doi: 10.3969/j.issn.0490-6756.2020.05.028

一种预测判断蛋白质 DNA 相互作用位点的新方法

王皆恒, 李 校
(四川大学生命科学学院 四川省分子生物与生物技术重点实验室, 成都 610064)

摘 要: 蛋白质-DNA 相互作用位点在各类生理生化反应中扮演重要角色. 本论文旨在构建一种可以准确预测"相互作用位点"的方法:PdDNA,其内容主要包括支持向量机和序列匹配器. 支持向量机通过提取相互作用位点中心残基的特征进行训练并分类,序列匹配器则通过蛋白质特征矩阵(PSSM)对氨基酸序列进行相关性评估,对二者结果进行归一化整合,得到最终的预测结果. 利用公开数据集 PDNA_62,我们的 PdDNA 预测准确率为 86.87%. 为进一步验证 PdDNA 可靠性,我们还自建了 PDNA_224 数据集,其预测准确率为 83.07%,处于较高水平. 因此 PdDNA 是一种有效的"蛋白质-DNA 相互作用位点"预测方法.

关键词: 蛋白质-DNA 相互作用位点预测; 支持向量机; 序列匹配算法
中图分类号: Q811.4 **文献标识码**: A **文章编号**: 0490-6756(2020)05-1009-06

A new method to predict Protein-DNA binding sites

WANG Jie-Heng, LI Xiao
(Sichuan Key Laboratory of Molecular Biology and Biotechnology,
College of Life Sciences, Sichuan University, Chengdu 610064, China)

Abstract: Protein-DNA binding sites play an important role in various physiological and biochemical reactions. In this paper, we establish a special method and algorithm based on Bioinformatics to forecast Protein-DNA binding sites, we call it PdDNA. According to our method we have 2 mainly algorithm: SVM-based predictor and sequence-based predictor. SVM-based predictor is trained and classified by extracting features of central residues at binding sites, and sequence-based predictor scores amino acid sequences for correlation by Position-Specific Scoring Matrix(PSSM). Normalization and integration of the two results to obtain the final forecast. According to our algorithm, it predicts DNA-binding sites with 86.87% accuracy when tested on PDNA_62 dataset. Otherwise, we established PDNA_224 data set, and PdDNA also has 83.07% accuracy at a high level. Therefore, PdDNA is an effective method for predicting "Protein-DNA binding sites ".

Keywords: Predict "Protein-DNA binding sites"; SVM-based predictor; Sequence-based predictor

1 引 言

对生命活动起到关键作用的两类生物大分子就是蛋白质和 DNA. 而这二者之间的相互作用则是大量生理生化反应的核心,它在基因表达调控[1]、组蛋白修饰[2],DNA 复制、修复和重组[3]等细胞过程中发挥着极其重要的作用. 蛋白质中有一类序列被称为"功能残基",它们只占庞大的功能蛋

收稿日期: 2020-02-25
基金项目: 国家自然科学基金(61001149)
作者简介: 王皆恒(1995-), 男, 山东济宁人, 硕士研究生, 研究方向为生物信息学. E-mail: wwjjhh@vip.qq.com
通讯作者: 李校. E-mail: lix@scu.edu.cn

白质序列中的一小部分,但却真正的参与了与其他生物大分子的相互作用以及各类生理生化反应. 也正因如此解析这类“功能残基”的真正功能和作用位点变得极为关键. 例如:如果想要深入地了解转录过程的机理,不可避免地需要首先了解转录过程中的 DNA 相互作用位点. 关于蛋白质功能残基位点的寻找主要有两种方法:第一种方法即为传统的通过位点突变来确定蛋白质功能残基的实验方法,实验方法的准确率较高但需要大量人力物力的投入,同时实验周期较长,完全无法满足目前生物学数据的增长速度. 由于蛋白质的序列结构和功能之间有着紧密的联系,若两种蛋白质的功能残基序列或结构相似,那么从生物意义上来讲这二者很有可能存在相似的生物学功能,因此仅仅通过单纯的数据计算也可以预测得到有意义的功能预测结果.

随着基因组/转录组测序技术的不断发展和成本不断降低,大量的生物学数据也需要这样一种高效的方法快速筛选需要的生物学数据,进而快速而精准的预测蛋白质上的功能残基位点.

