PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(12): 1331~1338
www.pibb.ac.cn

从氨基酸序列预测蛋白质折叠速率 *

郭建秀 饶妮妮 ** 刘广雄 李 杰 王云鹤
(电子科技大学生命科学与技术学院，成都 610054)

摘要 蛋白质折叠速率预测是当今生物物理学最具挑战性的课题之一．近年来，许多科研工作者开展了大量的研究工作来探索折叠速率的决定因素，许多参数和方法被相继提出．但氨基酸残基间的相互作用、氨基酸的序列顺序等信息对折叠速率的影响从未被提及．采用伪氨基酸组成的方法提取氨基酸的序列顺序信息，利用蒙特卡洛方法选择最佳特征因子，建立线性回归模型进行折叠速率预测．该方法能在不需要任何(显示)结构信息的情况下，直接从蛋白质的氨基酸序列出发对折叠速率进行预测．在 Jackknife 交互检验方法的验证下，对含有 99 个蛋白质的数据集，发现折叠速率的预测值与实验值有很好的相关性，相关系数能达到 0.81，预测误差仅为 2.54．这一精度明显优于其他基于序列的方法，充分说明蛋白质的序列顺序信息是影响蛋白质折叠速率的重要因素．

关键词 蛋白质折叠，折叠速率预测，伪氨基酸组成，蒙特卡罗方法
学科分类号 Q612，Q51，Q811.4
DOI: 10.3724/SP.J.1206.2010.00380

蛋白质是一类重要的生物大分子，是生命活动的主要承担者，在生物体内占有特殊的地位．每种蛋白质分子都有自己特有的氨基酸组成和排序，只有当这种氨基酸链折叠形成正确的三维空间结构才能具有正常的生物学功能．错误的折叠不仅会丧失其生物学功能，甚至会引起疾病，如疯牛病、Alzheimer's 综合症[1]等．蛋白质折叠问题，是分子生物学中心法则尚未解决的一个重大生物学问题，被列为 21 世纪生物物理学的重要课题．研究蛋白质折叠不仅具有重大的科学意义，而且在医学和生物工程领域具有极大的应用价值，如工业用酶、农药、医药的合理化设计等．

揭示蛋白质的折叠机理是一项具有挑战性的工作，其中重要的一项任务便是确定折叠速率的决定因素．尽管这一答案可以从各种生物实验中找到，如各种光谱技术、质谱和核磁共振[2-6]，但这些方法费时且昂贵．随着物理、数学的发展，特别是计算机技术的进步，寻找一种快速准确的理论计算方法来预测蛋白质的折叠速率越来越受到人们的重视．

近年来，许多科研工作者开展了大量的研究工作来探索折叠速率的决定因素，各种预测方法被相继提出．现有的预测方法大致可以分为三大类[7-8]．一类是基于三级结构的预测方法[9-15]，如接触序 CO[9]、长程序 LRO[10]、总接触距离 TCD[11]．然而，要得到这类方法所必须的三级结构信息，仍需要进行大量花费高、周期长的分子实验，无法达到快速预测的目的．第二类是基于二级结构的预测方法[16-20]，如二级结构含量 SSC[16]、有效长度 L_{eff}[17]．而这类方法所需的二级结构信息，要么和第一类方法类似，由分子实验得到，要么从一级序列预测得到，而这种逐级预测方法的预测精度将会受到二级结构预测精度的限制．第三类是基于一级序列的预测方法[21-30]，如长度 L[22]、组成信息 CI[25]、Fold-Rate[26]、N_{α}[27]、QRSM[28]．这类方法绕过了结构预测，能直接从氨基酸序列出发预测蛋白质的折叠速率，但这类方法在一定程度上仍需要知道部分结构信息，如结构分类信息．有关现有预测方法更详细的介绍请参考文献[7, 8]．

虽然人们考虑了各种结构因素与折叠速率的相

* 国家自然科学基金资助项目(30900318, 60571047).
** 通讯联系人.
Tel: 028-83206489, E-mail: raonn@uestc.edu.cn
收稿日期：2010-07-20，接受日期：2010-10-11

关性，但氨基酸的序列顺序信息和氨基酸间的相互作用从未被提及．在不使氨基酸残基数量发生变化的情况下，将氨基酸残基的顺序进行重排，折叠速率会发生巨大的变化．所有没有考虑这种变化的预测方法，似乎都不可避免地存在着不精确性．如果序列的顺序信息能被充分利用，不仅能使预测精度显著提高，还能揭示蛋白质折叠过程．然而，提取氨基酸的序列顺序信息是一项非常困难的工作．例如，对一个长度仅为 50 个氨基酸残基的蛋白质而言，不同的序列顺序组合数就有 $20^{50}\approx1.13\times10^{65}$ 之多；对一个长度为 360 个氨基酸残基的蛋白质，不同的顺序组合数将会是 $20^{360}\approx1.13\times10^{468}$．面对如此庞大的计算量，在目前的情况下，很难找到一种计算方法进行折叠速率预测，即使是建立抽样统计的数据集也无法实现．

