文章编号:1000－1638(2025)02-0174-13 DOI:10.13484/j.nmgdxxbzk.20250208

基于随机森林预测蛋白质变体折叠速率*

张非凡,张 颖,马莹雪,吕 军
(内蒙古工业大学理学院,呼和浩特 010051)

摘要:蛋白质单点突变引起的折叠速率变化的准确预测,对于探索序列如何编码折叠这一蛋白质折叠的基本问题有积极意义。搜集了 1329 个由实验测定的蛋白质单点突变体折叠速率数据,并采用 AlphaFold2 预测了所有变体的结构数据。为了比较不同模型之间的预测性能,选出其中 190 个变体作为盲测集,其余作为训练集。按照突变位点所处一级结构位置(N 端、中间和 C 端)、二级结构位置(螺旋、股和其他)以及三级结构位置(暴露、埋藏和部分埋藏)的不同,将变体蛋白分别归类到 27 个类别中。提取了残基的物化性质、取代得分以及接触势等 1325 个序列和结构特征。首先基于随机森林算法在每个类别的训练集上对特征进行重要性排序,并分别选择最优的 3 个特征,进一步将这些选出的特征再次输入到随机森林回归模型对变体相对于野生型的折叠速率改变量进行预测。结果表明,在盲测集上预测值与实验值之间的皮尔逊相关系数为 0.403,平均绝对误差为 0.613,优于现有的最好模型。

关键词:蛋白质单点突变;折叠速率;氨基酸性质;结构性质;随机森林算法
中图分类号:Q61;Q-03 **文献标志码**:A

在蛋白质合成过程中,特定的三维结构通常在翻译过程中或翻译完成后逐渐形成。蛋白质折叠速率差异极大,研究表明这种差异高达 9 个数量级[1]。在过去的研究中,研究者们提出多种基于蛋白质大小[2-4]、拓扑结构[5-7]、氨基酸组成[8-11]等方面的模型。这些模型在预测效果上表现出高度可靠性。然而,预测点突变对蛋白质折叠速度和稳定性影响的模型尚处于发展初期,该研究方向意义重大。首先,有效运用计算预测工具能大幅节省实验所需时间。更重要的是,它能作为设计感兴趣突变体的初步步骤,随后在实验中进行验证。此外,氨基酸替换等突变会影响蛋白质的多项特征,这些突变所带来的影响往往是不利的,如会导致亨廷顿病、朊病和阿尔茨海默病等。准确预测这些变化不仅有助于更好地理解疾病本质,而且在合成和设计蛋白质突变体方面具有潜在的重要价值,有望为治疗这些疾病提供新策略。

预测由单点突变引起的蛋白质变体折叠速率变化,相对于预测野生型蛋白质折叠速率更为复杂。因为单点突变对蛋白质整体大小影响微乎其微,并且对其三维结构造成的影响通常也较为有限。然而,即使是在氨基酸序列中单一氨基酸改变这种细微变化,也可能引起折叠速率显著改变,这些都增加了对突变蛋白质折叠动力学预测的复杂性。

以往研究中,Gromiha 研究小组使用了相同或者略有增加的数据集开发了多个预测工具,他们应

* **收稿日期**:2024-07-08;**修回日期**:2024-09-19
基金项目:内蒙古自然科学基金项目(2022LHMS03014,2024LHMS06018);内蒙古直属高校基本科研业务费项目(JY20220069);内蒙古自治区直属高校基本科研业务费项目(JY20250094)
作者简介:张非凡(1999－),男,山东惠民人,2022 级硕士研究生。E-mail:995730834@qq.com
通信作者:张 颖(1973－),女,辽宁昌图人,副教授,博士。主要从事计算生物学研究。E-mail:yzhang@imut.edu.cn

用的算法包括 FORA[12] 和 FREEDOM[13] 中的二次回归、KD-FREEDOM[14] 中基于规则的决策树、Folding RaCe[15] 中的多元线性回归以及 Unfolding RaCe[16] 中的氨基酸属性和多元线性回归。此外，Mallik 等[17] 提出了一个基于残基水平的共进化信息模型，使用相对共进化序参数预测蛋白质单点突变后折叠速率改变量。在 PON-Fold[18] 方法中，使用了机器学习方法，该方法用梯度提升算法实现。

郝冬磊[19] 研究发现，蛋白质变体折叠速率绝对增量的平均值，随突变残基在蛋白质中埋藏深度（使用相对溶剂可及性度量）不同而存在显著差异。Schwersensky 等[20] 研究显示，相比内部区域突变，表面区域突变蛋白质通常表现出更高的稳健性，即对功能和结构影响较小，尤其在小蛋白质中尤为明显，这揭示了蛋白质结构和功能对内部氨基酸序列的敏感性。

通过突变位点在野生型蛋白质中的相对溶剂可及性、二级结构以及序列位置分类，将变体划分为 27 个不同类别，采用随机森林算法预测单点突变对蛋白质折叠速率的影响，与现有预测工具相比展现出更优越的预测性能。

1　材料与方法

1.1　数据

1.1.1　蛋白质突变体数据集

基于数据库 K-Pro[21] 和文章 PON-Fold[18]，共收集 1329 个由实验测得的蛋白质单点突变体折叠速率数据，记为“PF1329”。数据集包含 43 种野生型蛋白质单点突变体，各蛋白质突变体数量从 2 到 68 个不等，该数据集是迄今为止使用过最大的变体蛋白数据集之一。大部分折叠速率的测量是使用停流荧光仪进行的，一些还利用了连续流动荧光仪、温度跳变荧光仪或停流圆二色谱。43 种野生型蛋白质结构数据下载自 PDB 数据库（https://www.rcsb.org/）[22]，突变体结构信息通过 AlphaFold2[23] 预测得到，它们的二级结构用 DSSP 程序分配[24]。蛋白质单点突变折叠速率改变量为 $\Delta\ln k_f=\ln k_f^{\mathrm{mut}}-\ln k_f^{\mathrm{wt}}$，其中 k_f^{mut} 和 k_f^{wt} 分别为突变型和野生型折叠速率。数据集上突变氨基酸类型的分布情况见图 1。

