计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(2):180−185 [doi: 10.15888/j.cnki.csa.006220]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多重进化矩阵的蛋白质特征向量构造方法①

杜月寒, 鹿文鹏, 刘毅慧, 成金勇

(齐鲁工业大学 (山东省科学院) 信息学院, 济南 250353)

摘　要: 特征向量的构造是蛋白质二级结构预测的一个关键问题. 现有的研究方法, 通常只使用 BLOSUM62 进化矩阵生成 PSSM 矩阵, 对蛋白质进化过程中存在的氨基酸残基突变现象缺乏考虑. 本文提出利用多重进化矩阵构造蛋白质特征向量, 其融合了不同进化时间的 PSSM 矩阵, 不仅能够很好地反映序列中氨基酸的位置信息, 而且能够反映序列进化过程中氨基酸位点发生突变产生的影响. 本文通过组合不同进化程度的矩阵来构造特征向量, 选用逻辑回归、随机森林和多分类支持向量机三种分类算法作为预测工具, 利用网格搜索法和交叉实验法优化参数, 在 RS126、CB513 和 25PDB 公用数据集上进行了若干组实验. 对比实验结果表明, 本文所提出基于多重进化矩阵的蛋白质特征向量构造方法能够有效提高蛋白质二级结构的预测精度.

关键词: 蛋白质结构预测; 多重进化矩阵; 逻辑回归; 随机森林; 多分类支持向量机

引用格式: 杜月寒,鹿文鹏,刘毅慧,成金勇.基于多重进化矩阵的蛋白质特征向量构造方法.计算机系统应用,2018,27(2):180–185. http://www.c-s-a.org.cn/1003-3254/6220.html

Protein Secondary Structure Prediction Based on Multiple Evolutionary Matrix

DU Yue-Han, LU Wen-Peng, LIU Yi-Hui, CHENG Jin-Yong

(School of Information, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China)

Abstract: The construction of feature vector is a key issue for protein secondary structure prediction. In the present methods, only the BLOSUM62 matrix is taken into account, which neglects the amino acid mutation of protein in the evolutionary process. In this study, we propose to construct feature vector by combining PSSM matrices of different evolutionary times, which cannot only reflect the position information, but also reflect the interaction of amino acids. Based on the feature vector, logistics, randomforest and M-SVM$_{CS}$ models are utilized to predict protein secondary structure on the public datasets (RS126, CB513, and 25PDB). The experimental result demonstrates that the method can achieve a better performance than traditional methods.

Key words: protein secondary structure prediction; multiple evolutionary matrix; logistics; randomforest; M-SVM$_{CS}$

蛋白质是生物体内生命活动的主要承担者, 是一切生命活动的基础, 它的生理功能除了体现在氨基酸构成上还体现在它的空间结构上[1]. 因此, 预测蛋白质结构是生物信息学领域的一个重要任务. 通常, 蛋白质结构包括 4 个层次[2]: 一级结构即氨基酸的排列顺序; 二级结构主要是由氢键维持的 α-螺旋和 β-折叠; 三级结构是完全折叠的蛋白质的空间结构残基的立体排列模式; 四级结构是多个蛋白质亚基组成的蛋白质复合体的结构 (即蛋白质之间的交互作用). 蛋白质二级结构是联系蛋白质一级结构和三级结构的纽带, 而且也是从一级结构预测其三级结构的关键步骤[3,4]. 当蛋白质二级结构预测正确率达到 80% 时, 就可以准确预测一个蛋白质分子的三维空间结构[5]. 可见, 蛋白质二级结构预测已经成为研究蛋白质结构和功能的重要手段.

① 基金项目: 国家自然科学基金 (61375013,61502259); 山东省自然科学基金 (ZR2013FM020)
收稿时间: 2017-04-25; 修改时间: 2017-05-11; 采用时间: 2017-05-25

由于已测定结构的蛋白质数量远远小于已知的蛋白质序列数量[6], 并且传统的生物实验测定蛋白质结构的方法花费昂贵且耗时时间较长. 因此, 采用数据驱动的方法 (如机器学习技术) 来预测未知的蛋白质的结构和功能广受青睐. 在过去的一段时间内, 很多方法被提出来用于蛋白质结构类的预测. 而影响蛋白质结构类预测效果的关键因素主要集中在两个方面上: 一是分类预测算法, Zhou 等人使用神经网络[7], Mandle 等人使用支持向量机[8], Wang 和 Peng 等使用深度卷积神经网络技术来进行蛋白质结构预测[9]; 二是蛋白质特征信息提取, 如 Chou 等人提出的伪氨基酸组成 (PseAA 法)[10–12], Cao 等人提出的基于简化 PSSM 与蛋白质结构位置信息的特征表示算法[13].