面对上述困难，在本文中，我们尝试采用伪氨基酸组成[31]的方法提取氨基酸序列的位置信息，利用蒙特卡洛方法选择最佳预测特征因子，建立线性回归模型进行折叠速率预测．该方法能在不需要任何(显性)结构信息的情况下，直接从蛋白质的氨基酸序列出发预测其折叠速率．对含有 99 个蛋白质的数据集，在 Jackknife 交互检验方法的验证下，折叠速率预测值与实验值有着很好的相关性，相关系数能达到 0.81，预测误差仅为 2.54．在相同的数据集下，我们与 5 个有代表性的基于序列的预测方法进行了比较．结果表明，无论是预测精度还是预测误差，我们的方法均优于其他的方法．这一结果也充分说明，蛋白质的序列顺序信息是影响蛋白质折叠速率的重要因素．

1 材料和方法

1.1 数据集

为尽可能多地利用已知折叠速率实验数据的蛋白质序列，我们从大量的文献[17, 25-28, 30]及数据库[32-33]中，收集到了许多的预测数据集．为避免过拟合，我们对这些数据集进行了整理，剔除掉重复序列后，最终确定了 99 个已知折叠速率的蛋白质序列作为实验数据集，如表 1 所示． 99 个蛋白质的氨基酸序列均来源于 PDB 数据库(http:// www.rcsb.org/pdb)．

Table 1 List of proteins used in this work

PDB code	Len	PDB code	Len	PDB code	Len	PDB code	Len	PDB code	Len	PDB code	Len	PDB code	Len
1PGB:B	16	1FMK:_	57	1G6P:A	66	1PBA:_	81	1FNF:94	94	1QTU:_	115	1I1B:_	151
1L2Y:A	20	1FEX:_	59	1CSP:_	67	1CEI:_	85	1N88:_	96	1HCD:_	118	2RN2:_	155
1PIN:A	32	1SHF:A	59	1NYF:_	67	1IMQ:_	85	1URN:A	97	1HMK:_	121	2A5E:_	156
1VII:_	36	1BDD:_	60	1PSF:_	69	1POH:_	85	2ACY:_	98	1ADW:A	123	1BEB:A	156
1E0L:_	37	1SHG:_	62	1UZC:_	69	1PNJ:_	86	1APS:_	98	2VIK:_	126	1RA9:_	159
1E0M:_	37	2PTL:_	62	1MJC:_	69	1NTI:A	86	1HNG:A	98	1EAL:_	127	2LZM:_	164
1K9Q:A	40	1TUD:_	62	2HQI:_	72	1K8M:_	87	1HX5:_	99	3CHY:_	128	1LOP:A	164
2PDD:_	43	1SRL:_	64	2A3D:_	73	1BRS:D	89	1RIS:_	101	1HEL:_	129	1PHP:_N	175
1ARR:_	53	1C8C:A	64	2AIT:_	74	1TIT:_	89	1YCC:_	103	1IFC:_	131	1PHP:_C	219
1BA5:_	53	1COA:_	64	1PKS:_	76	1FNF:90	90	1HRC:_	104	1OPA:A	134	1B9C:A	224
1PRB:_	53	1HZ6:A	65	1UBQ:_	76	1TEN:_	90	256B:A	106	1CBI:A	136	2BLM:A	260
1IDY:_	54	2CRO:_	65	1RFA:_	79	1GXT:A	91	1BKF:_	107	1DK7:_	146	1QOP:A	268
1ENH:_	54	2CI2:I	65	1AYE:_	80	1WIT:_	93	1BNI:A	110	1JOO:_	149	1L8W:A	338
1DIV:N	56	1C9O:A	66	1LMB:3	80	1DIV:C	93	1SCE:A	112	1A6N:_	151	1QOP:B	392
1PGB:_	56												

Len: The length of sequence.

1.2 伪氨基酸组成

为了提高蛋白质亚细胞定位的预测精度，Chou 等[31]提出了伪氨基酸组成的概念．根据 Chou 的伪氨基酸组成原理，蛋白质序列的位置信息可在一定程度上由一组序列相关因子 $\theta_1, \theta_2, \cdots\cdots, \theta_\lambda$ 反映，定义如下：

$$\begin{cases}\theta_1=\dfrac{1}{L-1}\sum_{i=1}^{L-1}\Theta(R_i,R_{i+1})\\ \theta_2=\dfrac{1}{L-2}\sum_{i=1}^{L-2}\Theta(R_i,R_{i+2})\\ \theta_3=\dfrac{1}{L-3}\sum_{i=1}^{L-3}\Theta(R_i,R_{i+3})(\lambda<L)\\ \cdots\cdots\\ \theta_\lambda=\dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\Theta(R_i,R_{i+\lambda})\end{cases} \quad (1)$$

