生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(5): 479~484
www.pibb.ac.cn

PIBB

基于 HNP 三态模型及相对熵方法的蛋白质折叠研究 *

苏计国 [1)] 王宝翰 [2)] 焦 雄 [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)

摘要 把 20 种氨基酸简化为 3 类：疏水氨基酸 (hydrophobic，H)、亲水氨基酸 (hydrophilic，P) 及中性氨基酸 (neutral，N)，每个氨基酸简化为一个点，用其 C_α 原子来代替. 采用非格点模型，以相对熵作为优化函数，进行蛋白质三维结构预测. 为了与基于相对熵方法的蛋白质设计工作进行统一，采用了新的接触强度函数. 选用蛋白质数据库中的天然蛋白质作为测试靶蛋白，结果表明，采用该模型和方法取得了较好的结果，预测结构相对于天然结构的均方根偏差在 0.30～0.70 nm 之间. 该工作为基于相对熵及 HNP 模型的蛋白质设计研究打下了基础.

关键词 氨基酸简化模型，相对熵，蛋白质折叠
学科分类号 Q615

从蛋白质一级序列预测其三维结构以及蛋白质折叠的机理是目前分子生物学研究的热点问题之一. 蛋白质结构预测的理论基础是 20 世纪 70 年代由 Anfinsen 提出的 Anfinsen 原理[1]. 目前蛋白质结构预测方法大致可分为两大类：一类是基于知识的预测方法，如同源模建[2~4]、Threading 方法[5,6]，以及这些方法与二级结构预测相结合的方法[7]等. 基于知识的蛋白质结构预测方法取得了比较好的结果，目前，同源模建方法被认为是进行蛋白质结构预测最为成功的方法[8,9]，Threading 方法也取得了不错的结果，对于小的蛋白质，均方根偏差的值在 0.30～ 0.75 nm 之间[6]. 另一类蛋白质结构预测方法是从头(ab initio) 折叠方法[10,11]，它是通过优化体系的能量函数来预测蛋白质结构，而不借助于任何其他的基于知识的信息. 该方法有助于理解各种物理的相互作用是如何导致蛋白质沿其能量曲面到达天然的三维结构的. 根据 Anfinsen 原理，蛋白质天然构象对应于体系自由能最低的状态[1]，但目前一般能量优化方法并没有完全考虑熵的效应，因此，预测结构并不对应自由能最小的态. 为此，我们小组提出了基于相对熵的蛋白质从头折叠方法，采用相对熵作为优化函数，来代替传统的体系能量，取得了好的结果[12,13]. 进而，把相对熵作为优化对象进行了蛋白质设计 (protein design) 的研究，也取得了比较满意的结果[14]. 该方法实质上是在按照 Boltzmann 分布的构象空间搜索最大可能的构象，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，它还具有势函数简单和计算快的优点.

蛋白质折叠的主要驱动力是疏水相互作用[15]，而构成蛋白质的 20 种氨基酸根据其亲疏水性质可简化为三类：疏水氨基酸(H)、亲水氨基酸(P)及中性氨基酸(N)，用简化的氨基酸组合来表示蛋白质序列，是缩小序列复杂度的一个有效方法[16]，在本工作中，采用该氨基酸简化模型以简化计算，那么，简单的疏水、中性、亲水 (HNP) 氨基酸三态模型是否能够反映蛋白质折叠的主要物理作用及折叠机制呢？本工作表明，该模型是切实可行的，预测精度与基于真实的 20 种氨基酸的结果相当. 该工作为蛋白质设计从两态 (HP) 模型发展到三态 (HNP) 模型提供了基础.

在文献[14]的蛋白质设计算法中需要计算接触强度在构象空间的系综平均值[14]，蛋白质折叠将为

* 国家自然科学基金 (10574009) 和国家教育部博士点基金 (20040005013)资助项目.

