的影响，不同残基对预测成功率的差异，以及基于不同的模型对这种差异有什么影响，不同的残基倾向于被预测成哪种残基类型，这些都是值得探讨的问题．蛋白质进化的过程显示，某些位置的残基相对于其他位置的残基具有较高的保守性，保守残基对维系蛋白质的结构和功能有重要意义．对于蛋白质设计，功能性残基的设计是关键的，该方法能否很好地预测功能性位点的残基类型，这将对蛋白质设计具有重要的实用价值．

为此，本文详细考察了基于 HNP 模型和相对熵的蛋白质设计方法，对不同折叠类型的蛋白质以及蛋白质中不同二级结构的适用情况，讨论了各种氨基酸类型对预测成功率的影响，并与基于 HP 模型方法的预测结果做了比较，进一步分析了造成这种差异的原因．另外，分析了不同残基预测结果中 3 种残基类型出现几率的差异性．最后还讨论了该方法对保守性残基和非保守性残基的预测情况．这些研究为该方法的应用和进一步的发展打下了良好的基础．

1　理论和方法

设蛋白质分子的序列空间为 $S=(s_1, s_2, \cdots, s_n)$，构象空间为 $r=\{r_1, r_2, \cdots, r_n\}$，其中，$r_i$ 是第 i 个残基的 C_α 原子的位置坐标，s_i 是第 i 个残基的类型．对于蛋白质设计，相对熵定义为[7]：

$$G(s)=\sum_r P_\alpha \ln(P_\alpha/P_0) \tag{1}$$

其中下标 α 表示目标结构．对于序列$\{s\}$，$P_0(r, s)$表示分子具有构象$\{r\}$的几率，可写为：

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

其中 $Z_0=\sum_{\{r\}} e^{-\beta H(r, s)}$，$P_\alpha$ 为分子具有特定构象$\{r^\alpha\}$的几率，可表示为：

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

其中 $Z_\alpha=\sum_{\{r_i\}} e^{-\beta H(r, s)}\prod_i \delta_{r_i, r_i^\alpha}$．容易证明，相对熵 $G\geqslant 0$ 且 $P_0=P_\alpha$ 时 G 有最小值 0[7]．相对熵 G 表示这两个几率函数相差的测度，它的极小化就相当于 P_0 的极大化．对于给定的目标结构$\{r_i^\alpha\}$，可以通过最陡下降法优化 G 搜寻最佳序列$\{s_i\}$．

蛋白质分子体系的哈密顿量采用简单的二体相互作用势，即：

$$H(r, s)=\frac{1}{2}\sum_{i,j\neq i}^{N} U(s_i, s_j)A(r_i-r_j) \tag{4}$$

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；$U(s_i, s_j)$为残基 s_i 和 s_j 之间的接触势；$A(r_i-r_j)$为接触强度函数，它与 $U(s_i, s_j)$一起表示残基间相互作用的强度和范围；r_i 为第 i 个残基的坐标．用最陡下降法优化相对熵可得数值迭代公式为[7]：

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

其中上标 k 表示第 k 步迭代，η 是介于 0 和 1 之间的可调参数，用来控制迭代的收敛速度，$\beta=1/RT$，T 为绝对温度，R 为普适气体常数，$\langle A(r_i-r_j)\rangle_0$ 表示接触强度在 P_0 分布下的系综平均值，r_i^α 为蛋白质目标结构 α 中第 i 个残基的坐标．

在本工作中，采用了氨基酸的简化分类模型．按照 Wang 等[17]的分类方法把 20 种氨基酸残基简化为 3 类：疏水残基(hydrophobic, H)、亲水残基(hydrophilic, P)及中性残基(neutral, N)，即 HNP 模型．疏水残基包括 Cys、Phe、Tyr、Trp、Met、Leu、Ile 和 Val，中性残基包括 Gly、Pro、Ala、Thr 和 Ser，亲水残基包括 Asn、His、Gln、Glu、Asp、Arg 和 Lys．残基之间的接触势采用 MJ 矩阵[18~20]，为了使接触势具有解析的形式，接触势 $U(s_i, s_j)$可展开为：