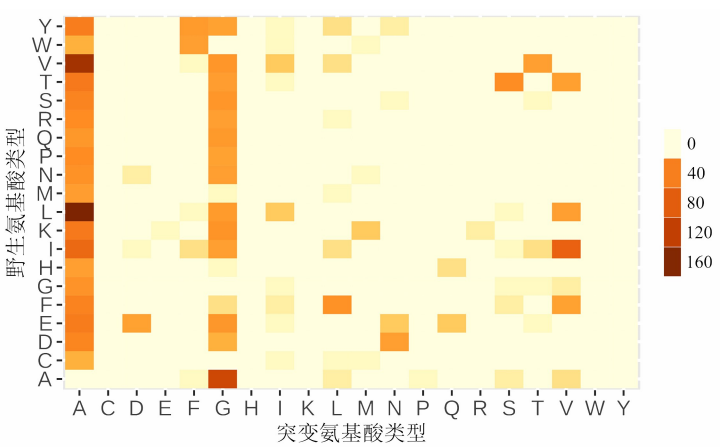

图 1　点突变氨基酸类型的分布

Fig. 1　Distribution of amino acid types for point mutations

如图 1 所示，数据集对某些残基有偏差，丙氨酸是最常见变异残基，占总变异残基 3/5 左右，其他用来替代的常见残基还有甘氨酸和缬氨酸。这种分布上的偏差部分原因是丙氨酸扫描突变分析在研究氨基酸侧链效应时广泛应用，加之这 3 种氨基酸体型较小，替换时不太可能引起蛋白质结构不稳定，使它们成为突变分析中的常见选择。对于一些氨基酸的取代，则没有这种情况，如半胱氨酸、组氨酸、赖氨酸、色氨酸、酪氨酸。造成这一现象的原因可能是这些氨基酸结构更为复杂，且在蛋白质结构

和功能中扮演着关键角色,替代这些氨基酸可能会对蛋白质稳定性造成较大影响。所有氨基酸类型都有发生变异,尽管数量差距很大,其中亮氨酸、缬氨酸、异亮氨酸、丙氨酸是最常改变的氨基酸类型。

1.1.2 数据集分类

由于突变位点在野生型蛋白质中不同位置对突变效应的差异性,根据突变位点所处一级结构(序列位置)(N,≤33%; M,33%－67%; C,≥67%),二级结构(H,S,O),三级结构(B,ASA≤20%; P,20%<ASA≤50%; E,ASA>50%)的位置对突变体进行分类,共产生 27 种类别。这种分类方法确保了在每个类别中突变效应相似,这对于确定统一规律性是有益的。每个类别根据突变位点所处位置进行命名,例如,突变位于序列位置≤30%、α-螺旋且 ASA≤20%的变体蛋白类型记为 NHB,其他类别以相同的方式命名。"PF1329"数据集分类后每一种类别中折叠速率改变量分布情况如图 2 所示。

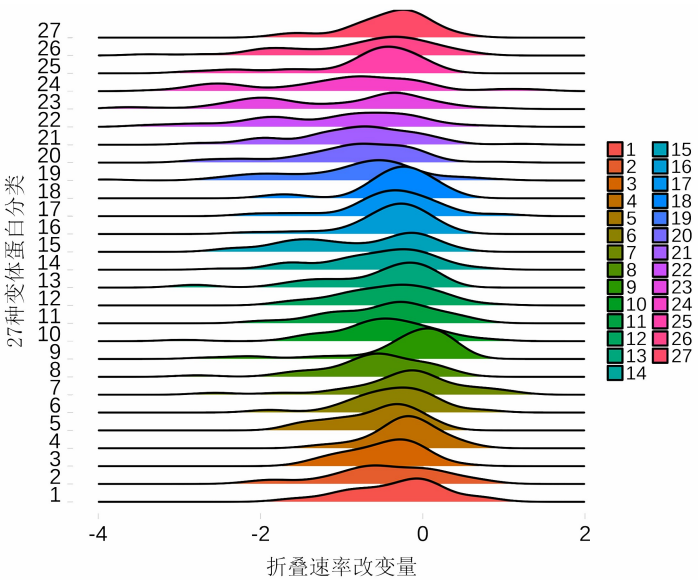

图 2 27 类变体折叠速率改变量分布的山脊图

Fig. 2 Ridge plot of distributions of folding rate changes for the 27 variant types

分析图 2 发现,变体蛋白质折叠速率改变量分布与分类情况紧密相关。例如,具有相同二级结构 S 或 O 的 13－15、22－24 分类和 16－18、25－27 分类中,尽管具有相同二级结构的两组类别分别处于不同三级结构 P 和 B,但具有相同二级结构的种类的折叠速率增量表现出高度相似性,在相同二级结构 S 下的 13－15、22－24 分类中,它们的折叠速率增量密度曲线呈现出类似的双峰密度曲线。这些类别折叠速率改变量分布范围广泛,但峰值幅度相对较低。在具有相同二级结构 O 的 6－18、25－27分类中,它们的折叠速率改变量都呈现出平坦而广阔的单峰密度分布,同样具有折叠速率改变量分布范围广泛、但峰值幅度相对较低的特点。可见在三级结构 P 和 B 分类中,二级结构的影响是主要的。在具有相同三级结构 E 的 1－9 分类中,折叠速率改变量与二级结构变化之间并未显现出明显关联性,反映出相同三级结构 E 在这一情况下扮演更为重要的角色,并在密度曲线上体现出高度一致性。通过对整体数据集分析发现,折叠速率受到一级结构影响较小,二级和三级结构的影响更为显著。