** 通讯联系人. Tel: 010-67392724, E-mail: cxwang@bjut.edu.cn

收稿日期：2005-11-28，接受日期：2006-01-27

该量的较精确计算提供依据，但在文献[12,13]中，由于蛋白质折叠和蛋白质设计所采用的简化模型和接触强度函数并不相同，因此无法利用蛋白质折叠的数据来计算接触强度在构象空间中的系综平均值，只采用了一个近似估计值[14]. 为了使蛋白质设计研究计算更加精确，很有必要把二者的模型和接触强度函数进行统一. 为此，在原工作的基础上，采用氨基酸简化模型，设计了一个新的适用于该模型的势函数，并且采用了新的接触强度函数. 用该方法对几个真实蛋白质作了测试，取得了较好的结果，预测构象相对于它们的天然构象的均方根偏差(RMSD) 在 0.30～0.75 nm 之间. 从而验证了该模型和方法的可行性，并为在该模型基础上的蛋白质设计研究奠定了基础.

1 理论与方法

设 $S=(s_1, s_2, \cdots, s_n)$ 表示蛋白质分子的序列空间，$\vec{r}=\{\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_n\}$表示构象空间，其中，$\vec{r}_i$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.

对于蛋白质折叠，相对熵定义为：

$$G(\{\vec{r}_i\})=\sum_{\{s_i\}} P_\alpha \ln (P_\alpha / P_0) \tag{1}$$

其中，P_0 为给定构象$\vec{r}=\{\vec{r}_i\}$，分子具有序列 $S=\{s_i\}$ 的几率，可表示为：

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

P_α 为给定构象$\vec{r}=\{\vec{r}_i\}$，分子具有天然序列 $S^\alpha=\{s_i^\alpha\}$的几率，可表示为：

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

(2)(3)式中的 $H(s,\vec{r})$为体系的哈密顿量.

容易证明，相对熵 $G(\{\vec{r}_i\})\geqslant 0$，且 $P_0=P_\alpha$ 时，G有最小值 0[14]. 给定序列 $S^\alpha=\{s_i^\alpha\}$，可以通过优化相对熵来寻找最佳结构，使得 P_0 接近 P_α. 采用最陡下降算法来优化相对熵，可得最陡下降公式：

$$\frac{\mathrm{d}\vec{r}_i}{\mathrm{d}t}=-\eta\frac{\mathrm{d}}{\mathrm{d}\vec{r}_i}G \tag{4}$$

其中，η 为可调参数，用来控制迭代收敛速度.

哈密顿量 H 采用接触势形式：

$$H=\frac{1}{2}\sum_{i,j\neq i} U(s_i, s_j)A(\vec{r}_i-\vec{r}_j) \tag{5}$$

其中，$U(s_i, s_j)$为残基 s_i 和 s_j 之间的接触势，$A(\vec{r}_i-\vec{r}_j)$为依赖于残基间距离的无量纲的接触强度函数. 代入最陡下降公式(5)得：

$$\vec{r}_i^{\,k+1}-\vec{r}_i^{\,k}=-\eta\beta\sum_{j\neq i}(U(s_i^\alpha, s_j^\alpha)-\langle U(s_i, s_j)\rangle_0)\frac{\partial}{\partial\vec{r}_i^{\,k}}A(\vec{r}_i^{\,k}-\vec{r}_j^{\,k}) \tag{6}$$

其中，k 为迭代次数，$\beta=1/RT$，T 为绝对温度，R 为普适气体常数；$\langle U(s_i, s_j)\rangle_0$ 为 $U(s_i, s_j)$在 P_0 分布下的系综平均值.

采用文献[17]的分类方法把 20 种氨基酸残基简化为三类：疏水残基(hydrophobic，H)、亲水残基(hydrophilic，P)及中性残基(neutral，N)，即 HNP 模型. 其中，疏水残基包括 Cys、Met、Phe、Ile、Leu、Val、Trp、Tyr，中性残基包括 Ala、Gly、Thr、Pro、Ser，亲水残基包括 Asn、Gln、Asp、Glu、Arg、Lys、His. 对于 HNP 模型，(5)式中的接触势 $U(s_i, s_j)$可以展开为以下形式：

$$U(s_i, s_j)=(1\ s_i\ s_i^2)\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}\begin{pmatrix} 1 \\ s_j \\ s_j^2 \end{pmatrix} \tag{7}$$

其中，s_i=1，−1 或 0，分别代表疏水残基、亲水残基和中性残基，a_{ij} 为常数，由接触势的具体数值来确定. 下面通过 M-J 矩阵[18~20]来确定接触势中的各常数的值.