其中，θ_1称为第一层相关因子，反映了相邻残基间的序列顺序相关性，如图 1a 所示；θ_2称为第二层相关因子，反映了相间一个残基的残基间的序列顺序相关性，如图 1b 所示；θ_3称为第三层相关因子，反映了相间两个残基的残基间的序列顺序相关性，如图 1c 所示；以此类推．一个蛋白质序列便可由一个$(20+\lambda)$维的向量 $X=(x_1, \cdots\cdots, x_{20}, \theta_1, \cdots\cdots, \theta_\lambda)$来表示．在向量 X 中，前 20 个特征因子代表了氨基酸的组成信息，后 λ 个特征因子表示了氨基酸序列的顺序信息．在本文中，相关函数定义为：

$$\Theta(R_i,R_j)=|H(R_j)-H(R_i)| \quad (2)$$

这里 $H(R_i)$定义为氨基酸残基 R_i 的疏水值．因为 Kauzmann 曾指出疏水作用力是驱使蛋白质折叠的主要作用力．

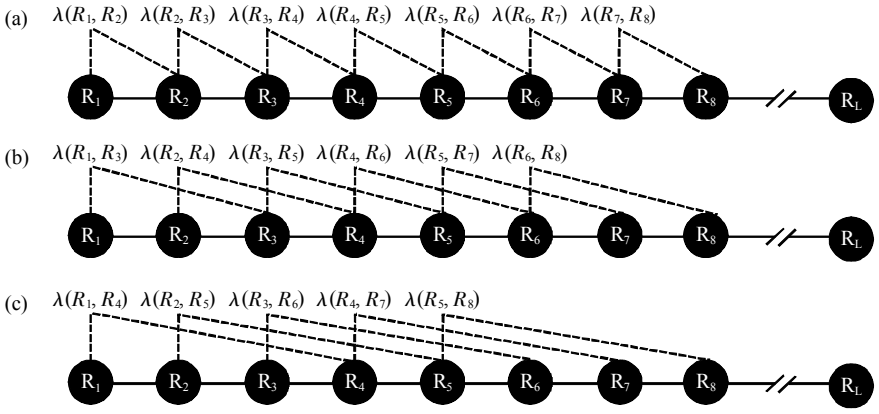

Fig. 1 A schematic drawing to show the sequence order correlation mode along a protein sequence

The first-tier panel (a) reflects the correlation mode between all the most contiguous residues, the second-tier panel (b) that between all the second-most contiguous residues, and the third-tier panel (c) that between all the third-most contiguous residues.

1.3 序列特征提取

已有研究表明[31]，λ 的最佳值应该使得预测算法能在 Jackknife 检验(详见 **1.4** 节)中取得最好的效果．对于本文中 λ 个序列顺序相关因子的选择，我们进行了多次实验，最终发现 λ 取 10 时预测效果最佳．此时我们便可用一个 20+λ=30 维的向量来表征一个蛋白质序列．

在进行蛋白质折叠速率预测时，我们需要从这个 30 维的向量中找出与折叠速率相关的特征因子．对于前 20 个代表氨基酸组成信息的特征因子，在先前的研究中[25]，我们利用了氨基酸的出现频率与蛋白质折叠速率的相关程度来挑选具有较高相关程度的氨基酸来预测蛋白质折叠速率．值得注意的是：我们的选取方法仅考虑了单个氨基酸与折叠速率的关系，忽略了氨基酸间的相互关系．然而，由于氨基酸残基间的相互影响，两个与折叠速率相关程度都较高的氨基酸组合后，与折叠速率的相关程度有可能会比其中任何一个都要低．如表 2 所示，4 个氨基酸都是与折叠速率相关程度较高的氨基酸(表的对角元素表示该氨基酸与折叠速率的相关程度)，但是它们的两两线性组合与折叠速率的相关程度并不都是增加的．例如，丙氨酸 A 与蛋白质折叠速率有较高的相关系数，为–0.39，但是它与其他任何一种氨基酸的组合都会使得相关程度有所降低，见表 2 的第一列所示．在我们先前的研究[25]中，选取了具有较高相关程度氨基酸的出现频率作为特征因子，对这些特征因子采用单纯的加减来预测蛋白质折叠速率，即线性组合中系数选为 1 或