一般的预测方法通常使用 BLOSUM62 矩阵构造特征向量, 对蛋白质进化过程中存在的氨基酸位点突变现象缺乏考虑. 本文提出一种新的特征表示方法, 对于一条蛋白质序列, 同时使用多种进化趋异度的矩阵来表示蛋白质序列, 更全面的考虑了残基替换的可能性. 不同的进化矩阵对不同相关程度的蛋白质序列的敏感性不同. 这使得多重进化矩阵这种蛋白质序列特征表示方法, 不仅可以很好地反映序列中氨基酸的位置信息, 而且全面考虑序列内部近相关和远相关蛋白质区域[14]之间的相互影响. 本文结合交叉验证法和网格搜索法来确定实验参数, 先在大范围大步距粗搜, 初步确定一个最优参数区间, 之后在此区间进行小步距精搜, 结合网格搜索法和交叉验证法共同确定实验参数. 在数据集 RS126、CB513 和 25PDB 上进行的多组实验, 表明本文所提出的基于多重进化矩阵的特征向量构造方法能够有效提高蛋白质二级结构预测精度.

1 相关知识

1.1 位置特异性矩阵

通过蛋白质序列的位置特异性打分矩阵而不是仅依靠序列来预测蛋白质结构, 是公认的提高预测精度的方法. 序列的多重比对反映了蛋白质家族的共同特征, 提取了结构的保守信息及家族中特定的残基替换模式, 同时多序列比对所携带的进化信息也表明了蛋白质进化过程中的相互作用[15].

本文主要采用了 BLAST 的本地化使用来获得蛋白质序列的 profile. 将 BLAST 软件包下载到本地以后, 可以通过命令行的形式去调用相应的可执行文件. 在这里, 我们使用 PSI-BLAST 程序 (h=0.001, j=3) 搜索和调整无冗余的 NR 数据库. 该程序将返回一个 20 维矢量的 PSSM[16], 其值是 20 个氨基酸保守的突变分数. 这样得到的 PSI-BLAST profile 是一个 $L\times20$ 的矩阵 (其中 L 是蛋白质序列的长度), 也称之为位置特异性打分矩阵 (Position-Specific Score Matrix, PSSM). PSSM 矩阵形式如公式 (1) 所示.

$$\mathrm{PSSM}_s=\begin{bmatrix} P_{1\to1} & P_{1\to2} & \cdots & P_{1\to j} & \cdots & P_{1\to20} \\ P_{2\to1} & P_{2\to2} & \cdots & P_{2\to j} & \cdots & P_{2\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i\to1} & P_{i\to2} & \cdots & P_{i\to j} & \cdots & P_{i\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{L\to1} & P_{L\to2} & \cdots & P_{L\to j} & \cdots & P_{L\to20} \end{bmatrix} \quad (1)$$

PSSM 矩阵的每一行代表在查询序列的相应位置发生在氨基酸替代的对数似然得分. 位于矩阵第 i 行第 j 列的元素 P_{ij} 表示在进化过程中查询序列的第 i 个位置的氨基酸突变成 j 类氨基酸的得分.

1.2 二级结构划分标准

蛋白质二级结构通常分为 8 类: G(3_{10}-helix), H(α-helix), I(π-helix), B(isolated β-bridge), E(β-stand), S(bend), T(hydrogen bonded turn) 和 rest(apparently random conformations). 主流的 PSSP 思想会将这 8 类结构归纳为 3 种构象 (H、E 和 C). 通常情况下, H、E 和 C 三种构象之间没有明确的界限而且也没有统一的标准去划分这三种构象. 然而在 1999 年, Cuff 和 Barton 两位学者证实了划分方案可以影响最后的预测精度, 所以人们希望找到一种划分方案可以获得更高的预测精度. 由此, 蛋白质二级结构字典法 (DSSP[17]) 获得了广泛的认可. 该方法依据已知的氢键相连的部分划分二级结构. 本文选用 DSSP 方法, 将 8 类结构明确归纳为: H、G 属于 Helices, 记作 H; E、B 属于 Sheets, 记作 E; G、S、T、C、I 属于 Coils, 记作 C.