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

其中，定义 s_i=1, 0, –1 分别代表疏水、中性和亲水残基，a_{ij} 为待定常数，由接触势的具体数值来确定．按照氨基酸的分组情况，把 MJ 矩阵的行、列进行调整，使得 H 类、N 类、P 类氨基酸分别调整在一起，在此基础上，可以把矩阵划分为 9 块，分别描述 H-H、H-N、H-P、N-H、N-N、N-P、P-H、P-N、P-P 相互作用，计算各块的算术平均值，可得到一个 3×3 的矩阵，近似认为该矩阵可表示 3 种类型氨基酸之间的相互作用势，即：

	H	N	P
H	–5.733	–3.639	–3.128
N	–3.639	–2.010	–1.669
P	–3.218	–1.669	–1.651

(7)

从 MJ 分块矩阵的算术平均值得到的势函数，满足关系[18, 19]：

$$U(i, i)+U(j, j) < 2U(i, j) \quad (8)$$

所以可以认为这一势函数是合理的．将(7)式中势函数的数值代入(6)式可得 a_{ij} 的值为：

$$a_{ij}=\begin{pmatrix} -2.010 & -0.985 & -0.644 \\ -0.985 & -0.237 & -0.035\,5 \\ -0.644 & -0.035\,5 & -0.157 \end{pmatrix} \quad (9)$$

在本文中，接触强度函数表达式为[11]：

$$A(r_i-r_j)=\begin{cases} (1+e^{r_{ij}-0.65})^{-1} & r_{ij} \leqslant 1(nm), \text{且 abs}(i-j)>1 \\ 0 & \text{其他情况} \end{cases} \quad (10)$$

接触强度系综平均值的计算采用迭代算法，由公式(5)得：

$$\frac{\mathrm{d}s_i}{\mathrm{d}t}=-\eta\beta\left\{\sum_{j\neq i}[A(r_i^{\alpha}-r_j^{\alpha})-\langle A(r_{ij})\rangle_0]\frac{\partial U(s_i s_j)}{\partial s_i}\right\} \quad (11)$$

当迭代收敛时，(11)式的值为 0，可以得到[11]：

$$\langle A(r_{ij})\rangle_{os}=\frac{1}{N}\sum_i\sum_{j\neq i}A(r_{ij}^{\alpha})\frac{\partial U/\partial s_i}{\sum_{j\neq i}\partial U/\partial s_i} \quad (12)$$

其中脚标 S 表示 $\langle A(r_{ij})\rangle_o$ 的值是与序列相关的．在第 k 次迭代中，由公式(5)计算得到的 s_i^{k+1}的值，可能不会恰好为整数±1 或 0，为了转换到整数，设定如下规则：

$$S_i^{k+1}=\begin{cases} +1 & S_i^{k+1} > 0.5 \\ 0 & \text{abs}(S_i^{k+1}) \leqslant 0.5 \\ -1 & S_i^{k+1} < -0.5 \end{cases} \quad (13)$$

从蛋白质数据库(PDB)中选取了 190 个不同结构类型的单链蛋白质，残基数介于 72 ～ 315，所有蛋白质的同源性小于 50%，其中包括 50 个全 α 型蛋白质，55 个全 β 型蛋白质，40 个 α + β 型蛋白质，45 个 α / β型蛋白质．采用非格点模型，蛋白质分子简化为由一组节点组成，每个节点代表一个氨基酸残基，它的坐标采用 Micheletti 等[21]的方法，即把在 C_α 与 C_β连线上且距离 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 残基没有 C_β原子，就用 C_α原子的坐标位置表示．对每个蛋白质的氨基酸序列，按照 HNP 模型的分类，将其转化为一个由±1，0 组成的序列作为标准序列．从任意初始序列开始进行迭代计算，当两次迭代计算得到的序列没有变化时即可认为迭代收敛了，将计算得到的序列与标准序列进行比较，可以得到预测的成功率(预测正确的残基数与蛋白质链长的百分比)．利用最陡下降法优化目标函数很容易陷入局部极小，预测结果依赖于初始数值的选取，为了消除这种影响，能够在更大空间中采样，因此对每个蛋白质都进行了独立的 5 万次计算，把平均成功率作为最后的结果．