–1，这虽然是一种非常简便的方法，但它往往会降低预测精度．例如，单独使用缬氨酸 V 和色氨酸 W 的出现频率 f_V 和 f_W 作为特征因子进行折叠速率预测时，预测精度分别为 0.30 和 0.21，但同时选择它们的出现频率的和 f_V+f_W 时，预测精度为 0.13．同样采用丙氨酸 A 和天冬酰胺 N 的出现频率 f_A 和 f_N 的预测精度分别为–0.39 和 0.21，但是同时采用二者的出现频率–f_A+f_N 的预测精度为–0.15. 因此，预测蛋白质折叠速率时，我们必须考虑氨基酸间的相互作用，即氨基酸的组合与蛋白质折叠速率之间的关系．

Table 2　The correlation coefficients between different combinations of amino acid and folding rates

Amino acid	A	N	V	W
A	–0.39			
N	0.16	0.21		
V	0.26	0.32	0.30	
W	0.24	0.29	0.33	0.21

The elements in leading diagonal are the correlation coefficient between the occurrence frequencies of single amino acid with folding rates.

本文中，从 30 个特征因子中选取哪些因子能获得最佳的预测精度呢？从理论上来说，要得到最佳的特征因子个数，我们需要把所有可能的氨基酸组合都试一遍，然后选择具有较高预测精度的特征因子作为代表因子．但是这种做法计算量十分庞大，需要的计算次数为 $\sum_{n=1}^{30} C_{30}^{n}$ 次．为了解决计算量的问题，对于给定的特征因子个数，我们利用蒙特卡洛方法来验证其平均预测精度．具体步骤为：a．给定初始特征因子个数 N；b．从 30 个特征因子中随机选取 N 个因子，利用 Jackknife 方法获得预测精度，利用蒙特卡洛方法，获得平均预测精度；c．N=N+1，返回(b)；d．重复上述步骤，使得特征因子个数取遍所有可能的情况．

图 2 给出了 200 000 次蒙特卡洛实验平均预测精度随着特征因子个数选取的变化情况．从图 2 中可以看到，当特征因子个数为 14 时，平均预测精度达到最高．因此我们选择 14 个特征因子来预测蛋白质折叠速率．

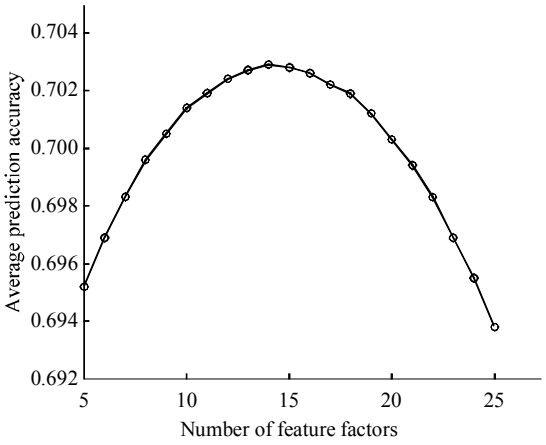

Fig. 2　Changing trends of the average prediction accuracy with different numbers of feature factors

x-axis: The number of feature factors chose in Monte Carlo. y-axis: The average prediction accuracy. The average prediction accuracy will be the largest when the number of feature factors is 14.

具体到应该选取哪 14 个特征因子才能获得最佳预测精度的问题时，理论上仍需先计算所有的组合形式，再从中选择最佳组合，这样的选择过程计算量仍十分庞大．本文中，我们再次利用蒙特卡洛方法来选取具有最佳预测精度的特征因子．我们从 30 个特征因子中随机选取 14 个特征因子进行蛋白质折叠速率预测，利用 Jackknife 检验方法获得预测精度．将预测精度大于事先设定的门限时的特征因子进行提取并统计．

图 3 给出了利用 1 000 000 次蒙特卡洛实验选取 14 个特征因子的统计直方图，其中预测精度门限设为 0.79．横坐标代表伪氨基酸组成的 30 个特

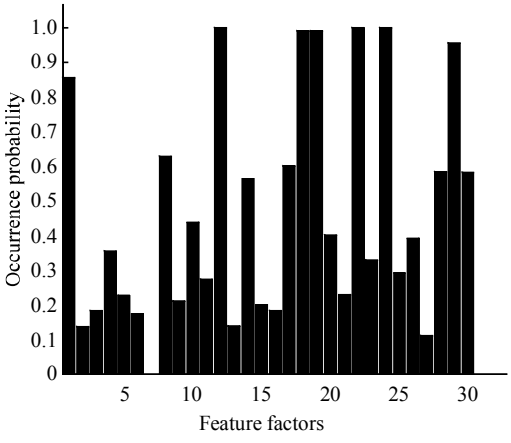

Fig. 3　The occurrence probability of different feature factors in Monte Carlo

x-axis: 30 feature factors of pseudo amino acid composition x_1, ……, x_{20}, θ_1, ……, θ_{10}. y-axis: Occurrence probability of each factor.

