文章编号:1000-1638(2025)04-0389-07 DOI:10.13484/j.nmgdxxbzk.20250406

参与天然态相互作用的关键残基总数决定蛋白质折叠速率*

刘 洋,王 坤,张 颖,吕 军
(内蒙古工业大学理学院,呼和浩特 010051)

摘要:理解影响蛋白质折叠速率的因素对于阐明蛋白质折叠的潜在机制至关重要。提出了一种新方法来发现影响蛋白质折叠速率的关键因素。该方法通过定义"有效长度"这一参数,来量化特定结构类蛋白质天然态中参与非局域相互作用的关键残基数量。通过在标准数据集上的验证,发现有效长度与蛋白质折叠速率之间存在显著的负相关性(Pearson相关系数$r=-0.91$)。基于有效长度的一元线性回归模型达到了迄今为止对蛋白质折叠速率的最高预测准确度。结果表明,蛋白质折叠速率主要受限于少数关键残基参与的非局域相互作用的形成。研究结果不仅为蛋白质折叠动力学研究提供了新的见解,也为蛋白质工程和设计提供了重要的理论指导。

关键词:蛋白质折叠速率;非局域相互作用;关键残基;相关性;预测
中图分类号:Q61;Q-03 **文献标志码**:A

蛋白质经历不断折叠和解折叠转变,其速率决定了它们在体内的稳态并调节它们的生物学功能。因此,深入解析决定蛋白质折叠速率的因素具有根本意义[1]。尽管深度学习在一定程度上解决了蛋白质结构预测的问题[2],但基本问题仍然存在[3-4],即蛋白质如何从氨基酸序列中快速地选择天然结构仍待进一步研究。

已有的研究表明,蛋白质的链长[5-6]和天然态结构拓扑[7-11]是决定其折叠速率的两个主要方面。理论上,蛋白质链长是由链节的数量决定的,通常将氨基酸(Amino acid,AA)残基总数计算为蛋白质链长。然而,Ivankov等[12]考虑到α-螺旋在折叠早期快速独立形成,这种块结构应该使得折叠链的有效长度小于残基总数,检验结果也表明有效链长与折叠速率的相关性更高。Chang等[13]基于各种氨基酸在折叠动力学或热力学中贡献不同的观察,建议折叠有效单位是原始序列中的必需氨基酸。在包含95个蛋白质折叠速率的数据集上,经遍历搜索得到了由10个氨基酸{CDGLPSTVWY}组成的最优氨基酸集,由此定义的有效长度与折叠速率也有很好的相关性[13]。另一方面,天然态残基间非局域相互作用在影响蛋白质折叠速率方面也扮演了重要角色[10-11]。Campos等[14]的研究指出,在实践中为了更好地预测蛋白质折叠速率,寻找对每种蛋白质折叠过渡态至关重要的少数关键残基和相互作用可能很重要。然而,从统计学角度针对单个蛋白质寻找影响其折叠速率的关键残基和相互作用是困难的。Gromiha等[15]和Harihar等[16]的统计表明,不同结构类的蛋白质有着不同的残基间接触

* **收稿日期**:2025-03-08;**修回日期**:2025-04-04
基金项目:内蒙古自然科学基金项目(2024LHMS06018,2022LHMS03014);内蒙古自治区直属高校基本科研业务费项目(JY20250094)
作者简介:刘 洋(2000—),女,内蒙古扎兰屯人,2022级硕士研究生。E-mail:3245203096@qq.com
通信作者:吕 军(1973—),男,内蒙古乌拉特前旗人,教授,博士。主要从事理论物理和理论生物学研究。E-mail:lujun@imut.edu.cn

范围,全α类蛋白质偏好中程接触,而全β类蛋白质更偏好长程接触,$\alpha+\beta$类和α/β类则处于二者之间。因此,从结构类特异性角度去寻找影响其折叠速率的关键残基可能是一个可行方案。

基于以上考虑,本文按蛋白质结构类分别搜索天然态相互作用中的关键残基,进而将参与天然态相互作用的关键残基数定义为蛋白质有效长度,研究了有效长度对蛋白质折叠速率的决定作用。

