doi:10.3969/j.issn.1672-5565.2013.02.02

基于序列疏水值震荡的折叠速率预测

胡 睿,史小红*,李晋惠
(西安工业大学理学院,陕西 西安 710032)

摘 要:蛋白质折叠速率的正确预测对理解蛋白质的折叠机理非常重要。本文从伪氨基酸组成的方法出发,提出利用序列疏水值震荡的方法来提取蛋白质氨基酸的序列顺序信息,建立线性回归模型进行折叠速率预测。该方法不需要蛋白质的任何二级结构、三级结构信息或结构类信息,可直接从序列对蛋白质折叠速率进行预测。对含有62个蛋白质的数据集,经过Jackknife交互检验验证,相关系数达到0.804,表示折叠速率预测值与实验值有很好的相关性,说明了氨基酸序列信息对蛋白质折叠速率影响重要。同其他方法相比,本文的方法具有计算简单,输入参数少等特点。

关键词:蛋白质折叠;折叠速率;疏水值震荡;伪氨基酸组成;回归分析
中图分类号:Q612; **文献标识码**:A **文章编号**:1672-5565(2013)-02-086-05

Prediction of Protein Folding-rate Based on the Hydrophobic Value Vibration

HU Rui, SHI Xiao-hong*, LI Jin-hui
(*School of science, Xi'an Technological University, Xi'an* 710032, *China*)

Abstract: Prediction of protein folding rates is important in understanding the overall folding the mechanism. This article gives a new method, which adopted hydrophobic value vibration to extract the sequence order information, established the linear regression model to predict the protein folding rate. This method can predict protein folding rate from amino acid sequence without any knowledge of the tertiary or secondary structure, or structural class information. Using Jackknife cross test to check the 62 proteins, the correlation coefficient is 0.804. It means that the predicted folding rates correlated well with the experimental values and the result implies that the sequence order information plays an important role in protein folding. Compared with other models, this method has advantages feathers in simple computation, less parameters and so on.

Key words: Protein Folding; Folding Rates; Hydrophobic Value Vibration; Pseudo-amino Acid Composition; Regression Analysis;

蛋白质折叠问题是计算生物学和生物信息学中的核心问题之一,对与理解蛋白质的折叠机制和分析蛋白质折叠的决定因素来说,预测蛋白质折叠速率非常重要,传统的实验方法来研究蛋白质折叠的方法有光谱,质谱,核磁共振等方法,但这些方法费时且昂贵[1]。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[2]。

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。现有的方法大致有以下几种:三级结构模型。通过已知的蛋白质拓扑结构构建模型预测,如相对接触距(CO)、长接触距(LRO)、总接触距(TCO)和绝对接触距(absolute contact order)等方法[3-6]。此类方法需要预先知道蛋白质的三级结构的拓扑结构信息,需要进行大量的分子实验,所以不可避免的周期很长,花费很多,背离预测的目的。二

收稿日期:2012-12-10;**修回日期**:2013-03-07.
基金资助:陕西省教育厅专项科研计划项目:基于图论模型的蛋白质结构预测问题的研究(2010JK596)。
作者简介:胡 睿,男,陕西省西安市,西安工业大学硕士研究生,E-mail:hurui66@163.com.
*通讯作者:史小红,女,陕西省西安市,西安工业大学副教授,博士。

级结构模型。通过真实和预测的二级结构来构建模型预测折叠速率,此类方法有二级结构含量的方法(SSC)[7],有效长度(Leff)方法等[8]。此类方法所需要的二级结构信息如果从分子实验得到出发,则有与从三级结构预测一样的缺点,费时费力;如果从序列预测二级结构出发,则不可避免的受到二级结构预测精度的影响,不利于再次预测折叠速率。通过序列信息来构建模型。如组成信息(CI)[9],Nα等[10],此类方法是从蛋白质序列信息出发建立模型,所以不需要知道蛋白质拓扑结构信息,但在一定程度上需要知道结构分类等信息。因为不需要蛋白质拓扑结构信息,所以通过序列信息构建模型预测蛋白质折叠速率从方法上避免了实验检测的一些困难,更接近预测的目的。