2 结果与讨论

把基于 HNP 模型与相对熵的蛋白质设计方法应用于 4 种结构类型的蛋白质，190 个蛋白质的预测成功率分布如表 1．对于全 α型、全β型、α + β型及α / β型蛋白质的预测平均成功率依次为 46.6%、45.5%、47.6%、45.1%．预测结果与国际上同类工作相近[16]，表明该方法可用于不同结构类型的蛋白质序列设计，具有很好的普适性．这说明氨基酸中性态的引入并不改变该方法普遍适用于不同结构类型蛋白质的特点．值得注意的是，随着中性态的引入，蛋白质设计的难度大大提高了，直接反映为相对于 HP 模型，其预测成功率降低了．但是考虑到对于 HNP 模型，一个随机设计的序列的成功率为 33%，应用基于相对熵与 HNP 模型的设计方法，与随机序列相比，成功率提高了大约 15%，说明该方法对于蛋白质设计是有效的，并且该方法的计算速度比较快．

Table 1 The distribution of success rate for all the proteins

Success rate /%	Number of proteins
0 ～ 40	20
40 ～ 45	65
45 ～ 50	73
50 ～ 55	23
55 ～ 65	9

进一步分析了不同二级结构中残基的预测情况．利用 Dssp 软件[22]，得到所有蛋白质中二级结构的分布情况，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)．其中 H 和 E 的预测成功率超过了 60%，远远高于对整个蛋白质的平均预测成功率，刘等[10]的研究中也表明对二者有很高的预测成功率．说明该方法对于 α-Helix、β-sheet 等规则的二级结构具有较高的预测成功率，而对于 Turn、Bend、无规卷曲等结构的预测成功率较低，这可能与无规卷曲的序列具有很高的序列简并性有关．如何提高无规卷曲序列的预测成功率是下一步工作中要考虑的问题．

比较不同氨基酸类型的预测成功率，最高的是亲水残基(55.3%)，次之为疏水残基(48.3%)，最低的是中性残基(37.1%)．进一步分析了 20 种残基对预测成功率的影响，发现不同的残基也有很大差异．残基预测成功率最高的依次是 G、R、E、N、K、D(图 1a)，与刘赟等[10]的预测结果(H、K、G、Q、Y、E)相比较，表明无论是采用 HP 模型还是 HNP 模型，基于相对熵的蛋白质设计方法都是对亲水残基的预测成功率较高．中性残基除了 G 外，预测成功率偏低．考虑到不同残基在蛋白质中的含量不同，含量高的残基的预测情况对平均成功率影响较大．把不同残基的成功率乘以相应残基在蛋白质中的含量百分比，发现残基 G、L、K、E、V、D 的预测成功率偏高(图 1b)，大部分疏水残基的预测成功率偏低，分析其原因，疏水相互作用倾向于把所有的疏水残基都集聚在蛋白质的内部，但是为了发挥特定的生物学功能，某些疏水残基在真实情况下可能处在蛋白质的表面，对于这些残基，我们的方法常常预测错误，这也是该方法有待于提高的地方．但就不同残基类型的预测成功率整体而言，中性残基的预测成功率偏低．因此基于相对熵的蛋白质设计方法如何体现不同残基的性质，以及是否有更好的残基分类方式能提高预测的成功率，都是值得考虑的问题．