1 材料与方法

1.1 蛋白质折叠速率数据

在PFDB数据库[17](http://lee.kias.re.kr/~bala/PFDB)的145个蛋白质折叠速率数据基础上,综合文献[18-20]的数据,去掉冗余,整理得到一个包含156个单结构域水溶性球蛋白质折叠速率的数据集。其中,全α类蛋白质有42个,全β类蛋白质有56个,$\alpha+\beta$类蛋白质有43个,α/β类蛋白质有15个。蛋白质折叠速率均为纯水中的实验值,记为k_f,单位为s^{-1}。折叠最快的蛋白质是长度为21 aa、人工设计的丙氨酸短肽,速率约为$5.4\times10^{6}\ s^{-1}$,接近折叠速率限[19]。折叠最慢的蛋白质是长度为397 aa的色氨酸合成酶β亚单位(TrpB,PDB代码2DH5),速率约为$1\times10^{-3}\ s^{-1}$。由于不同蛋白质折叠速率的最大差异达到9个数量级,因此,为了方便常常使用折叠速率的自然对数值进行研究,如无特别说明,下文所述折叠速率均指折叠速率的自然对数值。蛋白质的结构数据下载自PDB数据库(https://www.rcsb.org/)[21],二级结构用DSSP程序[22](https://swift.cmbi.umcn.nl/gv/dssp/index.html)分配。蛋白质结构类数据来源于SCOP 2[23]数据库(https://www.ebi.ac.uk/pdbe/scop/)。

1.2 有效长度的定义

给定结构类为C(=α,β,$\alpha+\beta$,α/β)的一个蛋白质,当其序列上第i个残基a_i同时满足:

条件一,残基a_i的C_α原子与序列上任一残基的C_α原子之间的空间距离小于$D_{cut}(C)$,且二者序列间隔大于$S_{cut}(C)$,这里$D_{cut}(C)$和$S_{cut}(C)$是接触阈值;

条件二,残基a_i是C类蛋白质关键残基集$A(C)$的成员。

则认为残基a_i在热力学或动力学上是重要的,能够显著影响该蛋白质的折叠速率,否则认为残基a_i对折叠速率影响很小,可以忽略。由此定义蛋白质的有效长度N_{eff}为

$$N_{eff}=\sum_{i=1}^{L}n_i,\ n_i=\begin{cases}1, & a_i\text{满足条件一和条件二}\\ 0, & \text{otherwise}\end{cases},$$

其中L为蛋白质链长,即氨基酸残基数。上述定义中存在3个待定参数$D_{cut}(C)$、$S_{cut}(C)$和$A(C)$。

1.3 结构类依赖的最优参数的确定

依据过渡态理论,蛋白质折叠速率的自然对数$\ln k_f$与折叠自由能垒呈线性关系,而折叠自由能垒与链长或有效长度呈幂律依赖关系[5-6,12-13],因此折叠速率预测方程为

$$\ln k_f=\beta_0+\beta_1(N_{eff})^{\nu} \tag{1}$$

这里取$\nu=0.5$,这与Thirumalai[5]基于玻璃动力学模型给出的链长与折叠速率的关系一致。

为了确定结构类C的最优接触阈值$D_{cut}(C)$和$S_{cut}(C)$以及最优关键残基集$A(C)$,在结构类C的蛋白质折叠速率数据集上,对接触阈值及关键残基集进行遍历搜索。$D_{cut}(C)$的搜索范围为5.5~9.0 Å,步长为0.5 Å,$S_{cut}(C)$的搜索范围为3~13 aa,步长为1 aa,$A(C)$的搜索范围为所有可能的氨基酸类型组合(总组合数为$C_{20}^{1}+C_{20}^{2}+\cdots+C_{20}^{20}=2^{20}-1$)。对于每一步搜索,均基于方程(1)采用最小二乘拟合方法,得到自代回回归的决定系数R^2和折叠速率的预测值与观测值之间的平均绝对误差(Mean absolute error,MAE)。最优结果由最大的R^2和较小的MAE来确定。

决定系数R^2和平均绝对误差MAE定义为

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2},$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\hat{y}_i\right|,$$