本文从蛋白质序列信息出发,构建模型预测蛋白质折叠速率。为了提取出序列信息中残基与残基之间的相互作用信息,参考了 Chou 的伪氨基酸组成方法[11],提出了氨基酸疏水值震荡的模型,利用 Matlab 采用遍取法选出最优预测因子,建立线性回归模型进行折叠速率预测。对含有 62 个蛋白质的数据集,在 Jackknife 交互检验方法的验证下,折叠速率的预测值与实验值有较好的相关性,相关系数能达到 0.804。在相同的数据集下,我们与前人引用该数据集所做的预测和其他三个有代表性的基于序列的预测方法进行了比较,结果表明,该方法在预测结果方面达到了一定的精度,说明蛋白质的序列信息是影响蛋白质折叠速率的重要因素[12]。

1　材料与方法

1.1　数据集

为了体现本文折叠速率预测方法的优良性,本文所采用的数据集是已被实验验证的 62 个蛋白质的折叠速率数据集[12]。该数据集曾经被 Ivankov 和 Finklstein 等其他人多次使用。如表 1 所示,该数据集可以从网址 http://mathbio.nankai.edu.cn/jzgao/folding_rate_database.htm 下载。

表 1　蛋白质折叠的动力学参数

Table 1　Kinetic parameters for the folding of proteins

序号	PDB	Ln(kf)	序号	PDB	Ln(kf)	序号	PDBLn(kf)	
1	1PIN	4.1	22	1PGB(57)	2.6	43	1QOP(268)	-1.1
2	2PDD	4.3	23	1FKB	0.7	44	1AON	0.3
3	2ABD	2.9	24	2CI2	1.7	45	1BRS	1.5
4	256B	5.3	25	1AYE	3	46	3CHY	0.4
5	1IMQ	3.2	26	1URN	2.5	47	2RN2	0
6	1LMB	3.7	27	1APS	-0.7	48	1RA9	2
7	1FNF(90)	-0.4	28	1RIS	2.6	49	1QOP(396)	-3
8	1WIT	0.2	29	1POH	1.2	50	1PHP(175)	1
9	1TEN	0.5	30	1DIV	2.6	51	1PHP(219)	-1.5
10	1SHG	0.6	31	2VIK	3	52	1BNI	1.1
11	1SRL	1.7	32	1A6N	0.5	53	2LZM	1.8
12	1PNJ	-0.5	33	1CEI	2.5	54	1UBQ	2.6
13	1SHF	2	34	2CRO	1.6	55	1SCE	1.8
14	1PSF	1.4	35	2A5E	1.5	56	1L2Y	5.4
15	1CSP	2.9	36	1TIT	1.6	57	1VII	5
16	1C9O	3.1	37	1HNG	0.8	58	1BDD	5.1
17	1G6P	2.7	38	1FNF(94)	2.4	59	1ENH	4.6
18	1MJC	2.	339	1IFC	1.5	60	1GXT	1.9
19	1LOP	2.9	40	1EAL	0.6	61	2ACY	0.4
20	1C8C	3	41	1OPA	0.6	62	1L8W	0.7
21	1HZ6	1.8	42	1CBI	-1.4			

1.2 疏水值震荡

本文参照 Chou 的伪氨基酸组成方法,定义了氨基酸疏水值震荡。根据伪氨基酸组成原理,蛋白质的位置信息用一组序列相关因子 $\theta_1,\theta_2,\theta_3\cdots\cdots\theta_\lambda$ 反映,称之为序列疏水值震荡,定义如下:

$$\begin{cases}\theta_1 = \ln[\sum_{i=2}^{L-1}\Phi(R_{i-1},R_i,R_{i+1})]\\ \theta_2 = \ln[\sum_{i=3}^{L-2}\Phi(R_{i-2},R_i,R_{i+2})]\\ \theta_3 = \ln[\sum_{i=4}^{L-3}\Phi(R_{i-3},R_i,R_{i+3})]\\ M\\ \theta_\lambda = \ln[\sum_{i=\lambda+1}^{L-\lambda}\Phi(R_{i-\lambda},R_i,R_{i+\lambda})]\end{cases} \quad (1)$$

其中,θ_1 为第一层疏水值震荡相关因子,反映了临近的三个残基间的相互作用,如图 1(a)所示;θ_2 为第二层疏水值震荡相关因子,反映了中心残基与其左右相隔 1 个残基之间的相互作用,如图 1(b)所示;以此类推,θ_λ 为第 λ 层疏水值震荡相关因子,反映了中心残基与其左右相隔 λ 个残基之间的相互作用。