在研究的绝大多数类别中,突变蛋白折叠速率改变量呈现出良好的正态分布特征。这一发现表明,尽管突变体包含不同类型氨基酸变异,但它们引起的折叠速率变化量在统计学上遵循一定的规律性。另外,相较于导致折叠速率提高的突变,折叠速率降低的突变在幅度上显得更加多变和广泛。这种分布可能与氨基酸突变对蛋白质三维结构造成的微弱扰动有关。根据蛋白质折叠理论[25],这些扰动可能会增加蛋白质的自由能,从而阻碍其正确折叠。另外,过往研究亦表明,由氨基酸突变导致的

局部结构不稳定性可能是造成折叠速率减缓的一个常见原因。通过深入分析这些数据，可以更好地了解单点突变对蛋白质功能的影响，从而为理解疾病机制和开发新的治疗方法提供有力的信息，这些洞见在蛋白质工程和药物设计等领域具有潜在的应用价值。

在对“PF1329”数据集进行训练集和盲测集划分时，考虑模型评估的准确性和模型对比的公平性，选取与 PON-Fold 方法中相同的盲测集，其他数据作为训练集。对训练集和盲测集中蛋白质单点突变的折叠速率改变量进行了分析，如表 1 所示。训练集中折叠速率增量的最小值为−5.23，折叠速率增量的中位数为−0.47，折叠速率增量的最大值为 2.61。在盲测集中，折叠速率增量的最小值为−4.01，折叠速率增量的中位数为−0.34，折叠速率增量的最大值为 1.13。通过分析这些数据，观察到单点突变往往导致蛋白质折叠速率显著降低，且这种现象发生范围更广、数量更多。相比之下，变体蛋白折叠速率提升范围较小，数量也较少。

表 1　训练集和盲测集的折叠速率增量

Table 1　Folding rate changes for training and blind test sets

数据集	数量	最小值	中位数	最大值
训练集	1139	−5.23	−0.47	2.61
盲测集	190	−4.01	−0.34	1.13

1.2　特征

1.2.1　氨基酸性质

从 AAindex 数据库（www.genome.jp/dbget-bin/www_bfind? aaindex）[26] 获取了共 707 种氨基酸属性。这一数据库被分为 3 个主要部分：AAindex1 包含了 566 种关于氨基酸各类物理化学指数的记录，例如疏水性、结构偏好性和电荷量等；AAindex2 提供了 94 种氨基酸相似性指数，用于评估序列比对中两个氨基酸的相似性得分；AAindex3 记录了 47 种氨基酸接触势，即两个氨基酸接触时的潜在能量指数。此外，我们参考 Gromiha[27] 的研究，从中获取了另外 49 种氨基酸属性，称为 T49（表 2）。

综合以上数据来源，累计得到 756 种氨基酸属性。这些丰富的数据包括了对本研究至关重要的物理化学特性、热力学参数、进化信息、结构倾向性和氨基酸间的相互作用势。对于 AAindex 数据库以及 T49 的处理方法如下所述：

（1）AAindex1 和 T49：AAindex1 和 T49 数据库记录了氨基酸的多种物化性质。对于特定突变位点，通过计算该位点突变氨基酸与相应野生型氨基酸相同性质的差值，得到差异性质 $\Delta P_1 = P_{\mathrm{mut}} - P_{\mathrm{wt}}$，其中 P_{mut} 表示突变位点氨基酸性质，P_{wt} 表示该位点野生型氨基酸性质。

（2）AAindex2：AAindex2 包含 94 种氨基酸相似性指数，这些指数形成一个矩阵，用以评估两个氨基酸相似度。对于特定突变位点，从该矩阵中检索并记录野生型氨基酸（A_{wt}）与突变氨基酸（A_{mut}）对应的相似性指数，并将此数值表示为 $\Delta S = S(A_{\mathrm{wt}} \rightarrow A_{\mathrm{mut}})$。

（3）AAindex3：AAindex3 提供了 47 种氨基酸接触势性质，它们测量蛋白质中氨基酸残基间的接触能量。本研究中，设定空间距离上界 $R_{\mathrm{cut}} = 8$Å 和序列距离下界 $l_{\mathrm{cut}} = 12$ 作为氨基酸间接触标准。基于这些界限，分别对突变位点的野生型和突变型蛋白质进行了计算，以获取相应的氨基酸接触势，并计算二者的差值，记为 $\Delta V = \sum_{i}^{N_1} P_i^{\mathrm{mut}} - \sum_{j}^{N_2} P_j^{\mathrm{wt}}$。这里，$P^{\mathrm{mut}}$ 表示突变型蛋白质在突变位点与满足条件的氨基酸的接触势，P^{wt} 表示野生型蛋白质与满足条件的氨基酸的接触势，N_1 和 N_2 分别表示野生型和突变型在指定条件下具有接触的氨基酸数目。