这里y_i是第i个蛋白质折叠速率的观测值，$\hat{y}_i$是预测值，$\bar{y}$是y_i的平均值，N为样本数。

2　结果与讨论

2.1　结构类依赖的最优接触阈值以及关键残基集

采用遍历搜索方法得到各结构类对应的最优接触阈值以及关键残基集如表1所示。由表1可知，不同结构类的最优参数不同。全α类$S_{\mathrm{cut}}(\alpha)$最小，仅为4 aa，属于中程接触，相比之下，全β类和$\alpha+\beta$类主要由长程相互作用主导，$S_{\mathrm{cut}}(\beta)=S_{\mathrm{cut}}(\alpha+\beta)=12$ aa，α/β类也属长程接触，$S_{\mathrm{cut}}(\alpha/\beta)=6$ aa。从关键残基集角度看，全α类为{CEHMTW}，$\alpha+\beta$类为{DFGHMST}，二者交集为{HMT}；全β类为{GKLPSWY}，α/β类为{IKMNPSWY}，二者交集为{KPSWY}。结果表明，全α类与$\alpha+\beta$类有更多相似，全β类与α/β类更为接近，而全α类与全β类仅有色氨酸W残基为二者的公共残基，且其接触阈值相差也很大。这一结果可能暗示出，α类蛋白质与β类蛋白质有着迥然不同的折叠动力学，也表明关键残基集有着较强的结构类特异性。

如果在定义有效长度时不做结构类划分，对于全体蛋白质的寻优结果为$D_{\mathrm{cut}}=7$ Å和$S_{\mathrm{cut}}=9$ aa，A={CIMLPSTVWY}。显然，不论是接触阈值还是关键残基集，对全体蛋白质的寻优结果均取了4个结构类的折中。关键残基集中{CIMLVW}为疏水残基，众所周知，疏水效应是蛋白质折叠的主要驱动力之一[1]。{Y}是苯环＋羟基结构，羟基主导极性而苯环辅助疏水性，整体表现为两亲性。{ST}为柔性残基，侧链中单键较多、结构较小，通常很灵活。{P}既不疏水，也缺乏柔性，但其也在关键残基集中，因为脯氨酸可能会在吡咯烷环的两种主要状态之间经历缓慢地异构化，进而显著地减慢蛋白质折叠[13]。平均而言，全β类比全α类蛋白质折叠更慢，α/β类比$\alpha+\beta$类折叠更慢，而脯氨酸P也入选到了全β类和α/β类的关键残基集中，未入选到全α类和$\alpha+\beta$类，这显然支持了关键残基是对折叠动力学很重要的氨基酸的观点。

由表1数据还可以看到，采用自代回方法得到了各结构类预测值与观测值之间的决定系数和平均绝对误差。α/β类有最大的决定系数$R^2=0.92$和最小的平均绝对误差$MAE=0.73$，原因可能是两方面的。一方面，α/β类蛋白质多属于多态折叠动力学，而多态折叠蛋白质的速率对大小有更强的依赖性[8]，我们所定义的有效长度已经捕获了该类蛋白质的绝大部分折叠动力学特征。另一方面，α/β类蛋白质数据量偏少。对全β类蛋白质折叠速率的预测精度最差，$R^2=0.69$，$MAE=1.56$，表明对于全β类蛋白质而言，有效长度仅捕获该类蛋白质不足70%的折叠速率信息，还有超过30%的信息是有效长度所不能解释的。此外，有效长度对全α类和$\alpha+\beta$类蛋白质折叠速率的解释度分别为79%和76%。由于有效长度并没有捕获蛋白质的局域接触信息，而局域接触可能对蛋白质折叠速率有一定的贡献[7,11,16]。因此，有效长度在解释全α类、全β类以及$\alpha+\beta$类蛋白质折叠速率时，还存在约20%～30%的不确定性，这些不确定性可能来源于有效长度定义的不完整性以及实验数据的测量误差等方面。

2.2　蛋白质折叠速率与有效长度的相关性

对于任意给定结构类的蛋白质，基于表1的参数计算其有效长度N_{eff}，在当前数据集上，折叠速率的自然对数与有效长度平方根之间的关系如图1所示。图1结果显示，蛋白质有效长度的平方根与折叠速率的自然对数高度负相关，皮尔森相关系数$r=-0.91$（双边t检验$P<0.001$），95%的置信区间（Confidence interval，CI）为（−0.93，−0.88）。采用自代回方法，得到由表1给出的在各结构类样本集上对应的回归决定系数和平均绝对误差，对于全体156个蛋白质，基于结构类特异参数，可以得到回归决定系数$R^2=0.83$，平均绝对误差$MAE=1.47$。由于有效长度N_{eff}定义为特定结构类蛋白质中参与天然态相互作用的关键残基数，因此可以得出结论，蛋白质折叠速率受限于有少数关键残基