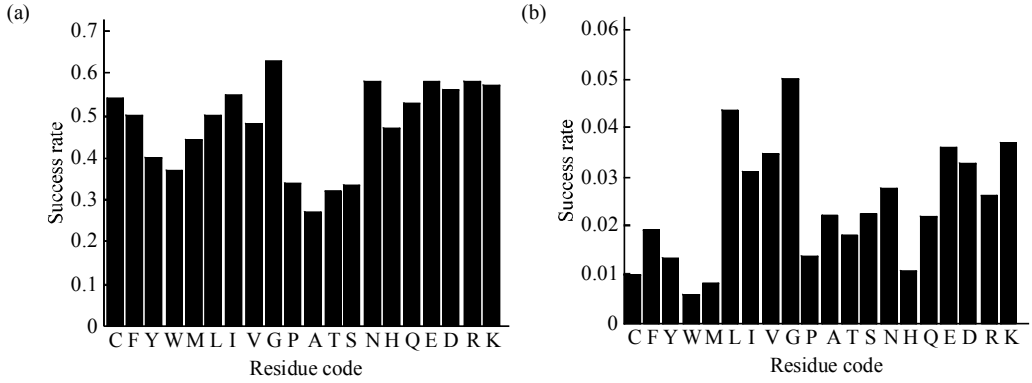

Fig. 1　Histogram of the success rate

(a) Success rate in identifying the correct HNP class of the 20 amino acids. (b) Success rate obtained through multiplying success rates of (a) by the fraction of each residue type in all proteins.

进一步考察不同残基被预测成 3 种残基类型可能性的差异．图 2 的分析结果显示，预测成疏水残基几率较高的残基，恰好为按照 Wang 等[17]分类标准所划定的 8 个疏水残基．在中性和亲水残基的预测结果中出现疏水残基的几率很低，5 个中性残基倾向于被预测成亲水残基，且 7 个亲水残基被预测成亲水残基的几率较大，高出疏水残基被预测成疏水残基的平均几率约 10%．不同残基被预测成中性残基的几率大概介于 20%～30%，即差异很小，表明基于相对熵的蛋白质设计方法对亲水残基的预测能力要强于对疏水和中性残基的预测．

关于蛋白质序列中保守残基的预测情况，选取 PDB 代码为 1JNP、1AEP 和 1APS 的蛋白质以及它们的同源蛋白作为研究对象，利用 ClustalX 1.83[23]对这 3 个蛋白质进行多重序列比对，分别得到 28 个、23 个和 10 个位点的残基是高度保守的．从任意初始序列开始进行迭代，对每个蛋白质都进行了独立的 5 万次计算，分别计算了保守残基和非保守残基的预测成功率，把平均成功率作为最后的结果．结果表明保守残基的预测成功率达到 52.4%，高于非保守残基的预测成功率(48.5%)．蛋白质功能性残基在进化上具有高度的保守性，而非保守位点的残基具有较高的简并性，因此对保守残基的预测成功率应该大于非保守残基，我们的模型反映了

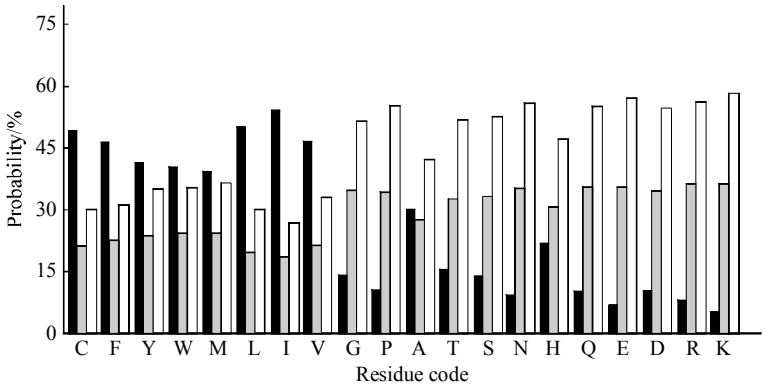

Fig. 2 Probability of HNP class in the prediction result of each amino acid

■: H; ▨: N; □: P.