根据氨基酸的分组情况，把 M-J 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，表示三种类型氨基酸之间的相互作用势，从而，得到(7)式中 a_{ij} 的值为：

$$\begin{array}{c|ccc} & H & N & P \\ H & -5.733 & -3.639 & -3.218 \\ N & -3.639 & -2.010 & -1.669 \\ P & -3.218 & -1.669 & -1.651 \end{array} \Rightarrow$$

$$a_{ij}=\begin{pmatrix} -2.010 & -0.985 & -0.644 \\ -0.985 & -0.237 & -0.035\,5 \\ -0.644 & -0.035\,5 & -0.157 \end{pmatrix} \tag{8}$$

为了使基于相对熵的蛋白质折叠方法和基于相对熵的蛋白质设计方法在接触强度上能统一起来，(5)式中接触强度函数 $A(\vec{r}_i-\vec{r}_j)$采用蛋白质设计工作中选用的表达形式[14]，为了能够适用于蛋白质折叠研究，对其作了适当修改，表达式如下：

$$A(\vec{r}_i-\vec{r}_j)=A_1+A_2=\alpha\frac{1}{1+\exp(r_{ij}-d)}+\varepsilon(-\frac{\sigma^6}{r_{ij}^6}+\frac{\sigma^{12}}{r_{ij}^{12}}) \quad (9)$$

其中，α、ε 和 σ 为可调参数，d 是一个残基接触距离附近的值，这里取 d=0.65 nm. 在实际模拟中，两个连续的残基之间的距离用 SHAKE 算法来约束，因而，任何两个连续的残基间的相互作用不计算. (9)式由两项组成，第一项 A_1 为蛋白质设计研究中所采用的形式[14]，它与接触势 $U(s_i, s_j)$一起考虑了蛋白质折叠的主要驱动力：疏水和亲水相互作用. 不同于蛋白质设计，在(9)式中接触强度表达式增加了第二项 A_2，目的是为防止一些残基紧密靠近，该项采用了一个类似于 van der Waals 势. 对于蛋白质折叠，附加这一项是必要的，在我们的测试中发现，如果去掉这一项，蛋白质折叠后的结构会聚成一团.

按照(9)式的接触强度公式编写程序并用真实蛋白质测试发现，应用此公式，去折叠的蛋白质无法重新正确折叠. 分析其原因，图 1 中点线显示了(9)式中第一项 A_1 表达式的函数曲线，从图 1 中可以看出，(9)式 A_1 的函数值在残基间距离 r 大于 1.00 nm 时趋近于 0. 而该项与 $U(s_i, s_j)$一起描述了残基间的疏水相互作用，当残基间距离 r 大于 1.00 nm 时，该项趋近于 0，这就意味着忽略了残基间长程相互作用对蛋白质折叠的影响，而残基间长程相互作用对蛋白质折叠又很重要，所以对于蛋白质折叠研究，采用(9)式是不太合理的. 为此，对(9)式做了一点修改，使该项的值随 r 的变化趋于缓和，修改后的接触强度形式如下：

$$A(\vec{r}_i-\vec{r}_j)=A_1+A_2=\alpha\frac{1}{1+\exp[(r_{ij}-d)/13]}+\varepsilon(-\frac{\sigma^6}{r_{ij}^6}+\frac{\sigma^{12}}{r_{ij}^{12}}) \quad (10)$$

(10)式相对于(9)式 A_1 项增加了参数 13，使得函数值随残基间距离的变化趋于缓和，如图 1 中三角线所示，该参数是通过真实蛋白质折叠测试优化得到的. 把修改后的接触强度函数(10)式重新用到蛋白质设计工作中，研究发现对结果精度影响很小.