在蛋白质转移至靶细胞形式功能的过程中,蛋白质的部分基团承载着前往靶细胞的定位信息,通过这些基团内的信息寻找靶细胞相关相互作用位点的过程也被称之为亚细胞定位[4]. 在亚细胞定位信息领域,通过支持向量机(SVM)进行亚细胞定位分析已经取得显著成果. 同时,支持向量机也可以对蛋白质中简单的四种超二级结构进行预测[5]. 经过改进的 Weighted SVM[6] 对于蛋白质磷酸化位点也有着不错的预测效果. 而多重支持向量机首尾相连进行大量特种归类也被称为神经网络算法[7]. 生物信息学分析着眼于序列比对、结构预测、分子进化等领域[8],随着数据量的增加神经网络算法势必能够为其他更深入的科学工作提供更可靠的实验数据分析.

目前,在 PDB 数据库[9] 中储存的蛋白质-DNA 复合体三维结构数据日益增加,这些结构数据也为功能残基位点预测提供了大量的数据来源. 在生物信息学中我们可以在这些数据库中挖掘提炼已有数据,对其进行不同特点和功能的分类和统计分析,从而得到潜在的结构或者序列共同点,并以此为基础整理成自动化算法对未知蛋白的功能残基位点进行预测. 在本论文当中我们共设计了两种预测蛋白质功能残基的计算方法:基于相似序列的一级结构预测、基于机器学习算法的支持向量机(SVM)预测. 并对其结果进行归一化整合,得到最佳的预测结果.

2 材料与方法

2.1 数据集构建

我们使用了 PDNA_62 和 PDNA_224 数据集来评估我们的预测能力,两个数据集的来源如下:

PDNA_62:PDNA_62 是一个由 Ahma and Sarai 建立的经典非冗余蛋白质-DNA 复合物数据集[10],主要集中了 Protein Data Bank (PDB)[9] 数据库中分辨率高于 3.5 Å 的蛋白质-DNA 复合物,并去除序列相似度高于 25%的冗余序列. PDNA_62 数据集中共包含 1 215 个 DNA 相互作用位点,以及 6 948 个非 DNA 相互作用位点.

PDNA_224:根据最新版本的 PDB 数据库(2019 年 7 月 4 日),检索 DNA 结合蛋白,并将 X 晶体衍射分辨率设置为 3.0 Å,这样从数据库里面得到 978 个蛋白质-DNA 复合物. 利用 PISCES software 过滤[11] 对候选复合物序列进行过滤,并剔除相似度高于 25%的冗余序列,再去除与 PDNA_62 中的同源序列,最终得到 224 条非冗余蛋白质序列. 其中共包含 3 778 个 DNA 相互作用位点和 53 570 个非 DNA 相互作用位点.

2.2 基于 SVM 的预测器

由于蛋白质中每个残基的特征可以用一个读码窗来描述,读码窗由中心残基(central residue)及其 n 个相邻残基的位置组合而成,若相邻残基不存在,则用 0 来表示. 假设读码窗的长度取 L,每个中心残基拥有 $P=24$ 个特征数,那么预测蛋白质-DNA 相互作用位点的特征总数是 $L\times P=L\times 24$,最后将这些特征量格式化为支持向量机的输入数据,得到预测结果. 我们所利用的是公用的 LibSVM 软件,该软件从 http://www.csie.nut.edu.tw/~cjlin/libsvm[12] 下载得到. LibSVM 将其输出结果转化成条件概率通过使 sigmoid 函数,sigmoid 函数的定义如下:

$$P(Y=1|x)=\frac{1}{(1+\exp(A\times sdv(x)+B)}$$

其中 x 是支持向量机的输入特征,sdv(x)代表输入特征 x 的阈值,$P(Y=1|x)$ 是条件概率结果,A 和 B 分别是 sigmoid 函数的斜率和偏移参数. 官方文档推荐 $A=-2.0$, $B=-0.5$. 此外,由于我们的数据集中正负样本差异过大,因此我们使用了随机过抽样(over-sampling)分析,并使用 LibSVM 的 RBF(radial basis function)内核进行超平面分类

器进行建模，以期尽可能消除样本数量对 SVM 带来的偏差.