征因子，纵坐标代表蒙特卡洛实验中每个特征因子的出现概率．我们可以从直方图中挑选出 14 个具有较高出现概率的特征因子作为预测蛋白质折叠速率的最佳特征因子．从图 3 中我们可以得到{x_1, x_8, x_{10}, x_{12}, x_{14}, x_{17}, x_{18}, x_{19}, x_{20}, θ_2, θ_4, θ_8, θ_9, θ_{10}}为预测蛋白质折叠速率的最佳特征因子．

此外，早期的研究发现，蛋白质的主链长度 L 及其变换形式 $L^{1/2}$、$L^{2/3}$、ln(L)与折叠速率有着很好的相关性[22, 34-36]，是影响折叠速率的重要因子．因此，在本文中，我们又引入了 ln(L)这一特征因子作为预测折叠速率的第 15 个特征因子．

1.4 评价方法

在统计预测中，下列三种方法经常被用来检验各种预测方法的有效性：独立集检验、子集抽样检验以及 Jackknife 检验．其中，对于一个给定的数据集，Jackknife 检验[37]总能得到一个唯一的结果，因而被证明为一种最客观有效的检验方法[38-39]．本文便采用 Jackknife 检验方法来评价我们的预测方法．Jackknife 检验的过程如下：在每次运行中，剔出一个蛋白质作为一个独立的检测样本，用剩下的数据作为训练样本，建立线性回归预测模型，推导出所有的计算参数，再用该模型对先前剔出的蛋白质进行预测，运行 N 次后，便能得到所有蛋白质的预测值．

本文中，我们采用了两个评价指标来表征预测方法的有效性，折叠速率预测值与实验值间的相关系数 R 和标准误差 σ，其定义如下：

$$R=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}} \quad (3)$$

$$\sigma=\sqrt{\frac{\sum_{i=1}^{N}(x_i-y_i)^2}{N-2}} \quad (4)$$

其中，N 为样本数，x_i、y_i 分别为第 i 个蛋白质折叠速率的预测值与实验值，$\bar{x}$、$\bar{y}$ 分别为平均预测值与平均实验值．R 越大，σ 越小，说明预测效果越好．此外，我们用 P 值来表征相关系数 R 的显著性水平．

2 预测结果与方法比较

2.1 预测结果

根据上面的特征选取方法，我们最终选取了一个 15 维的向量来表征一个蛋白质序列，用多元线性回归模型对折叠速率进行预测．在 Jackknife 方法的验证下，预测结果如图 4 所示．预测值与实验值的相关系数为 0.81($P < 10^{-4}$)，标准误差为 2.54．

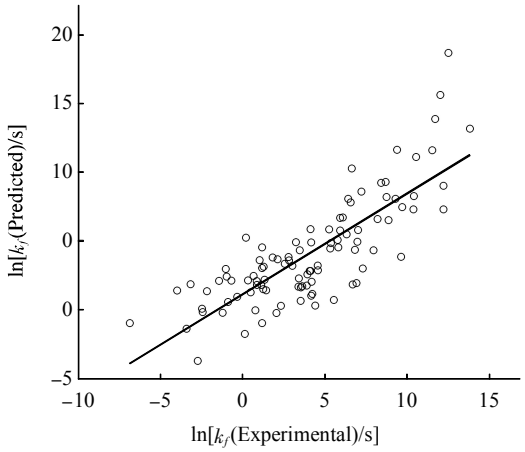

Fig. 4 Relationship between the experimental and predicted folding rates using linear regression model with Jackknife test for a set of 99 proteins

2.2 方法比较

为了评价我们提出的方法的优劣，与 5 个有代表性基于序列的方法[25-29]进行了比较．因我们的数据集包含了其他方法中所没用到的数据，为得到一个客观公正的比较结果，针对我们的数据集，把其他的方法重新进行了计算，结果如表 3 所示．从表 3 中可以看出，无论是相关系数 R，还是标准误差 σ，我们的方法都明显优于其他的预测方法．提

Table 3 Performances of different methods in predicting protein folding rate

Method	R	P	σ
Pred-PFR[1)]	0.78	8.15×10^{-19}	2.60
Fold-Rate[2)]	0.23	0.023	5.45
CI[3)]	0.71	3.63×10^{-17}	3.93
QRSMd[4)]	0.45	2.58×10^{-6}	11.80
N_α[5)]	0.40	3.24×10^{-5}	8.45
Current	0.81	6.79×10^{-24}	2.54

R: The correlation coefficient; P: The significant level; σ: The standard error. [1)] Result from the Pred-PFR web server at http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/; eight sequences are too short to be predicted (< 50 amino acids) and are excluded from evaluation. [2)] Result from the Fold-Rate web server at http://psfs.cbrc.jp/fold-rate/. [3)]Result from the CI web server at http://sdbi.sdut.edu.cn/FDserver. [4)] Result from the QRSM web server at http://210.60.98.17/FOLDRATE20r/foldrate20.htm. [5)] Result from the N_α web server at http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html.