表 2 49 种氨基酸性质

Table 2 49 types of amino acid properties

序号	特征编号	描述	序号	特征编号	描述
1	K0	可压缩性	26	am	处在 α-螺旋中间的能量
2	Ht	热力学转移疏水性	27	V0	氨基酸特定部分体积
3	Hp	周围的疏水性	28	Nm	平均中程接触
4	P	极性	29	Nl	平均长程接触
5	pHi	等电点	30	Hgm	结合的周围疏水性(球状与膜状)
6	pK	参考羧基(COOH 基团)的电离性质的平衡常数	31	ASAD	变性蛋白的溶剂可接触表面积
7	Mw	分子量	32	ASAN	天然蛋白的溶剂可接触表面积
8	Bl	体积	33	DASA	蛋白质展开的溶剂可接触表面积
9	Rf	色谱指数	34	DGh	展开过程的水合吉布斯自由能变化
10	Mu	折射指数	35	GhD	变性蛋白的水合吉布斯自由能变化
11	Hnc	标准化共识疏水性	36	GdN	天然蛋白的水合吉布斯自由能变化
12	Esm	短程和中程非键合能量	37	DHh	展开过程的水合焓变
13	El	长程非键合能量	38	－TDSh	展开过程的水合熵变
14	Et	总的非键合能量	39	DCph	展开过程的水合热容变化
15	Pa	α-螺旋倾向性	40	DGc_1	链展开过程的吉布斯自由能变化
16	Pb	β-螺旋倾向性	41	DHc_1	链展开过程的焓变
17	Pt	回转倾向性	42	－TDSc_1	链展开过程的熵变
18	Pc	线圈倾向性	43	DGc_2	展开过程的吉布斯自由能变化
19	Ca	螺旋接触面积	44	DHc_2	展开过程的焓变
20	F	均方根波动位移的平均值	45	－TDSc_2	展开过程的熵变
21	Br	埋藏性	46	v	体积(非氢侧链原子的数量)
22	Ra	溶剂可接触降低比例	47	s	形状(侧链中分支点的位置)
23	Ns	周围残基的平均数量	48	f	柔韧性(侧链二面角的数量)
24	an	处在 α-螺旋 N 端的能量	49	Pf-s	主链二面角概率
25	ac	处在 α-螺旋 C 端的能量			

在分析氨基酸性质改变量时,我们遇到一个问题:相同类型突变产生了完全相同的改变值,但却拥有不同的折叠速率改变量。这表明,除单个氨基酸的属性之外,邻域特征(即残基周围的环境)的影响也非常关键。因此,采用 AAindex1 数据库提供的 566 种氨基酸特征,计算的能量随着每个突变位置周围环境不同而变化的邻域特征。为了量化相邻残基的影响(记为 ΔP_{seq}),采用了一个窗口长度变化从 3 到 19 个残基的方法,ΔP_{seq}公式表示为

$$\Delta P_{\text{seq}} = P_{\text{mut}}(i) - \left[\left(\sum_{j=i-k}^{j=i+k} P_j(i)/(2k+1)\right)\right],$$

式中,P_{mut}为突变型氨基酸性质;P 为野生型蛋白质中氨基酸性质;i 为突变位点在蛋白质序列中的位置。窗口长度由 k 值确定,k 在两个方向上(突变位点的两侧)从 0 变化到 9。当 $k=0$ 时,窗口仅包含突变位点本身,不包含任何邻近残基;随着 k 值增加,所用窗口长度相应增长,例如 $k=1$ 时,使用

3 个残基窗口长度，以此类推。对于那些在原始数据中具有缺失值的特性，选择使用所有变体蛋白中该特性的平均值来补充这些缺失值。这一方法能够综合考虑突变本身的性质与其在蛋白质结构中邻域关系中的相关性，提高了研究突变对蛋白质功能影响的分辨能力。

1.2.2　蛋白质的结构特征

对于研究中涉及的变体蛋白质，使用 AlphaFold2 预测其结构数据。为定量分析这些变体结构的特征变化，采用 R 语言和相应生物信息学工具，帮助计算变体中特定结构参数的变化[9,28-30]。

(1) 组分约束二级结构序：$cSSO=\frac{1}{2N}\sum_{i,j}^{L}\delta_{ij}$，其中，$N$ 为二级结构数，L 为蛋白质链长，即氨基酸总数，i 和 j 为不同二级结构片段上的残基在蛋白质链上的序号，δ_{ij} 为接触数函数。d_{ij} 为两个残基的空间距离，残基位置由 C_a 原子的位置代替，距离单位为 nm，则 δ_{ij} 定义为 $\delta_{ij}=\begin{cases}1,\text{if}|i-j|>l_0 \text{ and } d_{ij}<d_0\\0,\text{otherwise}\end{cases}$，其中，$l_0$ 和 d_0 为接触阈值，l_0 为 i 残基与 j 残基的序列间隔阈值，d_0 为空间距离阈值。可以看出，cSSO其实是长程序 LRO 的二级结构版本。

(2) 绝对接触序：$ACO=\frac{1}{N}\sum_{i}^{N}\Delta S_{i,j}$，其中，$N$ 是接触总数，如果两个残基所包含的任意非氢原子之间的空间距离小于等于 6Å，就认为这两个残基存在一个接触(两原子间的空间距离不大于 R_{cut}，序列距离或残基间隔不小于 l_{cut}，这里定义 $R_{\text{cut}}=6$Å，$l_{\text{cut}}=1-2$)，$\Delta S_{i,j}$是接触残基 i 和残基 j 之间的序列间隔($\Delta S_{i,j}=|j-i|$)，nACO$=\log$(ACO)。

(3) 累积主链扭角：$CBTA=\sum_{i=1}^{L}|\phi_i|$，这里，$\phi_i$ 是主链上第 i 个主链扭角，由蛋白质的结构数据计算得到，L 为蛋白质的链长，“||”表示取绝对值。可以看到，CBTA 大体上和蛋白质的链长成比例，因此这个参数具有链长的属性，此外 CBTA 还具有蛋白质拓扑(形状)属性。

1.3　特征选择

1.3.1　初筛特征

通过氨基酸性质和蛋白质结构特征的整合，共获得 1325 条有关变体蛋白的特征信息。为防止过多特征对模型训练产生不利影响，我们执行一系列特征筛选步骤剔除冗余及相关性不高的特征。具体操作流程为：(1)特征分类，将所有特征依据性质类型划分为 5 大类——AAindex1 的物理化学指标、AAindex2 的氨基酸相似性指数、AAindex3 的接触能量指数、蛋白质序列的邻域特征(ΔP_{seq})和蛋白质的结构特征；(2)单一值过滤，在 5 个分类中分别移除那些单一值占比超过 50%的特征，即那些在大多数样本中保持不变的特征，因为它们提供的信息量极少且对模型的预测能力几乎没有帮助；(3)降低冗余，对每一类中的特征进行成对的相关性分析，以便识别高度相关(相关系数大于 0.8)的特征。对高度相关的特征集，只保留那些与预测目标相关性最强的特征，以减少数据冗余，保留最有价值的信息；(4)选取高相关特征，在每个特征类别中选取前 10 个与目标预测值相关性最高的特征。通过这种方式，5 个类别共筛选出 50 个候选特征。通过上述 4 步筛选过程，有效地从 1325 个特征中提炼出 50 个与目标关联度最高且冗余度最低的关键特征集合。