$\langle U(s_i, s_j)\rangle_0$ 的计算采用卢本卓等[12,13]基于平均场理论提出的近似方法，认为蛋白质能够折叠到天然态构象的必要条件是它当前状态的能量必须小于或等于某个平均能量，可取为蛋白质刚好变性时的能量，因为蛋白质构象空间主要由去折叠态所占据. 而刚好变性时的能量可通过忽略所有长程相互作用，只考虑序列上相邻残基之间的能量得到：

$$\langle U(s_i, s_j)\rangle_0=k_m\cdot\frac{2}{N}\overline{U},\ \ \overline{U}=\frac{1}{N-1}\sum_i^{N-1}U(s_i^{\alpha}, s_{i+1}^{\alpha}) \quad (11)$$

其中，k_m 为可调参数；上标 α 表示为给定序列，N 为蛋白质分子的残基个数.

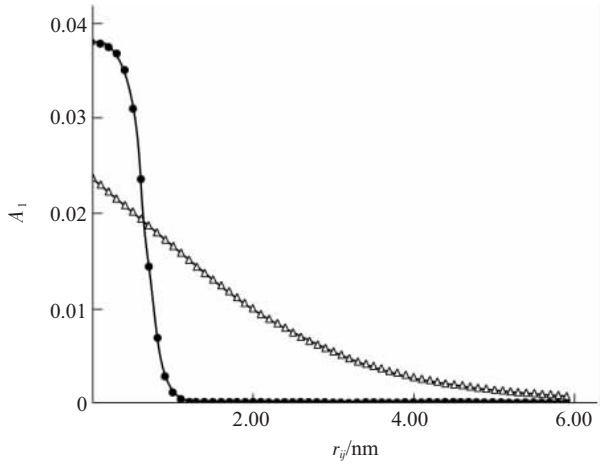

Fig. 1　The curve of the first term of the contact intensity function *vs* the distance between residues

2　结果与讨论

采用非格点模型，蛋白质分子简化为由一组节点组成，每一个节点代表一个氨基酸残基，它的坐标为相应残基的 C_a 原子的位置，并且把残基类型简化为三类，即疏水、中性和亲水残基. 根据上述算法，依据(6)式，编制了蛋白质折叠程序. 为了确定算法中的各个参数，我们首先把程序在牛胰岛素受体(BPTI)上作了测试并优化选取了各个可变参量，各参数的取值为：η = 0.13，T = 1，α = 0.038，ε = 2.080，σ = 0.380 nm，K_m= 7.2. SHAKE 约束的精度为 0.000 1 nm，当每个节点在连续两次迭代中的位置差都小于 0.000 1 nm 时，就认为迭代已收敛了. 若当初始结构的键长偏离要约束的长度太大时，就先采用一个谐振势进行键长约束，然后再使用 SHAKE 约束.

在蛋白质数据库(PDB)中选取了 14 个小蛋白质作为测试的靶蛋白，它们的代码分别是：1bpi、1ejg、1fcl、1ubq、1prb、1nmg、1ubi、2gb1、1a7f、1b17、1cq4、1kde、1stu、1e68. 其中，1a7f、1b17、1cq4 三个蛋白质有两条主链，其他蛋白质

为单链. 首先，把每个蛋白质充分去折叠成线团，二级结构被完全破坏(coil)后作为折叠过程的初始结构，然后用以上算法和程序进行折叠的模拟计算，得到了预测结构. 并计算了预测结构相对于天然结构的均方根偏差以及预测结构出现的残基天然接触数. 这些蛋白质的初始结构数据及预测结果列在表 1 中, 所得结果与最近报道的 ab initio 蛋白质折叠预测精度相近[21]. 对于文献[13]预测的目标蛋白用该方法也作了测试，二者结果的对比见表 2，从表 2 中可以看出，虽然采用了简化氨基酸模型，但其预测结果精度与文献[13]所得的结果精度相似，有些要好于文献[13]的结果. 以上结果表明，所采用的简化氨基酸模型和接触强度函数是可行的，为蛋白质设计的研究打下了基础.