2 基于多重进化矩阵的特征向量构造方法

在实际中, 蛋白质的结构是不断折叠式的, 某个残基不仅与它相邻的残基发生作用, 还可能与它在序列上相差较远的某些残基发生作用, 且蛋白质进化过程中氨基酸位点存在突变的可能. PAM 矩阵和 BLOSUM 矩阵就反映了蛋白质中存在的氨基酸突变.

因 PAM 矩阵和 BLOSUM 矩阵都是 PSI-BLAST

程序中的打分标准, 不同的打分矩阵对于评价氨基酸突变是不同的[18,19], 例如 PAM250 矩阵假设每 100 个氨基酸发生 250 次点突变, PAM 矩阵存在从 PAM1 到 PAM250 的情况. 由于 PAM 矩阵是基于近相关蛋白比对得到的打分矩阵, BLOSUM 矩阵是基于观测到的远相关蛋白对比得到的打分矩阵. 本文参考了 PAM 矩阵和 BLOSUM 矩阵之间的相互关系[20], 如图 1 所示, 设计了多重进化矩阵编码方式. 为更详细的描述氨基酸位点发生突变的可能性和序列内部近距离和远距离的氨基酸之间的相互影响, 选择了低趋异度矩阵 PAM30 和高趋异度矩阵 PAM250 和 BLOSUM62 矩阵三种不同趋异度的进化矩阵来表达蛋白质序列.

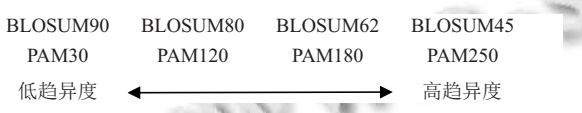

图 1 PAM 矩阵和 BLOSUM 矩阵概要

首先将蛋白质序列送入 PSI-BLAST 程序, 通过调整参数, 得到广泛使用的 BLOSUM62 矩阵、低趋异度矩阵 BLOSUM90 和高趋异度矩阵 PAM250. 将得到的三种不同趋异度的进化矩阵对齐特征维度, 组合得到 60 维的向量表示原来的蛋白质序列, 考虑临近残基的影响, 采用滑动窗口法对所得特征向量进行处理, 设置滑动窗口为 13, 得到一个 780 维向量表示原来的序列, 构成多重进化矩阵特征.

为了能够用计算方法进行训练和预测, 需要将相差较大的原始值进行规范化处理. 本文利用公式 (2) 把多重进化矩阵的元素标准化到 0~1 之间.

$$f(x)=\frac{1}{1+e^{-x}} \tag{2}$$

其中 x 是多重进化矩阵中元素的原始值.

3 蛋白质二级结构预测框架

为了在构造特征向量时能更好的反映蛋白质序列中氨基酸残基存在突变的可能性, 且考虑预测过程中存在分类器参数选择困难及可靠性差等问题, 本文提出基于多重进化矩阵的蛋白质二级结构预测方法, 其具体过程如下:

1) 首先要将 BLAST 本地化. 下载蛋白质 NR 数据库及 BLAST 程序本地软件包, 对 BLAST 进行本地配置.

2) 计算蛋白质序列的位置特异性打分矩阵 (PSSM) 矩阵, 设置 PSI-BLAST 程序的参数为 (-num_iterations: 3, -eavlue: 0.001, -matrix: BLOSUM62), 得到该参数条件下的 PSSM 矩阵.

3) 调整 PSI-BLAST 程序参数, 将 matrix 分别设置为 BLOSUM90 和 PAM250, 计算该参数条件下的 PSSM 矩阵.

4) 将 3) 中得到的三种进化矩阵对齐特征维度, 组合得到 60 维的向量来表示原来的蛋白质序列. 采用滑动窗口法处理向量, 设置滑动窗口为 13, 得到一个 780 维向量来表示原来的蛋白质, 构成多重进化矩阵特征, 对矩阵进行标准化处理.