1.3.2　精选特征

图 3 展示处理训练集及初筛特征流程。首先，对数据集进行组织归类。随后，对每一分类应用随机森林算法对初筛特征进行重要性评分，确定每项特征的相对重要性。基于重要性评分，从各分类中分别筛选出前 3 名的关键特征，这些精选特征接下来将被用于构建预测模型。

最终 27 个分类共选取 74 个特征(表 3)，值得关注的是，在这 74 个特征中，ISOY800107、FODM020101、QIAN880123、WERD780102、CHAM820102、KARP850101 和 KARS160120 都被选中两次，这 7 个性质分别为双弯曲的规范化相对频率、氨基酸在 pi-螺旋中的倾向性、窗口位置 3 处 β-折叠的权重、ε(i)到ε(ex)的自由能变化、溶于水中的自由能、对于没有刚性邻居的灵活性参数和基于

原子数的加权最小特征值。可见氨基酸的结构倾向性、自由能、灵活性参数和加权最小特征值在单点突变对蛋白质造成的影响中发挥重要作用。

表 3 27 类变体蛋白特征筛选结果

Table 3 Feature selection results for 27 variants of proteins

序号	特征编号	描述	变体种类	特征来源
1	AURR980102	α 螺旋末端 N 处的残基位置标准化频率	NOE	AAindex
2	CHOP780211	C 端非 β 区域的标准化频率	MHP	AAindex
3	PALJ810114	所有 β 类中转角的标准化频率	MSE	AAindex
4	ISOY800107	双弯曲的规范化相对频率	NSP、NOP	AAindex
5	CHOP780213	转角中第二个残基的频率	NHB	AAindex
6	FODM020101	氨基酸在 pi-螺旋中的倾向性	MOE、CSB	AAindex
7	RACS820103	特定结构环境中结构偏好程度	CHP	AAindex
8	RACS770103	侧链取向偏好程度	MHE	AAindex
9	GEOR030102	单跨膜蛋白连接数据集中氨基酸连接肽倾向性	NHP	AAindex
10	GEOR030105	小数据集中的连接体倾向性	MHE	AAindex
11	GEOR030101	全数据集中的连接体倾向性	NSP	AAindex
12	OOBM850101	优化的 β-结构-线圈平衡常数	MSP	AAindex
13	PUNT030102	跨膜螺旋环境中的偏向性	NSE	AAindex
14	QIAN880120	窗口位置 0 处的 β-折叠的权重	MSB	AAindex
15	QIAN880109.1	基于窗口位置 2 处的 α 螺旋权重计算的领域特征	CSP	ΔP_{seq}
16	QIAN880123	窗口位置 3 处的 β-折叠的权重	COE、CSB	AAindex
17	QIAN880139	窗口位置 6 处的无规则卷曲权重	NHP	AAindex
18	WERD780102	$\varepsilon(i)$到 $\varepsilon(ex)$的自由能变化	CHB、NSB、MOB	AAindex
19	WERD780103	氨基酸残基从 Ri 转变到 Rh 时伴随的自由能	CHE	AAindex
20	RADA880103	气相到环己烷相的转移自由能	MOE	AAindex
21	CHAM820102	溶于水中的自由能	NHB	AAindex
22	CHAM820102.1	基于溶于水中的自由能计算的领域特征	MSB	ΔP_{seq}
23	OOBM850104	每个原子优化的平均非键合能量	NHP	AAindex
24	ZIMJ680101	氨基酸的亲疏水性	MSE	AAindex
25	Hnc	标准化共识疏水性	CHB	T49
26	NADH010107	基于两状态模型(50%可接触性)的自信息值的亲水性等级	CSP	AAindex
27	EISD860102	基于原子的疏水矩	COB	AAindex
28	ROSM880103	通过螺旋形成导致侧链疏水性的损失	COP	AAindex
29	KLEP840101	净电荷	COB	AAindex
30	FAUJ880112	氨基酸中的负电荷量	NSE	AAindex
31	HUTJ700101	氨基酸的热容量	CSB	AAindex
32	GEIM800106	β-蛋白的 β-链指数	COP	AAindex
33	OOBM850105	优化的侧链相互作用参数	MHP	AAindex
34	RACS770101	C-α 的平均减小距离	MHB	AAindex
35	ANDN920101	C-α 化学位移	CHE	AAindex

表 3(续)