与文献[12,13]一样，新的优化算法只使用了两两连续的 C_α 原子之间的距离信息，不使用其他目标蛋白的已知信息，如天然的二级结构、二硫键等. 图 2 显示了 PDB 代码分别为 1ejg 和 1a7f 的蛋白质天然结构与预测结构的 C_a 骨架图. 其中，1ejg 蛋白有一条主链，而 1a7f 蛋白有两条主链. 从图 2 中可以看出，折叠预测构象明显地恢复了天然结构中的二级结构及主链走向，天然结构中所含有的 α 螺旋和部分 β 折叠片能很好地在预测结构中显示出来，说明新的模型及算法是有效的. 结果还表明，该算法不仅仅适用于单链蛋白，同样能够很好地适用于双链蛋白的折叠预测.

Table 1 Result of the protein folding prediction

PDB code	Residue number	NCN of the native structure	RMSD of the initial structure/nm	NCN of the folded structure	RMSD of the folded structure/nm
1bpi	58	180	1.74	96	0.70
1ejg	46	144	1.46	81	0.49
1fcl	56	179	1.00	100	0.56
1ubq	76	229	1.56	134	0.56
1prb	53	154	1.34	94	0.38
1nmg	67	215	1.65	110	0.65
1ubi	76	230	1.65	126	0.58
2gbl	56	170	1.48	95	0.57
1a7f	50	141	1.72	87	0.39
1b17	51	161	1.30	75	0.42
1cq4	71	198	1.30	94	0.58
1kde	65	215	1.90	124	0.65
1stu	68	257	1.73	124	0.66
1e68	70	236	1.73	169	0.37

NCN is the native contact number. If the distance of two residues (the residue position is represent by the coordinate of C_α atom) is less than 0.75 nm, they are considered in contact. The value of RMSD is obtained from the structures *vs* native structures of the PDB.

Table 2 The result comparison between literature [13] and this work

PDB code	Residue number	Result of literature [13]		Result of this work	
		NCN of folded structure	RMSD of folded structure/nm	NCN of folded structure	RMSD of folded structure/nm
1bpi	58	92	0.68	96	0.70
1fcl	56	96	0.56	100	0.56
1ejg	46	76	0.52	81	0.49
1ubq	76	135	0.55	134	0.56

NCN is the native contact number.

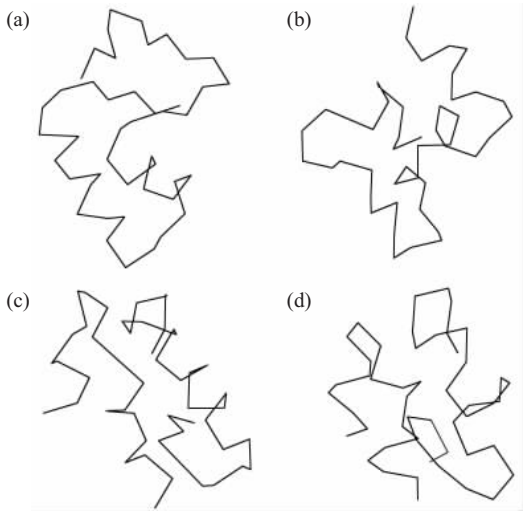

Fig. 2 The native structures *vs* the predicted structures of 1ejg and 1a7f

(a) The native structure of 1ejg. (b) The folded structure of 1ejg. (c) The native structure of 1a7f. (d) The folded structure of 1a7f.

3 结 论

基于相对熵的优化方法实质上是在满足波尔兹曼分布的构象空间中搜索能够匹配给定序列的构象，使得分布函数 P_0 趋近于 P_α，相对于体系的势能，我们的方法更接近于从自由能的角度研究蛋白质折叠问题. 基于相对熵的优化方法是完全基于物理原理的蛋白质从头折叠方法，本质上区别于基于知识的蛋白质预测方法，比如同源模建、Threading 方法等. 在我们的方法中，采用了残基简化模型和残基接触势 (M-J 矩阵)，它具有势函数简单、计算速度快的优点.