高的预测精度和减小的预测误差进一步说明，序列顺序信息对蛋白质折叠速率有着很大的影响．设计算法时应该充分考虑这种影响，以便较大幅度地提高预测方法的预测精度．

3 讨 论

在用蒙特卡洛实验方法进行序列特征选择时，我们发现，有 9 个氨基酸在折叠过程中对折叠速率起着重要的影响作用．按照它们在蒙特卡洛实验中出现次数的多寡，依次为 N(天冬酰胺)、W(色氨酸)、V(缬氨酸)、A(丙氨酸)、Q(谷氨酰胺)、I(异亮氨酸)、T(苏氨酸)、L(亮氨酸)和 Y(酪氨酸)．其中，色氨酸 W、缬氨酸 V、丙氨酸 A、异亮氨酸 I 以及亮氨酸 L 均为非极性疏水氨基酸，这一结果和文献[40]的报道一致．该报道称，增加蛋白质的疏水性会加速折叠进程，因为蛋白质的疏水部分可以促进收缩，从而稳定了过渡态系统．上述结果也表明，蛋白质序列的疏水氨基酸含量是决定折叠速率的一个重要因素[41]．其余的 3 种氨基酸虽然不是非极性疏水氨基酸，但它们均是极性不带电氨基酸，具有极高的柔韧性，能够加速折叠．而且天冬酰胺 N 和谷氨酰胺 Q 既是 H^+ 的供体，又是 H^+ 的受体，这种特性与蛋白质的结构和功能有着重要的联系，苏氨酸 T 是带有脂肪旋羟基侧链的氨基酸，其侧链能和适当的供体和受体基团形成氢键．所选的这 9 个氨基酸均非带电氨基酸，这一结果说明蛋白质中正、负电荷的侧链基团对折叠速率没有太大的影响．

对于序列顺序相关因子，我们发现 θ_2, θ_4, θ_8, θ_9 和 θ_{10} 是影响折叠速率的重要因素．其中 θ_2 和 θ_4 与折叠速率有着很强的相关性，这两个因子可以认为是蛋白质结构局部接触序信息的总和(α 碳原子残基间的截距分别取为 l_{cut}=2 和 l_{cut}=4)．这一结论已经在之前的研究[9, 42]中得到证实：残基间的相互接触影响着蛋白质的折叠速率．而序列相关因子 θ_8, θ_9 和 θ_{10} 与蛋白质的折叠速率并没有显著的相关关系，但它们却对最后的预测精度有着很大的影响．这一结果可能和文献[43−44]的研究结论一致．该研究认为，增加未卷入折叠核的那一部分序列的疏水性会加快蛋白质的折叠过程．θ_8, θ_9 和 θ_{10} 则可以认为是未卷入折叠核的那一部分序列的序列特征．最近的研究工作[45]表明，大约 15 个氨基酸易于折叠成一个螺旋结构，即当前的氨基酸仅和其前后 7 个相邻的氨基酸间产生影响，和其他氨基酸间的作用可以忽略不计．换句话说，第 8、9、10 个氨基酸不会和当前的氨基酸发生作用，将裸露在折叠核外面．在本文中，序列顺序相关因子 θ_8, θ_9 和 θ_{10} 正分别是第 8、9、10 位残基位置信息的总和．

尽管我们的方法提供了很好的预测质量(精度高、误差小)，但也有不足之处．一是没有考虑实验环境对折叠速率的影响，例如数据集中的“慢”折叠蛋白 1YCC:_[46]，它的折叠速率的测定温度为 40℃；蛋白质 1PRB:_[47]，它的折叠速率是在 50℃而非 25℃的常温环境下测得的；实验环境会对折叠速率的测定造成很大的影响，温度升高会加快折叠．采用不同实验条件下测得的实验数据进行折叠速率预测，势必会对最后的预测精度造成影响．而蛋白 1PGB:B[48]和 1L2Y:A[49]，是两个人工合成的多肽(β-hairpin 和 α-helix，长度分别为 16 和 20)，它们的折叠速率可能不同于天然蛋白质．二是该方法仅考虑了氨基酸的疏水性对折叠速率的影响，没有充分应用氨基酸其他的物理化学属性[50]对折叠速率所起的作用，可能会忽略掉一些重要的序列顺序信息，这些因素可能会影响到最后的预测精度．因此，寻找一个更大更精确的数据集以及选择更多的氨基酸属性来验证我们的方法，将是我们未来工作的一个重点．