序号	特征编号	描述	变体种类	特征来源
36	Pf-s	主链二面角概率	NOP	T49
37	pK	参考羧基(COOH 基团)电离性质的平衡常数	MOP	T49
38	s	形状(侧链中分支点的位置)	NHB	T49
39	LEVM760106	范德华参数 R0	MHP	AAindex
40	FAUJ880101	氨基酸形状指数	MHB	AAindex
41	CORJ870106	ALTLS 指数	NOB	AAindex
42	FAUJ880104	侧链的 STERIMOL 长度	MHB	AAindex
43	KARP850101	对于无刚性邻居的灵活性参数	NHE、CHB	AAindex
44	WEBA780101	高盐层析中的 RF 值	CSP	AAindex
45	HOPA770101	水合数	NSE	AAindex
46	DAYM780201	相对突变性	NSB	AAindex
47	ROBB760105	扩展信息测量	MSP	AAindex
48	LEVM760105	侧链的回旋半径	NSB	AAindex
49	MEEJ800101	HPLC 中的保留系数,pH 7.4	NOE	AAindex
50	KARS160120	基于原子数的加权最小特征值	MHE、COP	AAindex
51	WOLS870103	主要属性值	MOB	AAindex
52	AVBF000108	构成十肽的斜率	COE	AAindex
53	VASM830103	构象状态 E 的相对数量	COE	AAindex
54	RACS820101	A0(i)中的平均相对分数出现率	NSP	AAindex
55	RACS820106	ER(i)中的平均相对分数出现率	NOE	AAindex
56	NAKH920101	单跨膜蛋白 CYT 的氨基酸组成	CSE	AAindex
57	NAKH920103	单跨膜蛋白 EXT 的氨基酸组成	MOP	AAindex
58	MEHP950102	α-螺旋中的残基交换权重矩阵	CSE	AAindex
59	MEHP950103	β-链中的残基交换权重矩阵	CSE	AAindex
60	CSEM940101	残基替换的能力矩阵	MOP	AAindex
61	QU_C930101	主链偏好因子的交叉相关系数矩阵	MSE	AAindex
62	DOSZ010101	氨基酸相似性矩阵	NHE	AAindex
63	GIAG010101	从嗜热、中温菌到嗜冷菌的残基替代矩阵	MOE	AAindex
64	DOSZ010103	基于 THREADER 力场的氨基酸相似矩阵	NHE	AAindex
65	RISJ880101	结构相关蛋白中的替换矩阵	CHE	AAindex
66	BONM030103	氨基酸侧链之间相互作用的统计势	NOP	AAindex
67	SIMK990105	距离依赖的统计势	COB	AAindex
68	TOBD000102	基于大量伪装结构获得的优化衍生势	MSB	AAindex
69	SIMK990104	10～12 Å 距离依赖的统计势	MSP	AAindex
70	SKOJ000101	氨基酸对间相互作用的统计势	NOB	AAindex
71	MIYS850102	氨基酸从水转移到蛋白质环境的准化学转移能	NOB	AAindex
72	ΔcSSO	组分约束二级结构序	CHP	文献[7]
73	ΔnACO	绝对接触序的自然对数	MOB	文献[7]
74	ΔCBTA	累积主链扭角 CBTA	CHP	文献[29]

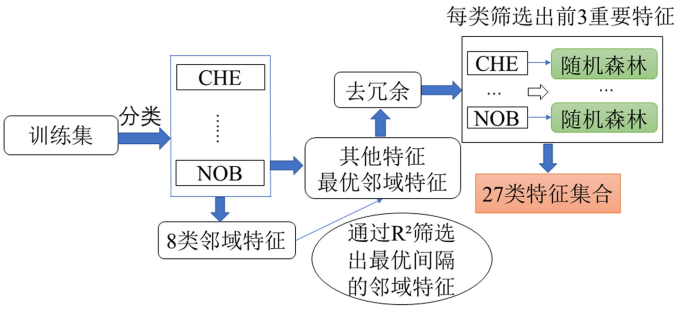

图 3 特征选择流程图

Fig. 3 Flowchart of feature selection

1.4 模型

1.4.1 随机森林模型

本研究选用随机森林算法作为构建预测模型的主要方法。随机森林是一种基于多个决策树构建而成的集成学习技术,它通过集合众多决策树的预测结果来提高整个模型的预测准确度和泛化能力。此算法的一个显著优点是它能够高效处理不同种类的数据集,并且在执行多样化的预测任务时表现出色,显示了它的强大适应性和处理能力。

图 4 展示模型训练和应用流程,使用训练集进行 10 折交叉验证确保模型的鲁棒性和可靠性。训练完成后,对盲测集采用相同分类方法,并根据分类结果匹配对应特征集进行预测。最后,每个类别数据分别通过训练完成的随机森林模型进行预测分析。

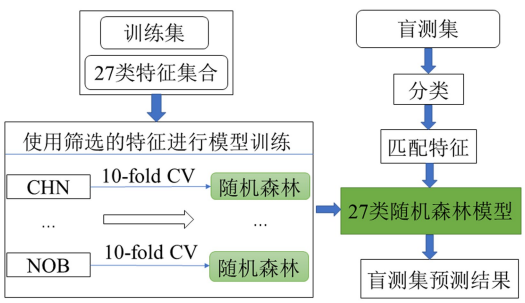

图 4 随机森林模型的训练及应用过程

Fig. 4 Training and application process of random forest model

2 结果与讨论

采用 10 折交叉验证来评估使用 5 个和 3 个特征进行模型训练的性能差异,并运用 4 种评价指标进行综合比较。测试结果表明,在训练集上,基于 5 个特征构建的模型在所有选用的评价指标上均优于使用 3 个特征的模型(表 4)。

表 4 特征选择与评估指标结果对比

Table 4 Comparison of feature selection and evaluation metrics results

评价指标	训练集		盲测集	
	5 个特征	3 个精选特征	5 个特征	3 个精选特征
PCC	0.875	0.844	0.381	0.403
MSE	0.193	0.226	0.718	0.694
MAE	0.298	0.331	0.610	0.613
R^2	0.735	0.691	0.102	0.132

然而,盲测集上的预测结果表明,采用 3 个特征的模型实际展示出更优的性能(表 4)。这表明使

用 5 个特征的模型在面对训练数据时可能过分拟合数据，没有掌握到足以泛化到未知数据的通用规律。鉴于上述测试结果，最终模型构建中选用 3 个关键特征，旨在创建一个既高效又具有泛化能力的预测模型。

最终该模型在面对较为困难的预测任务以及有限并且带有偏差的数据集时，依然展现出令人满意的性能。如图 5 所示，在盲测集测试中，模型训练结果的皮尔森相关系数(PCC)达到 0.403，均方误差(MSE)和平均绝对误差(MAE)分别为 0.613 和 0.694。值得注意的是，某些变异类型在训练数据中出现较少或完全没有，这在一定程度上限制模型的性能。