5) 利用网格搜索法和 K 折交叉验证来优选实验参数. 选取强分类器多分类支持向量机 M-SVM$_{CS}$ 来说明实验过程:

① 设定网格搜索的变量 (c, p) 的范围以及搜索步距, 选择使分类准确率最高的一组 c 和 p;

② 在寻得了局部最优参数之后, 再在这组参数附近选择一个小区间, 采用小步距进行二次精搜, 再次选择使分类准确率最高的一组 c 和 p;

③ 涉及的所有参数对都用 7 折交叉验证进行实验, 按数据集条数平均分成 7 份, 每次选择其中 6 份做训练集, 剩下的 1 份做测试集, 重复 7 次;

④ 上述提到的分类准确率的参数对按照以下原则确定: 若参数选择过程中有多组 c 和 p 对应于最高的验证分类准确率, 则选取能够达到最高验证分类准确率中参数 c 最小的那组 c 和 p 作为最佳的参数; 如果对应最小的 c 有多组 p, 就选取搜索到的第一组 c 和 p 作为最佳参数对;

6) 按照 5) 中获得的最优参数模型, 输入结构未知的蛋白质序列特征, 预测各个位点残基二级结构.

4 实验与结果分析

4.1 数据集

为了检验模型的预测精确性, 选择数据集要慎重, 需要结合机器学习和生物学方面的知识. 伴随着 PDB 等主要蛋白质结构数据库中的蛋白质结构资源的日益丰富, 可用的蛋白质二级结构预测的样本也越来越多. 出于对实验结果的公平及公正性的考虑, 本文选择三个广泛应用的低同源性数据集 RS126[21]、CB513[22]和 25PDB[23]作为本文的实验数据集, 序列相似性均低于 25%. RS126 数据集含有 126 条非同源蛋

白质序列. CB513 数据集含有 513 条非同源蛋白质序列. 25PDB 数据集含有 1673 条非同源蛋白质和从 PDB 中下载和扫描的高分辨率结构域.

4.2 参数选择

为了证明本文提出的多重进化矩阵是一种有效特征向量表示方法, 本文选择了两种弱分类器 Logistics、RandomForest 和一种强分类器 MSVMpack 进行实验. 其中 Logistics 和 RandomForest 来自 WEKA 软件, M-SVM$_{CS}$ 来自 MSVMpack 软件[24]. 三种分类器都是通过网格搜索法来挑选实验参数. 为了对分类器参数进行优化, 且保证优化结果的可靠性, 本方法结合七折交叉验证与网格搜索法来确定实验参数.

经过多组实验, 对所获得的实验结果进行对比, 选择其中最好的一组作为最优参数. 对于数据集 CB513 和 25PDB 我们将针对不同分类算法得到的最优参数汇总如表 1 所示.

表 1 最优参数结果表

数据集	方法	Logistics	RandomForest	M-SVM$_{CS}$
RS126	BLOSUM62法	M: 20	I: 210 K: 10	c: 0.7, p: 10
	多重进化矩阵法	M: 13	I: 200 K: 30	c: 0.75, p: 16
CB513	BLOSUM62法	M: 15	I: 200 K: 30	c: 0.7, p: 12
	多重进化矩阵法	M: 15	I: 250 K: 30	c: 0.5, p: 20.5
25PDB	BLOSUM62法	M: 25	I: 250 K: 10	c: 0.5, p: 20
	多重进化矩阵法	M: 25	I: 250 K: 30	c: 0.4, p: 20

4.3 结果评价标准

关于蛋白质二级结构预测结果的评价标准有很多种. 目前在国际上大多使用以下几种标准:

(1) 整体预测准确率 Q_3

目前应用最广泛的准确率, 它指的是被正确预测的 3 种二级结构 (残基) 的总百分比, 可由公式 (3) 计算得出.

$$Q_i = \frac{P_{\mathrm{H}} + P_{\mathrm{E}} + P_{\mathrm{C}}}{N_{\mathrm{H}} + N_{\mathrm{E}} + N_{\mathrm{C}}} * 100\% \quad (3)$$

其中, N_{H}、N_{E} 和 N_{C} 分别表示序列中二级结构为 H、E 和 C 的残基的总个数, P_{H}、P_{E} 和 P_{C} 分别表示被正确预测为 H、E 和 C 构象的残基个数.

(2) 三态预测准确率 Q_i

我们用 Q_i 来表示每种二级结构被正确预测为 H, E 或 C 构象的预测准确率. 可由公式 (4) 计算得出:

$$Q_i = \frac{P_i}{N_i} * 100\%, i \in \{\mathrm{H}, \mathrm{E}, \mathrm{C}\} \quad (4)$$

其中, P_i 是待预测序列中被正确预测的处于 i 构象的残基数目, N_i 是待预测序列中被正确预测的处于 i 构象的残基数目, i 属于 H 构象、E 构象或 C 构象.