4 结 论

在考虑了氨基酸间的相互作用、氨基酸的序列顺序信息对折叠速率的影响后，我们提出了一个新的蛋白质折叠速率预测方法．该方法能够直接从氨基酸的序列出发，在不需要任何结构信息的情况下，对折叠速率进行预测．在新方法中，我们用 Chou的伪氨基酸组成方法提取序列的顺序信息，用蒙特卡洛实验方法进行特征筛选，用线性回归方法进行折叠速率预测．在 Jackknife 方法的验证下，我们的方法取得了很好的预测效果，预测值与实验值的相关系数能达到 80.7%，预测误差仅为 2.54．这一精度要明显优于其他的预测方法，也充分说明氨基酸的序列顺序信息是影响折叠速率的重要因素．

参 考 文 献

[1] Taubes G. Misfolding the way to disease. Science, 1996, **271**(5255): 1493−1495

[2] Fabian H, Naumann D. Methods to study protein folding by stopped-flow FT-IR. Methods, 2004, **34**(1): 28−40

[3] Zeeb M, Balbach J. Protein folding studied by real-time NMR

spectroscopy. Methods, 2004, **34**(1): 65−74

[4] Maity H, Maity M, Krishna M M G, *et al*. Protein folding: the stepwise assembly of foldon units. Proc Nat Acad Sci USA, 2005, **102**(13): 4741−4746

[5] Xiao H, Hoerner J K, Eyles S J, *et al*. Mapping protein energy landscapes with amide hydrogen exchange and mass spectrometry: Ⅰ. A generalized model for a two-state protein and comparison with experiment. Protein Sci, 2005, **14**(2): 543−557

[6] Maxwell K L, Wildes D, Zarrine-Afsar A, *et al*. Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. Protein Sci, 2005, **14**(3): 602−616

[7] 郭建秀, 马彬广, 张红雨. 蛋白质折叠速率预测研究进展. 生物物理学报, 2006, **22**(2): 89−95
Guo J X, Ma B G, Zhang H Y. Acta Biophys Sin, 2006, **22**(2): 89−95

[8] Gromiha M M, Selvaraj S. Bioinformatics approaches for understanding and predicting protein folding rates. Current Bioinformatics, 2008, **3**(1): 1−9

[9] Plaxco K W, Simons K T, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol, 1998, **277**(4): 985−994

[10] Gromiha M M, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol, 2001, **310**(1): 27−32

[11] Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J, 2002, **82**(1): 458−463

[12] Nölting B, Schälike W, Hampel P, *et al*. Structural determinants of the rate of protein folding. J Theor Biol, 2003, **223**(3): 299−307

[13] Weikl T R, Dill K A. Folding kinetics of two-state proteins: Effect of circularization, permutation, and crosslinks. J Mol Biol, 2003, **332**(4): 953−963

[14] Ivankov D N, Garbuzynskiy S O, Alm E, *et al*. Contact order revisited: influence of protein size on the folding rate. Protein Sci, 2003, **12**(9): 2057−2062

[15] Mirny L, Shakhnovich E. Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct, 2001, **30**(1): 361−396

[16] Gong H, Isom D G, Srinivasan R, *et al*. Local secondary structure content predicts folding rates for simple two-state proteins. J Mol Biol, 2003, **327**(5): 1149−1154

[17] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Nat Acad Sci USA, 2004, **101**(24): 8942−8944

[18] Fleming P J, Gong H P, Rose G D. Secondary structure determines protein topology. Protein Sci, 2006, **15**(8): 1829−1834

[19] Huang J T, Cheng J P, Chen H. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins, 2007, **67**(1): 12−17

[20] Prabhu N P, Bhuyan A K. Prediction of folding rates of small proteins: empirical relations based on length, secondary structure content, residue type, and stability. Biochemistry, 2006, **45**(11): 3805−3812

[21] Shao H, Peng Y, Zeng Z H. A simple parameter relating sequences with folding rates of small helical proteins. Protein Pept Lett, 2003, **10**(3): 277−280

[22] Galzitskaya O V, Garbuzynskiy S O, Ivankov D N, *et al*. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins, 2003, **51**(2): 162−166

[23] Huang J T, Jing T. Amino acid sequence predicts folding rate for middle-size two-state proteins. Proteins, 2006, **63**(3): 551−554

[24] Gromiha M M. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model, 2005, **45**(2): 494−501

[25] Ma B G, Guo J X, Zhang H Y. Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. Proteins, 2006, **65**(2): 362−372

[26] Gromiha M M, Thangakani A M, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res, 2006, **34**(suppl_2): 70−74

[27] OuYang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci, 2008, **17**(7): 1256−1263

[28] Huang L T, Gromiha M M. Analysis and prediction of protein folding rates using quadratic responde surface models. J Comput Chem, 2008, **29**(10): 1675−1683

[29] Shen H B, Song J N, Chou K C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomedical Science and Engineering, 2009, **2**(3): 136−143

[30] Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem, 2009, **30**(5): 772−783

[31] Chou K C. Prediction of protein cellular attributes using pesudo-amino acid composition. Proteins, 2001, **43**(2): 246−255