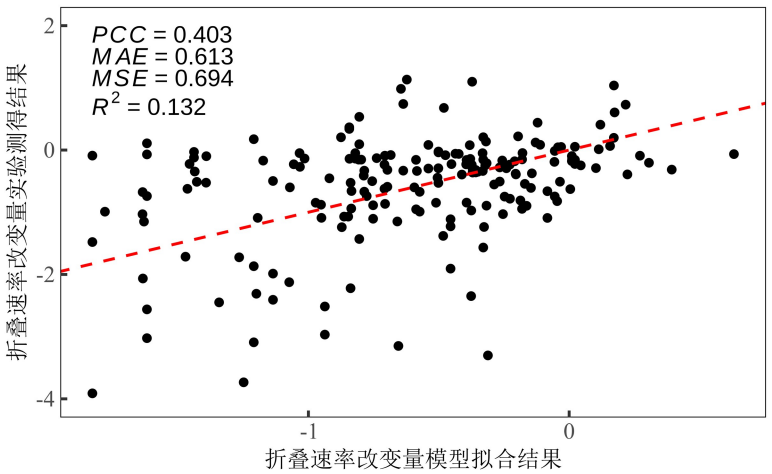

图 5　盲测集上不同模型的预测性能比较

Fig. 5　Comparison of predictive performance among different models on blind test sets

与现有的几种用于变体蛋白折叠速率预测模型进行比较时，只有两种模型：PON-Fold[18] 和 Folding RaCe[15] 可用于比较，其他工具要么不可用，要么不利于大规模的预测。在特征选取上，Folding RaCe 使用相对溶剂可及性、二级结构信息和序列位置等因素，通过多元线性回归建立模型来预测折叠速率的变化。PON-Fold 则通过从 1161 个特征筛选出 21 个重要的特征，使用这些特征来训练模型，这些特征覆盖了氨基酸的相对溶剂可及性、序列中位置和 C 分数、各种氨基酸的性质及邻域特征等。在本研究中，不仅与现有模型进行比较，还将训练集和盲测集应用XGBoost算法进行训练和预测分析。结果显示，与随机森林模型相比，XGBoost 在预测准确性上表现不佳，且所需训练时间更长。基于这些发现，最终选择将随机森林模型作为本研究的核心工具，并在文章中对其进行深入讨论。盲测集选取中，沿用与 PON-Fold 和 Folding RaCe 相同的变体蛋白质集，以保证对比结果的准确性和公平性，模型预测对比结果如表 5 所示。

表 5　随机森林与 XGBoost、PON-Fold 和 Folding RaCe 性能对比

Table 5　Performance comparison of random forest, XGBoost, PON-Fold, and Folding RaCe

评价指标	随机森林	XGBoost	PON-Fold	Folding RaCe
PCC	0.403	0.180	0.330	0.170
MAE	0.613	0.708	0.672	0.952
MSE	0.694	1.000	0.817	1.632
R^2	0.132	−0.250	−0.021	−1.040

在相同盲测集下,随机森林模型相对于 XGBoost、PON-Fold 和 Folding RaCe 模型展示出相对较好的性能。PCC 分别提高 124.89%、22.12%和 137.06%,MAE 和 MSE 分别下降了 13.43%、30.60%、8.77%和 15.02%、35.61%、57.45%。在训练集中使用 1139 个变体蛋白的信息,相比 PON-Fold 和 Folding RaCe 的 762 和 734 组成的训练集,拥有的变体蛋白数据集更大更复杂。因此,我们的模型更准确地说明导致蛋白变体折叠速率改变的关键特征是性能提升的主要原因。

对变体蛋白性质筛选过程中,发现将变体蛋白基于突变位点位置进行细分极为关键。以 AAindex 数据库 FINA910101 性质为例,未分类的 1329 个变体蛋白样本集中,该性质与变体蛋白折叠速率变化的 PCC 仅为 0.13。而在执行细致分类后,与变体蛋白折叠速率变化的 PCC 显著增加至0.50。这表明,变体蛋白分类能够显著提高预测变体蛋白折叠速率变化的准确度。

从筛选结果看,氨基酸结构倾向性、自由能、电荷量和疏水性是预测变体蛋白折叠速率改变量的重要因素。此外,邻域特性和蛋白质结构信息也在调控变体蛋白折叠速率改变中起到重要作用,特别是识别同类突变中不同变体的折叠速率变化时,邻域特征和蛋白质结构信息具有决定性作用。这是因为同类突变中氨基酸性质变化是相同的,只依赖氨基酸性质改变无法对同类突变中不同变体折叠速率变化进行区分,因此,邻域特征和结构信息的改变提供了一个区分这些折叠速率差异的有效途径。

在现有研究基础上,进一步考虑诸如蛋白质空间形态、蛋白质序列顺序的变化情况以及突变氨基酸对邻近氨基酸影响等,这些因素对变体蛋白折叠速率改变可能均有贡献,从而影响到变体蛋白折叠速率。如何增加这些信息来提升模型对蛋白质突变导致折叠速率变化的预测性能,进而探索蛋白质单点突变所引起变体蛋白折叠速率改变的本质原因,除考虑特征选取之外,探索更加精细和科学的变体蛋白分类方法也是十分值得关注的问题。从上述可知,新特征选取以及变体蛋白分类方式优化都可能是变体蛋白研究领域的关键着力点。

3 结论

蛋白质单点突变引起折叠速率变化是复杂的生物化学问题,折叠速率是理解蛋白质折叠机制的关键探针。已有研究表明,蛋白质氨基酸性质以及拓扑结构改变是引起蛋白质折叠速率改变的关键因素。随机森林是机器学习中一种强大的分类和回归方法,该方法将样本数据多样性、树的拓扑及属性的重要性耦合起来,形成综合判决结果,采用多树结果投票得到最终结论。采用随机森林对变体蛋白折叠速率改变量进行预测,结果表明,随机森林对单点突变导致的折叠速率变化有较高的拟合精度。

参考文献:

[1] 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.