在本工作中，我们把 20 种氨基酸简化为疏水、亲水及中性氨基酸的基础上，采用基于相对熵的优化方法进行蛋白质从头折叠的研究，蛋白质折叠的主要驱动力是疏水相互作用，残基简化模型能够较好地反映氨基酸之间的疏水相互作用，基于该模型和相对熵的方法预测蛋白质结构取得了好的结果，预测精度与采用真实的 20 种氨基酸的方法结果相当，表明该简化模型是可行的. 本工作为基于相对熵的蛋白质设计从两态 (亲水、疏水) 模型扩展到三态 (亲水、疏水及中性) 模型提供了模型依据.

本文采用了新的接触强度函数，使得蛋白质折叠与蛋白质设计所采用的模型和接触强度函数统一起来，为利用蛋白质折叠数据进行蛋白质设计中接触强度的系综平均值的计算打下了基础.

参 考 文 献

1 Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, **181** (4096): 223～230

2 Warme P K, Momany F A, Rumball S V, *et al*. Computation of structures of homologous proteins. Alpha-lactalbumin from lysozyme. Biochemistry, 1974, **13** (4): 768～782

3 Johnson M S, Overington J P, Blundell T L. Alignment and searching for common protein folds using a data bank of structural templates. J Mol Biol, 1993, **231** (3): 735～752

4 Fischer D, Rice D W, Bowie J U, *et al*. Assigning amino acid sequences to 3D protein folds. FASEB J, 1996, **10** (1): 126～136

5 Sippl M J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structure. J Comput Aided Mol Des, 1993, **7** (4): 473～501

6 Venclocas C, Zemla A, Fedelis K, *et al*. Some measures of comparative performance in the three CASPs. Proteins, 1999, **37** (s3): 231～237

7 Skolnick J, Kolinski A, Ortiz A R. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol, 1997, **265** (2): 217～241

8 Hardin C, Pogorelov T V, Luthey-Schulten Z. Ab initioprotein structure prediction. Current Opinion in Structural Biology, 2002, **12** (2): 176～181

9 Sanchez R, Sali A. Comparative protein structure modeling in genomics. J Comput Phys, 1999, **151** (1): 388～401

10 Zhou Y Q, Karplus M. Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. J Mol Biol, 1999, **293** (4): 917～951

11 Huang E S, Samudrala R, Ponder J W. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol, 1999, **290** (1): 267～281

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 Lu B Z, Wang C X, Wang B H. A new minimization method for real protein folding prediction. Chin J Chem Phys, 2003, **16** (2): 117～121

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

15 Dill K A. Theory for the folding and stability of globular proteins. Biochemistry, 1985, **24** (6): 1501～1509

16 王 骏, 王炜. 蛋白质表示简化的物理研究. 物理学进展, 1997, **20** (3): 301～309
Wang J, Wang W. Progress in Physics, 1997, **20** (3): 301~309

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 interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 1985, **18** (3): 534～552

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 Bonneau R, Strauss C E M, Baker M. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins, 2001, **43** (1): 1～11

Protein Folding Study Based on The HNP Model and The Relative Entropy Approach*

SU Ji-Guo[1], WANG Bao-Han[2], JIAO Xiong[1], CHEN Wei-Zu[1], WANG Cun-Xin[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 Twenty kinds of amino acids are simplified into 3 types: hydrophobic amino acids (H), hydrophilic amino acids (P) and neutral amino acids (N). Each residue is reduced to a bead which locates in the position of the C_α atom. The off-lattice model is adopted and the relative entropy is used as a minimization function to predict the tertiary structure of a protein. A new contact intensity function is given to consist with protein design research based on the relative entropy. Testing on several real proteins from Protein Data Bank (PDB) shows the good results obtained with the model and method. The root mean square deviations (RMSD) of the predicted structures relative to the native structures range from 0.30 to 0.70 nm. A foundation for studying protein design using the HNP model and the relative entropy was made.

Key words simplified amino acid model, relative entropy, protein folding

---

*This work was supported by grants from The National Natural Science Foundation of China (10574009) and Specialized Research Fund for The Doctoral Program of Higher Education (20040005013).

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

Received: November 28, 2005 Accepted: January 27, 2006