为了是其预测结果能够更好地与后续序列匹配预测器结果整合，我们将上述方程进行反函数处理，使最终结果为 SVM 预测阈值.

$$sdv(x) = \frac{\ln(\frac{(1-P)}{P}) - B}{A}$$

2.3　基于序列匹配的预测器

PDNA_62 和 PDNA_224 中的序列文件保存为 FASTA 格式，但其并不能很好地反映序列之间的替换关系. 因此利用 PSI-BLAST，根据 BLOSUM62 替代矩阵对数据库进行初始化打分，生成 PSSM 蛋白质特征矩阵(图 1).

Last position-specific scoring matrix computed, weighted observed percentages rounded down, information per position, and relative weight of gapless real matches

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V		
1 M	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	3	-3	5	-1	-4	-3	-2	5	-1	3	0.61	0.30
2 E	-3	-2	-2	1	-6	1	7	-4	-2	-5	-5	-1	-4	-5	-3	-2	-3	-5	-4	-4	1.61	0.80
3 K	-3	3	-1	-3	-5	-1	-1	-4	-3	-5	-4	7	-3	-5	-3	-2	-3	-5	-4	-4	1.41	0.82
4 P	-2	-3	-4	-3	-3	-3	-3	-4	-4	-5	-4	-3	-4	-6	8	-2	-3	-6	-5	-4	2.62	0.82
5 Y	-4	-2	-4	-5	-4	-3	-4	-5	2	-3	-3	-4	-3	3	-5	-4	-4	0	9	-3	1.94	0.82
6 A	1	4	-2	-3	-4	0	-1	-2	-3	-4	-4	4	-1	-4	1	0	0	-5	-4	-3	1.94	0.82
7 C	-2	-5	-5	-6	11	-5	-6	-5	-5	-3	-3	-5	-3	-4	-5	-3	-3	-4	-1	-3	2.79	0.82
8 P	-1	-1	0	3	-1	-1	0	0	-1	-1	-1	-1	-1	-1	4	-1	-1	-1	-1	-1	2.84	0.82
9 V	0	0	0	-1	0	-1	0	-1	-1	1	0	-1	1	-1	0	-1	0	-1	-1	3	2.62	0.82
10 E	2	0	1	0	3	0	6	2	1	3	3	0	2	3	2	1	0	3	2	2	2.62	0.82
11 S	-1	-2	0	0	-4	2	6	-1	0	-4	-3	-1	-3	-4	-2	0	-2	-3	-2	-3	2.62	0.82
12 C	-2	-5	-5	-6	11	-5	-6	-5	-5	-3	-3	-5	-3	-4	-5	-2	-3	-4	-3	-3	2.79	0.82
13 D	-2	-2	2	4	-4	-2	-1	3	-3	-5	-5	-2	-4	-5	-2	4	-1	-5	-4	-4	2.62	0.82
14 R	-3	5	-2	-3	-5	-1	-1	-4	-2	-5	-4	6	0	-5	-3	-2	-2	-5	-4	-4	2.62	0.82
15 R	-1	3	-1	-2	-2	6	-1	-2	-2	-5	-4	0	-3	-5	-2	2	-1	-1	-4	-4	2.62	0.82
16 F	-4	-5	-5	-6	-4	-5	-5	-5	-3	-2	-1	-5	-2	9	-6	-4	-4	-1	1	-3	2.79	0.82
17 S	0	-2	0	-2	-2	-2	-2	-1	-3	-3	-4	-1	-3	-4	-2	6	0	-5	-4	-3	2.62	0.82
18 D	-3	4	-2	-1	-5	6	2	-3	2	-4	-3	1	-2	-4	-3	-2	-1	-4	-3	-2	2.62	0.82
19 S	-1	-2	-1	-4	-2	-1	-3	-3	-4	-1	4	-1	1	-2	-2	4	-1	-4	-3	-2	2.62	0.82
20 S	0	-3	-1	3	3	-2	-2	3	-1	-3	-5	-3	-4	-3	-3	4	-1	-5	-3	-3	2.62	0.82
21 N	-1	-2	5	2	-4	0	2	-2	5	-4	-5	-2	-3	-4	-4	1	0	-5	-2	-3	2.62	0.82
22 L	-3	-4	-6	-6	-3	-4	-5	-6	-5	0	6	-5	1	-1	-5	-4	-3	-4	-3	-1	2.79	0.82
23 T	2	0	1	2	4	1	1	4	3	1	4	6	1	5	3	1	3	5	4	2	2.62	0.82
24 R	-1	4	-1	-3	-4	0	-2	-3	-1	0	-3	3	-1	-4	-4	1	2	-5	-4	-1	2.62	0.82
25 H	-4	-2	-1	-3	-5	-1	-2	-4	10	-5	-5	-3	-4	-3	-4	-3	-4	-4	0	-5	2.79	0.82
26 I	-2	1	-3	-4	-4	1	-2	-5	-2	2	0	0	8	-2	-4	-3	-2	-4	-3	1	2.62	0.82
27 R	-3	8	-2	-4	-5	0	-2	-4	-1	-4	-4	1	-2	-5	-4	-3	-3	-5	-4	-4	2.62	0.82