[32] Fulton K F, Bate M A, Faux N G, *et al*. Protein Folding Database (PFD 2.0): an online environment for the International Foldeomics Consortium. Nucleic Acids Res, 2007, **35**(Database issue): D304−D307

[33] Bogatyreva N S, Osypov A A, Ivankov D N. KineticDB: a database of protein folding kinetics. Nucleic Acids Res, 2009, **37**(Database issue): D342−346

[34] Gutin A M, Abkevich V I, Shakhnovich E I. Chain length scaling of protein folding Time. Phys Rev Lett, 1996, **77**(27): 5433−5436

[35] Thirumalai D. From minimal models to real proteins: time scales for protein folding kinetics. J Phys I, 1995, **5**(11): 1457−1467

[36] Finkelstein A V, Badretdinov A Y. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des, 1997, **2**(2): 115−121

[37] Chou K C, Zhang C T. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol, 1995, **30**(4): 275−349

[38] Chou K C, Shen H B. Recent progress in protein subcellular location prediction. Anal Biochem, 2007, **370**(1): 1−16

[39] Chou K C, Shen H B. Cell-PLoc: a package of Web servers for

predicting subcellular localization of proteins in various organisms. Nature Protocols, 2008, **3**(2): 153-162

[40] Viguera A R, Vega C, Serrano L. Unspecific hydrophobic stabilization of folding transition states. Proc Nat Acad Sci USA, 2002, **99**(8): 5349-5354

[41] Cranz-Mileva S, T. Friel C, E. Radford S. Helix stability and hydrophobicity in the folding mechanism of the bacterial immunity protein Im9. Protein Eng Des Sel, 2005, **18**(1): 41-50

[42] Muñoz V, Eaton W A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Nat Acad Sci USA, 1999, **96**(20): 11311-11316

[43] Northey J G B, Nardo A A D, Davidson A R. Hydrophobic core packing in the SH3 domain folding transition state. Nat Struct Biol, 2002, **9**(2): 126-130

[44] Dias C L, Ala-Nissila T, Wong-ekkabut J, *et al*. The hydrophobic effect and its role in cold denaturation. Cryobiology, 2010, **60**(1): 91-99

[45] Lee S, Lee B C, Kim D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins, 2006, **62**(4): 1107-1114

[46] Mines G A, Pascher T, Lee S C, *et al*. Cytochrome c folding triggered by electron transfer. Chemistry & Biology, 1996, **3**(6): 491-497

[47] Wang T, Zhu Y, Gai F. Folding of a three-helix bundle at the folding speed limit. J Physi Chem B, 2004, **108**(12): 3694-3697

[48] Muñoz V, Thompson P A, Hofrichter J, *et al*. Folding dynamics and mechanism of β-hairpin formation. Nature, 1997, **390**(13): 196-199

[49] Qiu L, Pabit S A, Roitberg A E, *et al*. Smaller and faster: the 20-residue trp-cage protein folds in 4 μs. J. Am Chem Soc, 2002, **124**(44): 12952-12953

[50] Kawashima S, Ogata H, Kanehisa M. AAindex: Amino acid index database. Nucleic Acids Res, 1999, **27**(1): 368-369

Predicting Protein Folding Rate From Amino Acid Sequence*

GUO Jian-Xiu, RAO Ni-Ni**, LIU Guang-Xiong, LI Jie, WANG Yun-He

(*School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu* 610054, China)

Abstract Prediction of protein folding rate is one of the most important challenges in contemporary biophysics. Over the past few years, many researchers have devoted great efforts to reveal the major determinants of protein folding rate, and many parameters and methods have been proposed successively. However, the interaction of amino acids and the sequence order information have never been considered as a property for predicting protein folding rates. It was proposed a novel method, which adopted Chou's pseudo-amino acid composition to extract the sequence order information, used Monte Carlo method to choose the optimal feature factors, and established the linear regression model to predict the protein folding rate. This novel method can predict protein folding rate from amino acid sequence without any knowledge of the tertiary or secondary structure, or structural class information. Using the Jackknife cross validation test, for the largest dataset yet studied including 99 proteins, it was found that the predicted folding rates correlated well with the experimental values; the correlation coefficient is 0.81, and the standard error is 2.54. The prediction quality is excelled with most existing sequence-based methods. The result implies that the sequence order information plays an important role in protein folding.

Key words protein folding, prediction of folding rate, pseudo-amino acid composition, Monte Carlo method

DOI: 10.3724/SP.J.1206.2010.00380

*This work was supported by a grant from The National Natural Science Foundation of China (30900318, 60571047).

**Corresponding author.

Tel: 86-28-83206489, E-mail: raonn@uestc.edu.cn

Received: July 20, 2010 Accepted: October 11, 2010