Φ 函数定义为:

$$\Phi(R_{i-\lambda},R_i,R_{i+\lambda}) = \begin{cases}1\left[\begin{matrix}H(R_{i-\lambda}) < H(R_i) > H(R_{i+\lambda}),\\ H(R_{i-\lambda}) < H(R_i) > H(R_{i+\lambda})\end{matrix}\right]\\ 0,\text{其他}\end{cases} \quad (2)$$

其中,H(Ri)定义为氨基酸残基 Ri 的疏水值

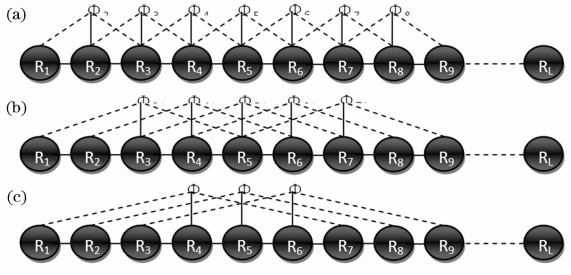

图 1 蛋白质序列疏水值震荡信息的提取

注:(a)表示第 1 层序列疏水值震荡因子,(b)表示第 2 层序列疏水值震荡因子,(c)表示第 3 层序列疏水值震荡因子。

Fig. 1 A drawing to show the hydrophobic value vibration correlation mode in a protein sequence

Notes: (a) reflects the correlation mode between all the most contiguous residues (b) reflects the correlation mode between all the second-most contiguous residues (c) reflects the correlation mode between all the third-most contiguous residues

1.3 序列预测因子的选定

本文选取预测因子时,除了选取 λ 层疏水值震荡因子外,另外还选取了构成蛋白质的 20 种残基的出现几率,组成 20 + λ 维向量。其中 λ 是由数据集中链长最小蛋白质的蛋白质决定,在本文中,λ 取值为 9。所以此时本文所用的预测因子总数为 29 个。另外,考虑到蛋白质主链长度 L 的变换形式 Ln(L)与折叠速率有很好的相关性,所以本文引入 Ln(L)这一特征因子作为预测蛋白质折叠速率的第 30 个因子。如何选取最优预测因子以达到最优结果,这是一个需要解决的问题。从理论上来说,要获得最优预测因子的解,就需要把所有可能的因子组合都试一遍,然后选择预测精度最高的因子作为最终结果。但这种选取方法计算量非常大,需计算 $\sum_{i=1}^{30}C_{30}^{i}$ 次,在本文所处的实验环境下是不可能实现的。为了解决计算量的问题,本文选择折中的方法,先对 20 个氨基酸因子进行遍取,选出预测精度最高的氨基酸因子组。然后将选出的氨基酸因子与疏水值震荡因子和 Ln(L)组合,再对组合后的因子进行遍取。

在对 20 种氨基酸进行遍取后,如图 2(a)所示,当氨基酸种类为 10 时,预测精度最高。这第 10 组的编号为 63919,对应氨基酸编号为{x1、x3、x6、x8、x10、x11、x14、x17、x18、x19},氨基酸分别为{A、D、G、I、L、M、Q、T、V、W}。

将这 10 组氨基酸数作为预测因子与疏水值震荡因子和 Ln(L)组合后,对重新组合的 20 个预测因子进行遍取,得出的结果如图 2(b)所示,当因子组为第 10 组的时候,预测精度最高。这第 10 组的编号为 55567,对应因子组编号为{1、3、5、6、7、8、11、18、19、20}。对应因子为{x1、x6、x10、x11、x14、x17、θ_1、θ_8、θ_9、Ln(L)}。

2 预测结果与方法比较

2.1 预测结果

根据上面所选出的预测因子选取方法,得出了一个由 10 维向量来表示一个蛋白质序列所含信息。用多元线性回归的方法对所选向量进行预测,在 Jackknife 交互检验的验证下,预测结果如图 3 所示,预测值与实验值的相关系数为 0.804。