根据本文第 4 节的方法, 我们在 RS126、CB513 和 25PDB 数据集上进行实验. 在 RS126 数据集上三个独立分类器得到的整体预测准确率分别为 67.86%、67.90% 和 73.90%, 其各项独立指标如表 2 所示. 在 CB513 数据集上三个独立分类器得到的整体预测准确率分别为 65.53%、71.32% 和 75.50%, 其各项独立指标如表 3 所示. 在 25PDB 数据集上三个独立分类器得到的整体预测准确率分别为 68.57%、72.62% 和 76.72%, 其各项独立指标如表 4 所示.

表 2 RS126 数据集使用 BLOSUM62 矩阵预测结果 (%)

分类模型	Q_H	Q_E	Q_C	Q_3
Logistics	62.85	61.69	64.87	63.59
RandomForest	56.27	53.69	72.83	67.90
M-SVM$_{CS}$	74.48	61.17	80.21	73.90

表 3 CB513 数据集使用 BLOSUM62 矩阵预测结果 (%)

分类模型	Q_H	Q_E	Q_C	Q_3
Logistics	79.29	36.36	69.84	65.53
RandomForest	71.35	49.93	81.62	71.32
M-SVM$_{CS}$	79.01	63.16	79.20	75.50

表 4 25PDB 数据集使用 BLOSUM62 矩阵预测结果 (%)

分类模型	Q_H	Q_E	Q_C	Q_3
Logistics	78.84	51.87	69.17	68.57
RandomForest	74.16	54.17	81.50	72.62
M-SVM$_{CS}$	80.25	68.69	78.17	76.72

然后, 我们组合三种不同进化趋异度的矩阵, 作为三个独立分类器的输入向量, 通过网格搜索法和 7 折交叉法优选实验参数, 获得优化参数模型, 输入结构未知的蛋白质序列特征, 预测各个位点残基二级结构. 对数据集 RS126 使用三种分类器获得的整体预测准确率分别是 66.40%、68.08% 和 74.05%, 各类别的预测准确率如表 5 所示. 对数据集 CB513 使用三种分类器获得的整体预测准确率分别是 69.18%、71.89% 和 75.92%, 各类别的预测准确率如表 6 所示. 通过对比表 2 和表 5 可以看出, 相比于传统的实验方法, 多重进化矩阵这种表示方法在 RS126 数据集上分别高出了 -1.37%、0.18% 和 0.15%. 通过对比表 3 和表 6 可以看出, 相比于传统的实验方法, 多重进化矩阵这种表示方法在 CB513 数据集上分别高出了 3.65%、0.57% 和

0.42%. 而对于数据集 25PDB 得到的整体预测准确率分别是 70.57%、73.16% 和 78.05%, 各类别的预测准确率如表 5 所示. 通过对比表 4 和表 7 可以看出, 相比于传统的实验方法, 多重进化矩阵这种表示方法在 25PDB 数据集上分别高出了 2.00%、0.54% 和 1.33%. 各表中整体预测准确率提高的值用粗体显示.

表 5 RS126 数据集使用多重进化矩阵预测结果 (%)

分类模型	Q_H	Q_E	Q_C	Q_3
Logistics	79.53	49.42	65.82	62.22
RandomForest	60.13	53.85	80.89	68.08
M-SVM$_{CS}$	73.69	61.08	80.91	74.05

表 6 CB513 数据集使用多重进化矩阵预测结果 (%)

分类模型	Q_H	Q_E	Q_C	Q_3
Logistics	78.77	50.54	71.31	69.18
RandomForest	72.83	50.87	82.22	71.89
M-SVM$_{CS}$	79.26	62.38	80.39	75.92

表 7 25PDB 数据集使用多重进化矩阵预测结果 (%)

分类模型	Q_H	Q_E	Q_C	Q_3
Logistics	64.65	58.53	81.22	70.57
RandomForest	74.39	54.26	82.76	73.16
M-SVM$_{CS}$	81.14	69.36	80.25	78.05

为了更为直观的体现本文方法的有效性, 本文将在 RS126 数据集、CB513 数据集和 25PDB 数据集上使用不同算法得到的整体预测准确率表示成图 2、图 3 和图 4, 从中可以看出, 本方法相对于原 BLOSUM62 蛋白质序列表示方法, 除去对 RS126 数据集使用 Logistics 分类器得到的结果有所下降, 在其他对比实验中得到的整体预测准确率均有提高. 对于这种现象, 由于:

(1) 相对于 CB513 数据集和 25PDB 数据集, RS126 数据集数据量比较少, 包含的蛋白质种类少 .