28 I	-2	-3	-2	-3	-2	-3	-3	-4	-4	3	-2	-3	-2	-3	-3	-1	6	-4	-2	-1	2.62	0.82
29 H	-4	-2	-1	-3	-5	-1	-2	-4	10	-5	-5	-3	-4	-3	-4	-3	-4	-4	0	-5	2.79	0.82
30 T	-2	-3	-1	-3	-3	-3	-3	-4	-4	-4	-3	-2	-3	-4	-3	0	7	-5	-4	-2	2.62	0.82
31 G	-2	-4	-2	-3	-5	-4	-4	7	-4	-6	-6	-3	-5	-5	-4	-2	-4	-5	-5	-5	2.79	0.82
32 Q	-2	-2	-2	0	-6	2	7	-4	2	-5	-5	-1	-4	-5	-3	-2	-3	-5	-4	-4	2.62	0.82
33 K	-3	1	-1	-3	-5	0	-1	-4	-3	-5	-5	7	-3	-5	-3	-2	-2	-5	-4	-4	2.62	0.82
34 P	-3	-4	-4	-3	-5	-3	-3	-4	-3	-3	-4	-3	-4	-6	8	-2	-3	-6	-5	-4	2.79	0.82
35 F	-4	-4	-3	-5	-4	-4	-4	-5	0	-3	-3	-4	-3	5	-5	-4	-4	0	9	-3	2.62	0.82
36 Q	0	4	-1	-2	-5	2	-1	-3	-1	-3	-4	4	-2	-5	1	-1	1	-5	-4	-4	2.62	0.82
37 C	-2	-6	-5	-6	11	-5	-6	-5	-5	-3	-3	-5	-3	-4	-5	-3	-3	-4	-4	-3	2.79	0.82
38 R	-2	2	-1	4	-5	0	5	-3	-2	-5	-4	0	-3	-5	0	-2	-1	-5	-4	-4	2.62	0.82
39 I	-2	-2	-2	-1	-5	1	6	-4	0	2	-3	-1	-3	-4	-3	-2	-2	-5	-3	1	2.62	0.82
40 C	-2	-6	-5	-6	11	-5	-6	-5	-5	-3	-3	-5	-3	-4	-5	-3	-3	-4	-2	-3	2.79	0.82
41 M	-2	-2	0	2	-1	-1	-1	2	-2	-4	-3	-2	3	-4	-3	5	-1	-5	-4	-3	2.62	0.82
42 R	-3	6	-2	-3	-5	0	-2	-4	-1	-4	-4	6	0	-5	-3	-2	-3	-5	-4	-4	2.62	0.82
43 N	-1	0	1	-2	0	6	0	-2	-1	-5	-4	-1	-3	-4	-3	2	-1	-4	-2	-4	2.62	0.82
44 F	-4	-5	-5	-5	-4	-5	-5	-5	-3	-2	-1	-5	-2	9	-6	-4	-4	-1	1	-3	2.79	0.82
45 S	0	0	1	-2	-3	-2	-2	-1	-3	-3	-4	-2	-3	-4	-3	6	0	-5	-2	-3	2.62	0.82
46 R	-3	4	-2	-1	-3	5	2	-3	1	-3	-3	0	-2	-4	-4	-1	-1	-3	-3	-1	2.62	0.82
47 S	-2	-2	0	-4	-3	-1	-3	-4	-3	-1	4	-3	1	-1	-1	3	-1	-4	-2	-2	2.62	0.82
48 D	-1	-3	-1	3	2	-2	-2	4	1	-3	-5	-2	-4	-5	-3	3	-1	-5	-3	-3	2.62	0.82
49 H	-1	-2	3	0	-4	-1	-2	-3	8	-4	-4	-2	-3	-4	-3	1	0	-5	-1	-2	2.62	0.82
50 L	-3	-4	-5	-6	-3	-4	-5	-6	-5	0	6	-4	1	-1	-5	-4	-3	-4	-3	-1	2.79	0.82
51 T	-3	0	1	-3	-4	0	0	-4	-3	-3	-4	6	-2	-5	-3	-2	3	-5	-3	-2	2.62	0.82
52 T	0	2	-1	-3	-4	1	-1	-2	-1	0	-2	2	-3	-4	-3	1	4	-5	-4	-1	2.62	0.82
53 H	-4	-2	-1	-3	-5	-1	-2	-4	10	-5	-5	-3	-4	-3	-4	-3	-4	-4	0	-5	2.79	0.82
54 I	-3	2	-3	-4	-4	2	-2	-5	-3	2	0	0	8	-3	-4	-3	-2	-4	-3	0	2.62	0.82
55 R	-3	8	-2	-4	-5	0	-2	-4	-2	-4	-3	1	-2	-5	-4	-3	-3	-5	-4	-4	2.62	0.82
56 T	-2	-3	-2	-3	-3	-3	-3	-4	-4	1	-3	-3	-2	-4	-3	0	7	-5	-4	-1	2.62	0.82
57 H	-4	-2	-1	-3	-5	-1	-2	-4	10	-5	-5	-3	-4	-3	-4	-3	-4	-4	0	-5	2.79	0.82
58 T	-2	-3	-1	-3	-3	-3	-3	-4	-4	-4	-3	-3	-3	-4	-3	0	7	-5	-4	-2	2.62	0.82
59 G	-2	-4	-2	-2	-3	-4	-4	7	-4	-6	-6	-4	-5	-5	-4	-2	-4	-5	-5	-5	2.79	0.82
60 E	-3	-2	-2	0	-6	0	7	-4	-2	-5	-5	-1	-4	-5	-3	-2	-3	-5	-4	-5	2.62	0.82