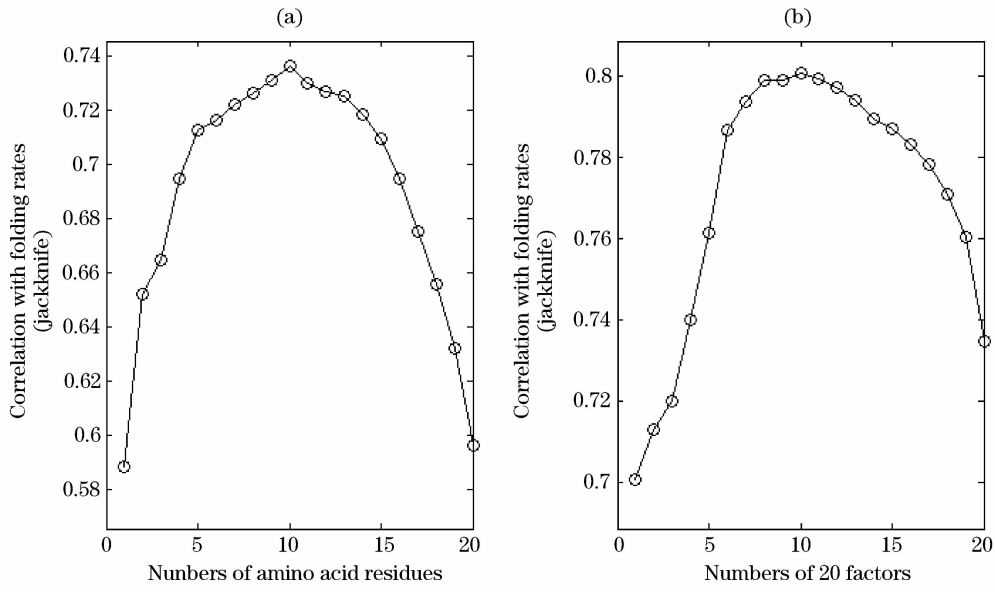

图 2　用遍取法选择预测因子

(a)表示 20 种氨基酸的遍取结果,(b)表示 20 种预测因子的遍取结果

Fig. 2　Using all – pick method choose factors

(a) shows the results of amino acid residues. shows the results of 20 factors

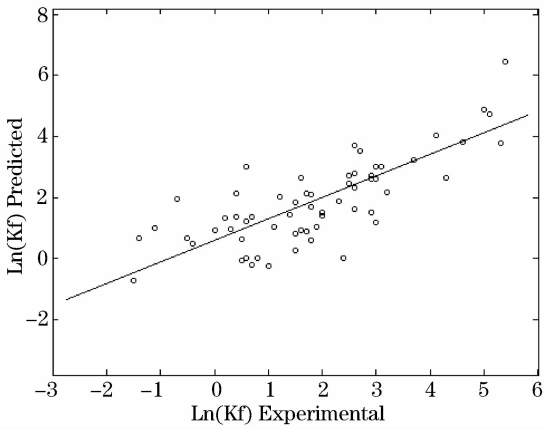

图 3　62 个蛋白质折叠速率实验值与预测值的线性回归

Fig. 3　Relationship between the experimental and predicted folding rates using linear regression model with Jackknife test for a set of 62 proteins

2.2　方法比较

为了将本文中的预测结果与其他方法比较,除了选取了文献运用该数据集所得出的结果外,还选取了其他 3 个基于序列的方法。针对本文的数据集,用这三种方法进行计算,结果如表 2 所示。虽然 PPFR[13] 方法预测结果 0.82 要好于本文方法,但是该方法用到的参数比较多,还用到了一些其他软件所预测的二级结构信息。而本文方法仅仅利用了序列信息,并不需要其他的预测软件支持,模型参数少,计算简单。

表 2　不同预测方法结果对比

Table 2　Performances of different methods in predicting protein folding rate

Method	R
CI[9]	0.72
Fold – Rate(a)	0.34
PPFR[13]	0.82
文献[12]方法	0.778
本文方法	0.804

Notes:(a) Result from the Fold – Rate web server at http://psfs.cbrc.jp/fold – rate/