(2) Logistics 分类器的分类精度受样本数据量的影响, 当样本数量较小时, 结果存在的风险较大.

综合这两种因素, 我们认为, 对于逻辑回归分类器, 使用多重进化矩阵反而分类精度有所下降这种现象是正常的, 不能否认多重进化矩阵是一种有效的蛋白质序列特征表示方法. 我们将对于三个数据集的使用 M-SVM$_{CS}$ 分类器得到的结果汇总, 如图 5 所示. 从表 2 至表 7 和图 5 可以看出, 在整体预测准确率上, 本文方法比 BLOSUM62 矩阵表示方法在不同数据集上分别提高了 0.15%、0.42% 和 1.33%, 对于数据集 RS126 提升较小, 对于 25PDB 数据集提升较大. 说明多分类支持向量机比较适用于大样本数据集, 而对于小样本数据集效果并不明显.

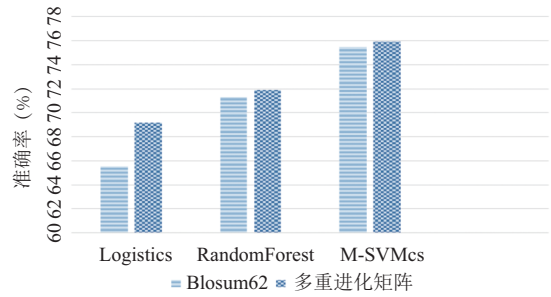

图 2 不同方法在数据集 RS126 的整体预测准确率

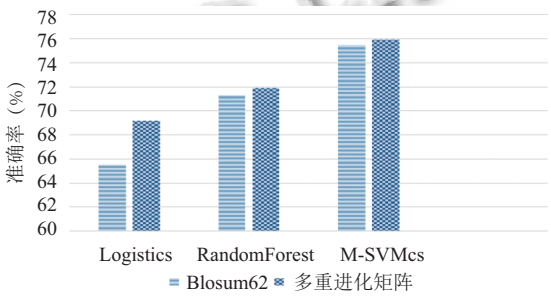

图 3 不同方法在数据集 CB513 的整体预测准确率

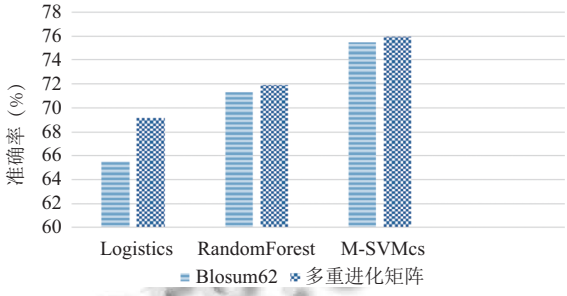

图 4 不同方法在数据集 25PDB 的整体预测准确率

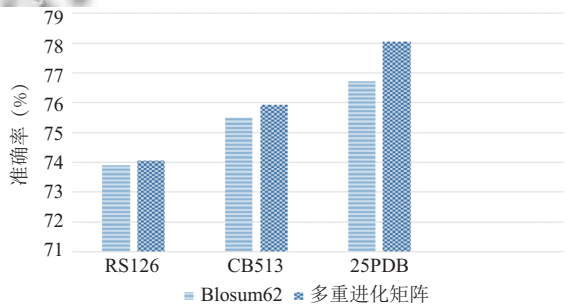

图 5 M-SVM$_{CS}$ 分类器在不同数据集的整体预测准确率

5 结语

本文根据蛋白质序列不同进化趋异度之间的关系, 组合 PAM 矩阵和 BLOSUM 矩阵, 设计了一种新的方法来构成特征向量表示蛋白质序列信息; 选用 Logistics、

RandomForest 和 M-SVM$_{CS}$ 机器学习模型作为预测工具, 采用交叉验证法和网格搜索法相结合来确定实验参数, 预测各个位点残基二级结构. 在数据集 RS126、CB513 和 25PDB 上开展的对比实验, 表明本文所提出基于多重进化矩阵的蛋白质特征向量构造方法能够有效提高蛋白质二级结构的预测精度.