参与的残基间非局域相互作用的形成,且关键残基具有结构类依赖性。换句话说,蛋白质天然态相互作用中的关键残基数决定折叠速率,该结论与最近Campos等[14]的研究观点一致。

表1 依赖于蛋白质结构类的最优接触阈值及关键残基集的搜索结果

Table 1 Search results for optimal contact cutoffs and key residue sets dependent on structural classification of proteins

结构类C	样本数/个	平均折叠速率/s^{-1}	$D_{cut}(C)$/Å	$S_{cut}(C)$/aa	$A(C)$	R^2	MAE
α	42	7.88	7.5	4	CEHMTW	0.79	1.46
β	56	2.60	6	12	GKLPSWY	0.69	1.56
$\alpha+\beta$	43	3.47	7	12	DFGHMST	0.76	1.35
α/β	15	−0.44	6	6	IKMNPSWY	0.92	0.73
All	156	3.97	7	9	CIMLPSTVWY	0.75	1.72

如果在定义有效长度时不做结构类划分,我们发现如此定义的有效长度的平方根与折叠速率的自然对数之间也有较高的负相关性,$r=-0.87$(双边t检验$P<0.001$),95% 的CI为(−0.90,−0.82),决定系数和平均绝对误差数据在表1中给出。这一结果再次表明,天然态非局域相互作用中的关键残基数是蛋白质折叠速率的决定因素。然而,不进行结构类划分时也得到较高的线性相关系数,那么进行结构类划分的必要性是否还存在?也就是说,结构类特异的相关系数−0.91比非结构类特异的相关系数−0.87是否显著地小(相关性显著地大)?为此我们进行了两个重叠相关系数之间的单边差异性检验,结果显示,结构类特异的相关性显著大于非结构类特异的相关性($t=-3.39$,$P<0.0004$)。该结果显示,如果能够定义蛋白质特异的有效长度,可能会实现蛋白质折叠速率更高精度的预测,正如Campos等[14]所指出的,寻找对每种蛋白质折叠过渡态至关重要的少数关键残基和相互作用可能很重要。然而,针对单个蛋白质寻找影响其折叠速率的关键残基和相互作用在统计上是困难的。因此,我们给出的基于结构类的方法应该是一个有效的折中方案。

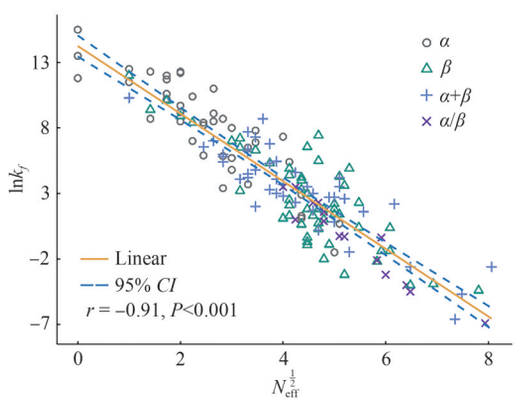

图1 蛋白质折叠速率自然对数与有效长度平方根之间的相关性

Fig. 1 Correlation between the natural logarithm of the protein folding rate and the square root of the effective length

此外,在当前数据集上搜索到的非结构类特异性关键残基集{CIMLPSTVWY}与Chang等[13]给出的最优氨基酸集{CDGLPSTVWY}有2个不同的残基,表明关键残基集可能存在一定的数据集依赖性。尽管如此,我们看到有效长度方法可以大幅改善折叠速率的预测精度,说明该研究方法是可行的。为了确定关键残基集对数据集的依赖性强度,我们也进行了数据的随机采样分析,结果表明,随着折叠速率数据集的扩大,关键残基集所包含的残基类型也逐渐趋于稳定。

2.3 与典型折叠速率预测模型的比较

在156个蛋白质折叠速率数据集上采用jackknife方法,基于方程(1)应用最小二乘拟合参数β_0和β_1,并预测折叠速率,预测性能评价指标为决定系数R^2和平均绝对误差MAE。选择截至目前已发表的9个典型的折叠速率预测模型进行性能比较。其中链长$L^{1/2}$模型[5],绝对接触序(Absolute contact order,ACO)模型[8],长程序(Long-range order,LRO)模型[9],折叠链有效长度(Effective length of folding chain,L_{eff})模型[12],基于组分有效链长(Composition-based effective chain length,n_e)模型[13]和折叠子漏