[2] FINKELSTEIN A V,BADRETDINOV A. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold[J]. Folding & Design,1997,2(2):115-121.

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

[4] 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.

[5] PLAXCO K W,SIMONS K T,BAKER D. Contact order,transition state placement and the refolding rates of sin-

gle domain proteins[J]. Journal of Molecular Biology,1998,277(4):985-994.

[6] ZHANG L X,SUN T T. Folding rate prediction using *n*-order contact distance for proteins with two- and three-state folding kinetics[J]. Biophysical Chemistry,2005,113(1):9-16.

[7] 李彦儒. 从有限构象搜索空间角度预测蛋白质折叠速率[D]. 呼和浩特:内蒙古工业大学,2020.

[8] 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[J]. Proteins-Structure Function and Bioinformatics,2006,65(2):362-372.

[9] HUANG L T,GROMIHA M M. Analysis and prediction of protein folding rates using quadratic response surface models[J]. Journal of Computational Chemistry,2008,29(10):1675-1683.

[10] 高建召,胡刚,王奎,等. 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.

[11] 高建召. 基于序列的蛋白质折叠速率与膜蛋白功能分类研究[D]. 天津:南开大学,2010.

[12] HUANG L T,GROMIHA M M. Real value prediction of protein folding rate change upon point mutation[J]. Journal of Computer-Aided Molecular Design,2012,26(3):339-347.

[13] HUANG L T,GROMIHA M M. First insight into the prediction of protein folding rate change upon point mutation[J]. Bioinformatics,2010,26(17):2121-2127.

[14] HUANG L T. Finding simple rules for discriminating folding rate change upon single mutation by statistical and learning methods[J]. Protein and Peptide Letters,2014,21(8):743-751.

[15] CHAUDHARY P,NAGANATHAN A N,GROMIHA M M. Folding RaCe:A robust method for predicting changes in protein folding rates upon point mutations[J]. Bioinformatics,2015,31(13):2091-2097.

[16] CHAUDHARY P,NAGANATHAN A N,GROMIHA M M. Prediction of change in protein unfolding rates upon point mutations in two state proteins[J]. Biochimica et Biophysica Acta(BBA)-Proteins and Proteomics,2016,1864(9):1104-1109.

[17] MALLIK S,DAS S,KUNDU S. Predicting protein folding rate change upon point mutation using residue-level coevolutionary information[J]. Proteins,2016,84(1):3-8.

[18] YANG Y,CHONG Z,VIHINEN M. PON-Fold:Prediction of substitutions affecting protein folding rate[J]. International Journal of Molecular Sciences,2023,24(16):13023.

[19] 郝冬磊. 蛋白质单点突变折叠速率改变的统计分析[D]. 呼和浩特:内蒙古工业大学,2016.

[20] SCHWERSENSKY M,ROOMAN M,PUCCI F. Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness[J]. BMC Biology,2020,18(1):146.

[21] TURINA P,FARISELLI P,CAPRIOTTI E. K-Pro:Kinetics data on proteins and mutants[J]. Journal of Molecular Biology,2023,435(20):168245.

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

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

[24] TOUW W G,BAAKMAN C,BLACK J,et al. A series of PDB-related databanks for everyday needs[J]. Nucleic Acids Research,2015,43(D1):364-368.

[25] ANFINSEN C B. Principles that govern the folding of protein chains[J]. Science,1973,181(4096):223-230.

[26] KAWASHIMA S C,POKAROWSKI P,POKAROWSKA M,et al. AAindex:Amino acid index database,progress report 2008[J]. Nucleic Acids Research,2008,36(Suppl 1):202-205.

[27] GROMIHA M M. A statistical model for predicting protein folding rates from amino acid sequence with structural class information[J]. Journal of Chemical Information and Modeling,2005,45(2):494-501.

[28] IVANKOV D N,GARBUZYNSKIY S O,ALM E,et al. Contact order revisited:Influence of protein size on the

folding rate[J]. Protein Science: A Publication of the Protein Society, 2003, 12(9): 2057-2062.

[29] LIANG H, WANG L L, ZHANG Y, et al. Prediction of protein folding rates from the amino acid sequence-predicted backbone torsion angles[J]. Letters in Organic Chemistry, 2017, 14(9): 643-654.

[30] 胡秀珍. 蛋白质规则二级结构中亲疏水氨基酸紧邻关联特性[J]. 内蒙古大学学报(自然科学版), 2002, 21(4): 395-400.

(**责任编委** 刘俊杰)

Predicting the Folding Rate of Protein Variants Based on Random Forest

ZHANG Feifan, ZHANG Ying, MA Yingxue, LÜ Jun

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

Abstract: The accurate prediction of changes in protein folding rates caused by single-point mutations is of positive significance for exploring the fundamental question of how sequences encode protein folding. We collected 1329 experimentally measured protein single-point mutant folding rate data and used AlphaFold2 to predict the structural data of all variants. To compare the predictive performance between different models, 190 variants were selected as a blind test set, with the remainder used as the training set. Variants were categorized into 27 classes based on the location of the mutation site in the primary structure (N-terminal, middle, and C-terminal), secondary structure (helix, strand, and others), and tertiary structure (exposed, buried, and partially buried). A total of 1325 sequence and structure features were extracted, including the physicochemical properties of residues, substitution scores, and contact potentials. Based on the random forest algorithm, features were first ranked by importance on the training set of each category and the top 3 features were selected, which were then re-entered into the random forest regression model to predict the changes in folding rate of variants relative to the wild-type. The results showed that the Pearson correlation coefficient between the predicted values and experimental values on the combined blind test set was 0.403, and the mean absolute error was 0.613, which is superior to the existing best model.

Key words: protein single-point mutation; folding rate; amino acid property; structural property; random forest algorithm