3　结果分析与讨论

在用遍取法选取特征因子时,发现有 6 个氨基酸在折叠过程中对折叠速率有重要影响,它们分别为 A(丙氨酸)、G(甘氨酸)、L(亮氨酸)、M(蛋氨酸)、Q(谷氨酰胺)、T(苏氨酸)。其中 A(丙氨酸)、L(亮氨酸)、M(蛋氨酸)为非极性疏水氨基酸。根据文献[14]所描述,增加蛋白质的疏水性会使蛋白质的折叠过程加快,而蛋白质的疏水部分会加强蛋白质的收缩机制。剩下的 G(甘氨酸)、Q(谷氨酰胺)、T(苏氨酸)均为极性不带电荷氨基酸,其中 Q(谷氨酰胺)与 T(苏氨酸)的侧链可以成为适当受体与供体形成氢键,而氢键是影响蛋白质折叠的重要因素,能够加速折叠。

在选取疏水值震荡因子时,发现因子 θ_1、θ_8、θ_9 对折叠速率有重要影响。其中 θ_1 能够反映出蛋白质序列的疏水震荡的剧烈程度。疏水震荡越剧烈,程度越强,越能够加速蛋白质的折叠速率。θ_8、θ_9 则可以认为是与折叠核外相互作用的那一部分序列的序列特征。根据文献[2]、[15]的研究表明,一个氨基酸和它的前 7 个与后 7 个氨基酸比较容易形成一个螺旋结构,即当前氨基酸仅与其前后 7 个近邻的氨基酸产生影响,而 θ_8、θ_9 代表的含义分别为当前氨基酸与其前后第 8 个,第 9 个氨基酸之间的作用,并未参与到折叠核的形成,所以包含与折叠核的相互作用的那一部分序列的序列特征。

虽然本文的方法在预测折叠速率方面达到了一些精度,但是还有一些不足之处。其主要方面是在提取蛋白质序列的特征信息时,仅仅考虑了氨基酸疏水值对折叠速率的影响。而氨基酸其他的物理化学属性都没有考虑在内,这肯定会导致一些蛋白质序列信息的缺失。那么如何利用氨基酸的其他属性来进一步提高预测精度,将会是以后的一个工作重点。

4 结论

本文在参考了伪氨基酸组成方法的基础上,提出了利用氨基酸疏水值震荡的方法来提取蛋白质序列信息,能够在不需要其他结构信息的情况下,直接从蛋白质序列来预测折叠速率。在选取特征因子中,采用折中的方法来遍取筛选。Jackknife 交互检验的结果也说明本文的方法在折叠速率预测方面达到了一定的精度,相关系数为 80.4%。分析表明蛋白质序列的疏水值震荡和氨基酸的疏水信息对折叠速率的影响显著。

参考文献(References)

[1] 郭建秀, 马彬广, 张红雨. 蛋白质折叠速率预测研究进展[J]. 生物物理学报, 2006, 4(2):89-95.

[2] 郭建秀, 饶妮妮, 刘广雄, 李杰, 王云鹤. 从氨基酸序列预测蛋白质折叠速率[J]. 生物化学与生物物理进展,2010, 37(12): 1331-1338.

[3] Plaxco K W, Simons K T, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins [J]. Journal of Molecular Biology, 1998, 277(4): 985-994.

[4] 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]. Journal of Molecular Biology, 2001, 310(1): 27-32.

[5] Zhou H Y, Zhou Y Q. Folding rate prediction using total contact distance[J]. Biophysical Journal, 2002, 82(1): 458-463.

[6] Ivankov D N, Garbuzynskiy S O, Alm E, Plaxco K W, Baker D, Finkelstein A V. Contact order revisited: influence of protein size on the folding rate[J]. Protein Science, 2003, 12(9): 2057-2062.

[7] Gong H, Isom D G, Srinivasan R, Rose G D. Local secondary structure content predicts folding rates for simple, two-state proteins[J]. Journal of Molecular Biology, 2003, 327(5): 1149-1154.

[8] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure[J]. Proceedings of the National Academy of Sciences, 2004, 101(24): 8942-8944.

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

[10] Ou Y Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence[J]. Protein Science, 2008, 17(7): 1256-1263.

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

[12] 高建召,胡刚,王奎,沈世镒,基于序列和局部信息熵的蛋白质折叠速率预测模型[J]. 工程数学学报, 2010,27(6):959-966.

[13] Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequence using hybrid sequence representation[J]. Journal of Computational Chemistry, 2009, 30(5): 772-783

[14] Viguera A R, Vega C, Serrano L. Unspecific hydrophobic stabilization of folding transition states. Proceedings of the National Academy of Sciences of the USA,2002, 99(8): 5349-5354.

[15] 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.