在下一步的工作中, 我们可以从下面几点做出改进: (1) 深入研究蛋白质信息特征提取算法, 加入对蛋白质二级结构特征信息的描述; (2) 尝试利用特征选择算法优选特征, 降低特征向量维度, 提高分类器计算速度. 在分类算法上进行可能的改进也是下一步研究的重点.

参考文献

1 Jones DT. Protein structure prediction in the postgenomic era. Current Opinion in Structural Biology, 2000, 10(3): 371–379. [doi: 10.1016/S0959-440X(00)00099-3]

2 泽瓦勒贝, 鲍姆. 理解生物信息学. 李亦学, 郝沛, 译. 北京: 科学出版社, 2012.

3 Floudas CA. Computational methods in protein structure prediction. Biotechnology and Bioengineering, 2007, 97(2): 207–213. [doi: 10.1002/(ISSN)1097-0290]

4 Khoury GA, Smadbeck J, Kieslich CA, *et al*. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnology, 2014, 32(2): 99 –109. [doi: 10.1016/j.tibtech2013.10.008]

5 张海霞, 唐焕文, 张立震, 等. 蛋白质二级结构预测方法的评价. 计算机与应用化学, 2003, 20(6): 735–740.

6 Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology, 2007, 8(12): 995–1005. [doi: 10.1038/nrm2281]

7 Cai YD, Zhou GP. Prediction of protein structural classes by neural network. Biochimie, 2000, 82(8): 783–785. [doi: 10..1016/S0300-9084(00)01161-5]

8 Mandle AK, Jain P, Shrivastava SK. Protein structure prediction using support vector machine. International Journal on Soft Computing, 2012, 3(1): 67 –78. [doi: 10.5121/ijsc]

9 Wang S, Peng J, Ma JZ, *et al*. Protein secondary structure prediction using deep convolutional neural fields. Scientific Reports, 2016, 6: 18962. [doi: 10.1038/srep18962]

10 Lu Z, Szafron D, Greiner R, *et al*. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics, 2004, 20(4): 547 –556. [doi: 10.1093/bioinformatics/btg447]

11 Chou KC, Cai YD. Predicting protein localization in budding yeast. Bioinformatics, 2005, 21(7): 944–950. [doi: 10.1093/bioinformatics/bti104]

12 Cai YD, Chou KC. Predicting 22 protein localizations in budding yeast. Biochemical and Biophysical Research Communications, 2004, 323(2): 425 –428. [doi: 10.1016/j.bbrc.2004.08.113]

13 Wang JR, Wang C, Cao JJ, *et al*. Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features. Gene, 2015, 554(2): 241–248. [doi: 10.1016/j.gene.2014.10.037]

14 梅娟, 赵吉, 傅毅. 基于图聚类和序列信息的蛋白质远同源性探测. 计算机与应用化学, 2015, 32(8): 945–950.

15 Wang L, You ZH, Xia SX, *et al*. Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier. Journal of Theoretical Biology, 2017, 418: 105 –110. [doi: 10.1016/j.jtbi.2017.01.003]

16 Ben-Gal I, Shani A, Gohr A, *et al*. Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics, 2005, 21(11): 2657–2666. [doi: 10.1093/bioinformatics/bti410]

17 Sebastiani F. Text categorization. Rivero LC, Doorn JH, Ferraggine VE. Encyclopedia of Database Technologies and Applications. Hershey, US: Idea Group Reference, 2005. 683–687.

18 Ortuño FM, Valenzuela O, Prieto B, *et al*. Comparing different machine learning and mathematical regression models to evaluate multiple sequence alignments. Neurocomputing, 2015, 164: 123–136. [doi: 10.1016/j.neucom.2015.01.080]

19 Lal D, Verma M. Large-scale sequence comparison. Keith JM. Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution. New York: Springer, 2017. 191–224.

20 乔纳森•佩夫斯纳. 生物信息学与功能基因组学. 孙之荣, 译. 北京: 化学工业出版社, 2006.

21 Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, 1993, 232(2): 584–599. [doi: 10.1006/jmbi.1993.1413]

22 Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins Structure Function and Bioinformatics, 1999, 34(4): 508–519. [doi: 10.1002/(ISSN)1097-0134]

23 Kurgan LA, Homaeian L. Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognition, 2006, 39(12): 2323–2343. [doi: 10.1016/j.patcog.2006.02.014]

24 http://www.Loria.fr/lauer/MSVMpack.