这一特性．保守残基对于维系蛋白质结构和功能起着关键的作用，保守位点的正确设计是整个蛋白质设计的核心，本文研究表明，基于相对熵的蛋白质设计方法对于保守残基的设计成功率高于非保守残基，这对于该方法的实际应用具有重要的意义．

需要指出的是，在我们的工作中，势函数的准确性直接决定了结果的可靠程度．本文选取了统计势 MJ 矩阵来描述残基之间的相互作用，它很大程度上已经包含了蛋白质体系相互作用的主要特征，并突出反映了疏水效应在蛋白质折叠中的重要作用．而蛋白质折叠的主要驱动力是疏水相互作用，因而 MJ 矩阵包含了蛋白质折叠的主要机制．研究发现，该简化模型和 MJ 矩阵能够有效地用于蛋白质折叠和蛋白质设计研究，揭示了蛋白质折叠和蛋白质设计的一些普遍的规律，取得了好的结果[7~14]．但是，实际上蛋白质折叠是很多精细的相互作用协调作用的结果，包括共价键、静电、范德华、氢键、疏水等作用．MJ 矩阵由于其统计特性，只能反映蛋白质折叠主要的普遍特征，并不能很精确地反映这些特异作用的信息，因而有必要把现在的模型进一步细化，并采用更加精细的相互作用势，以获得更加精确的预测结果．当然，模型越精细，计算量就越大，如何协调模型精确性和计算量是一个值得研究的问题．在我们的后续工作中将会对这些问题作进一步的探讨．

3 结　论

本文在已有工作的基础上，考察了基于 HNP 模型与相对熵的蛋白质设计方法对不同结构类型蛋白质的普适性．选取 4 种不同结构类型的 190 个蛋白质进行了预测，得到的平均预测成功率与当前国际同类工作的结果相当，且发现，该方法对不同的结构类型无明显的偏好性，表明该方法适用于不同结构类型的蛋白质序列设计．通过对蛋白质二级结构中残基的预测发现，规则的二级结构比无规卷曲的结构预测成功率偏高．进一步分析了不同残基对预测成功率的影响以及不同残基预测结果中 3 种残基类型出现几率的差异性，都表明基于相对熵的蛋白质设计方法对亲水残基的预测成功率较高．此外，该方法对蛋白质中保守残基的预测成功率要高于非保守残基．

基于相对熵的蛋白质设计方法完全基于物理原理，用相对熵为优化目标函数，克服了单纯优化哈密顿量带来的困难，解决了同时在序列空间和构象空间搜索带来的计算量大的难题．研究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．但需要指出的是，该方法预测的成功率在很大程度上依赖于$<A(r_i-r_j)>_0$的计算，有必要尝试更精确的取值，以提高预测的成功率．另外该方法对各种残基之间相互作用势的计算是粗糙的，这也是一个需要改进的方面．如何把该方法发展到含有更多种氨基酸类型的非格点模型，还有待进一步的探讨．

致谢 感谢中国科学院生物物理研究所王宝翰老师对本工作给予的指导．

参 考 文 献

1 Shakhnovich E I, Gutin A M. A new approach to the design of stable proteins. Protein Engineering, 1993, **6** (8): 793～800

2 Shakhnovich E I. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett, 1994, **72** (24): 3907～3910

3 Shakhnovich E I. Protein design: a perspective from simple tractable models. Folding & Design, 1998, **3** (3): R45～R58

4 Micheletti C, Maritan A, Banavar J R. A comparative study of existing and new design techniques for protein models. J Chem Phys, 1999, **110** (19): 9730～9738

5 Deutsch J M, Kurosky T. New algorithm for protein design. Phys Rev Lett, 1996, **76** (10): 323～326

6 Seno F, Vendruscolo M, Maritan A, *et al*. Optimal protein design procedure. Phys Rev Lett, 1996, **77** (9): 1901～1904

7 Liu Y, Wang B H, Wang C X, *et al*. A new approach for protein design based on the relative entropy. Science in China, 2003, **46** (6): 659～669

8 刘　赟, 王存新, 王宝翰, 等. 基于格子模型的蛋白质设计方法. 生物化学与生物物理进展, 2004, **31** (2): 172～175
Liu Y, Wang C X, Wang B H, *et al*. Prog Biochem Biophys, 2004, **31** (2): 172～175