斗模型(Foldon funnel model,FFM)[19],这6个模型在折叠速率预测研究过程中不同时间节点上具有突出的理念性价值和高预测性能。有3个模型分别为有效累积扭角(Effective cumulative torsion angles,$CBTA_{eff}$)模型[24]、耦合二级结构数和非局域相互作用(Coupling of secondary structure number and nonlocal interaction,CSNI)模型[25]、多因素耦合参数(Multi-factor coupled parameters,C_p)模型[26],是近年来由我们研究组提出的高预测性能的模型。所有模型的jackknife预测结果见表2。

表2　在当前数据集上典型蛋白折叠速率预测模型的性能比较

Table 2　Comparison of the performance of typical protein folding rate prediction models on the current dataset

评价指标	模型									
	本文模型	$L^{1/2}$[5]	ACO[8]	LRO[9]	L_{eff}[12]	n_e[13]	FFM[19]	$CBTA_{eff}$[24]	CSNI[25]	C_p[26]
MAE	**1.48**	2.31	2.21	2.43	1.90	2.02	2.03	1.88	1.83	1.55
R^2	**0.82**	0.57	0.60	0.53	0.69	0.67	0.67	0.71	0.72	0.79

如表2所示,在当前数据集上,本文提出的集成了非局域结构拓扑、蛋白质大小以及氨基酸组分信息的有效长度模型获得最优的预测结果,R^2值达到0.82。而仅包含蛋白质大小信息的链长模型R^2值仅为0.57,包含了蛋白质大小和二级结构信息的L_{eff}模型R^2值为0.69,包含了蛋白质大小和氨基酸组分信息的n_e模型R^2值为0.67,仅包含非局域结构拓扑信息的LRO模型R^2值为0.53,包含了蛋白质大小和结构拓扑信息的ACO模型R^2值为0.60。FFM模型[19]是基于第一性原理的一个热力学加动力学模型,该模型以二级结构为折叠单元,有清晰的物理图像,提供了对折叠路径的理解,其R^2值为0.67。$CBTA_{eff}$、CSNI和C_p这3个模型有超过0.7的拟合质量,其中C_p模型表现最好,R^2值达到0.79,但C_p模型的参数构造复杂度高于本文模型,而拟合质量却低于本文模型。

3　结论

在深度学习解决蛋白质结构预测的时代,如何更好地了解决定蛋白质折叠速率的因素,发现隐藏在蛋白质序列中的折叠密码具有重要意义。影响蛋白质折叠动力学的因素是复杂的,而对于一个实际的复杂系统,在无法获知其完整动力学方程的情况下,通过掌握少量控制变量来预测其行为往往是可行的[27]。考虑到残基间的非局域相互作用和有效折叠单元数量是影响蛋白质折叠速率的关键因素,本文通过定义有效长度将这两个方面的因素综合在一起,结果证实蛋白质天然态中参与非局域相互作用的关键残基数量决定了折叠速率。尽管本文现有的结果可能仍存在一定的数据集依赖性,但所呈现的研究方法具有一定的启发性和可行性。

参考文献:

[1] NASSAR R,DIGNON G L,RAZBAN R M,et al. The protein folding problem:The role of theory[J]. Journal of Molecular Biology,2021,433(20):167126.

[2] JUMPER J,EVANS R,PRITZEL A,et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature,2021,596(7873):583-589.

[3] OUTEIRAL C,NISSLEY D A,DEANE C M. Current structure predictors are not learning the physics of protein folding[J]. Bioinformatics,2022,38(7):1881-1887.

[4] CHEN S J,HASSAN M,JERNIGAN R L,et al. Protein folds vs. protein folding:Differing questions,different challenges[J]. Proceedings of the National Academy of Sciences of the United States of America,2023,120(1):e2214423119.

[5] THIRUMALAI D. From minimal models to real proteins: Time scales for protein folding kinetics[J]. Journal de Physique, 1995, 5(11): 1457-1467.

[6] NAGANATHAN A N, MUÑOZ V. Scaling of folding times with protein size[J]. Journal of the American Chemical Society, 2005, 127(2): 480-481.

[7] 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.

[8] IVANKOV D N, GARBUZYNSKIY S O, ALM E, et al. Contact order revisited: Influence of protein size on the folding rate[J]. Protein Science, 2003, 12(9): 2057-2062.

[9] 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.

[10] DMITRII E M, PLAXCO K W. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics[J]. Protein Science, 2003, 12(1): 17-26.

[11] WANG J, PANAGIOTOU E. The protein folding rate and the geometry and topology of the native state[J]. Scientific Reports, 2022, 12(1): 6384.

[12] 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 of the United States of America, 2004, 101(24): 8942-8944.

[13] CHANG L, WANG J, WANG W. Composition-based effective chain length for prediction of protein folding rates[J]. Physical Review E, 2010, 82(5 Pt 1): 051930.

[14] CAMPOS L A, MUÑOZ V. Targeting the protein folding transition state by mutation: Large scale (un) folding rate accelerations without altering native stability[J]. Protein Science, 2024, 33(7): e5031.

[15] GROMIHA M M, SELVARAJ S. Influence of medium and long range interactions in different structural classes of globular proteins[J]. Journal of Biological Physics, 1997, 23(3): 151-162.

[16] HARIHAR B, SARAVANAN K M, GROMIHA M M, et al. Importance of inter-residue contacts for understanding protein folding and unfolding rates, remote homology, and drug design[J]. Molecular Biotechnology, 2025, 67(3): 862-884.

[17] MANAVALAN B, KUWAJIMA K, LEE J. PFDB: A standardized protein folding database with temperature correction[J]. Scientific Reports, 2019, 9(1): 1588.

[18] GARBUZYNSKIY S O, IVANKOV D N, BOGATYREVA N S, et al. Golden triangle for folding rates of globular proteins[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(1): 147-150.

[19] ROLLINS G C, DILL K A. General mechanism of two-state protein folding kinetics[J]. Journal of the American Chemical Society, 2014, 136(32): 11420-11427.

[20] SUBRAMANIAN S, GOLLA H, DIVAKAR K, et al. Slow folding of a helical protein: Large barriers, strong internal friction, or a shallow, bumpy landscape?[J]. The Journal of Physical Chemistry B, 2020, 124(41): 8973-8983.

[21] BURLEY S K, BHIKADIYA C, BI C, et al. RCSB protein data bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D[J]. Protein Science, 2022, 31(1): 187-208.

[22] KABSCH W, SANDER C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers, 1983, 22(12): 2577-2637.

[23] ANDREEVA A, KULESHA E, GOUGH J, et al. The SCOP database in 2020: Expanded classification of repre-

sentative family and superfamily domains of known protein structures[J]. Nucleic Acids Research, 2020, 48(D1): D376-D382.

[24] LI Y R, ZHANG Y, LV J. An effective cumulative torsion angles model for prediction of protein folding rates[J]. Protein and Peptide Letters, 2020, 27(4): 321-328.

[25] 徐素杰,张颖,吕军. 耦合二级结构数和非局域相互作用预测蛋白质折叠速率[J]. 内蒙古工业大学学报(自然科学版),2021,40(2):92-100.

[26] 李兰,张颖. 基于多因素耦合参数拟合蛋白质折叠速率[J]. 内蒙古工业大学学报(自然科学版),2023,42(2):103-108.

[27] LUO L F, LÜ J. Data-driven prediction in complex systems of virus evolution and global warming[J]. 内蒙古大学学报(自然科学版),2025,56(1):1-7.

(责任编委　刘俊杰)

Total Number of Key Residues Involved in Native-State Interactions Determined Folding Rates of Proteins

LIU Yang, WANG Kun, ZHANG Ying, LÜ Jun

(*College of Science*, *Inner Mongolia University of Technology*, *Hohhot* 010051, *China*)

Abstract: Understanding the factors influencing protein folding rates is essential for elucidating the mechanisms underlying protein folding. A novel method is proposed to identify the critical factors affecting protein folding rates. This approach introduces a parameter termed "effective length", which quantifies the number of key residues involved in non-local interactions within the native state, and is specific to the structural classification of proteins. Validation on a benchmark dataset demonstrates a strong negative correlation between effective length and protein folding rates (Pearson correlation coefficient, $r=-0.91$). A univariate linear regression model based on effective length achieves the highest prediction accuracy for protein folding rates to date. The results suggest that protein folding rates are primarily constrained by the formation of non-local interactions involving a limited number of key residues. These findings provide new insights into protein folding dynamics and offer valuable theoretical guidance for protein engineering and design.

Key words: protein folding rate; non-local interaction; key residue; correlation; prediction