9 Wang Y H, Wang B H, Liu Y, *et al*. A generalized approach for protein design based on the relative entropy. Chinese Science Bulletin, 2004, **49** (5): 426～431

10 刘　赟, 王忆华, 王宝翰, 等. 四种结构类型的蛋白质设计方法. 生物物理学报, 2004, **20** (4): 307～314
Liu Y, Wang Y H, Wang B H, *et al*. Acta Biophysica Sinica, 2004, **20** (4): 307～314

11 Jiao X, Wang B H, Su J G, *et al*. Protein design based on the relative entropy. Phys Rev E, 2006, **73** (6): 061903

12 Lu B Z, Wang B H, Chen W Z, *et al*. A new computational approach for real protein folding prediction. Protein Engineering, 2003, **16** (9): 659～663

13 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
Lu B Z, Wang C X, Wang B H. Chin J Chem Physics, 2003, **16** (2): 117～121

14 苏计国, 王宝翰, 焦　雄, 等. 基于 HNP 三态模型及相对熵方法的蛋白质折叠研究. 生物化学与生物物理进展, 2006, **33** (5): 479～484
Su J G, Wang B H, Jiao X, *et al*. Prog Biochem Biophys, 2006, **33** (5): 479～484

15 Shakhnovich E I. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett, 1994, **72** (24): 3907～3910

16 Rossi A, Micheletti C, Seno F, *et al*. A self-consistent knowledge-based approach to protein design. Biophysical Journal, 2001, **80** (1): 480～490

17 Wang J, Wang W. Modeling study on the validity of a possibly simplified representation of proteins. Phys Rev E, 2000, **61** (6): 6981～6986

18 Miyazawa S, Jernigan R L. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 1985, **18** (3): 524～552

19 Maiorov V N, Crippen G M. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol, 1992, **227** (4): 876～888

20 Miyazawa S, Jernigan R L. Residue-residue potentials with a favorable contact pair term and unfavorable high packing density term, for simulation and threading. J Mol Biol, 1996, **256** (3): 623～644

21 Micheletti C, Seno F, Maritan A, *et al*. Design of proteins with hydrophobic and polar amino acids. Proteins, 1998, **32** (1): 80～87

22 Kabsch W, Sander C. Dictionary of protein secondary structure: Patter recognition of hydrogen-bonded and geometrical features. Biopolymers, 1983, **22** (12): 2577～637

23 Thompson J D, Gibson T J, Plewniak F, *et al*. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 1997, **25** (24): 4876～4882

Application of Protein Design Method Based on The HNP Model and Relative Entropy for Different Protein Systems*

QI Li-Sheng[1], SU Ji-Guo[1,2], CHEN Wei-Zu[1], WANG Cun-Xin[1]**

([1]*College of Life Science and Bioengineering, Beijing University of Technology, Beijing* 100124, China;
[2]*College of Science, Yanshan University, Qinhuangdao* 066004, China)

Abstract The application of the protein design method based on the HNP model and the relative entropy theory is discussed for four structural classes of real proteins, and the results are compared with that of the HP model. Testing on 190 proteins shows that this method is generally effective for the different structural classes of proteins. Further studies show that the success rate of this method on regular secondary structures is higher than that on the random coil. Additionally, the success rate for different types of amino acids is also analyzed. It is found that the success rate on the hydrophilic residues is higher than those of the other two types. Furthermore, the success rate of this method on the conserved residues is higher than the non-conserved residues. The reasons resulting in the difference of the success rate on different systems were also analyzed. All analyses mentioned above make the foundation for the development and the application of this method in the future.

Key words protein design, relative entropy, structural classes, simplified amino acid model

---

*This work was supported by grants from The National Science Foundation of China (10574009, 30670497) and Beijing Natural Science Foundation (5072002).

**Corresponding author.
Tel: 86-10-67392724, E-mail: cxwang@bjut.edu.cn

Received: January 21, 2008 Accepted: April 15, 2008