	K	Lambda
Standard Ungapped	0.1366	0.3276
Standard Gapped	0.0410	0.2670
PSI Ungapped	0.1964	0.3179
PSI Gapped	0.0601	0.2670

图 1　蛋白质 1TC3 中 C 段肽链的 PSSM 特征矩阵

Fig. 1　PSSM matrix of C-chain in protein 1TC3

使用初始化后的 PSSM 数据文件进行计算. 序列的提取与 SVM 中的信息窗口类似，使用一个长度为 n 的读码框(n=3、5、7、9、11……)，若两边数据不足则用 0 补齐. 读码框在已知(B)和未知(A)序列中同时滑动，并在相应位点进行计算. 相应的计算公式我们使用了锌指结构预测中 CHDEs 氨基酸得分的计算公式[13].

$$Score(\alpha,\beta) = \sum_{j=1}^{9}\left(\sum_{i=1}^{20}(\alpha_{ni}\log\left(\frac{\beta_{ni}}{P_i}\right)+\beta_{ni}\log\ (\frac{\alpha_{ni}}{P_i}))\right)$$

此公示中 α、β 表示两条序列，j 表示长度为 n 的读码框中第 j 位点的匹配的分，i 则表示每个位点的氨基酸与 20 个标准氨基酸进行匹配，α_{ni}、β_{ni} 则表示这些氨基酸的匹配的分，关于单个位点与 20 个氨基酸得分对应关系，依旧沿用 BLOSUM62 氨基酸得分矩阵，具体的计算方法如下. P_i 则代表氨基酸

i 出现的背景频率. 最终将所有读码框的得分相加,总得分即能够代表功能未知序列(A)和功能已知序列(B)之间的相似关系.

$$\alpha_{i,j}=e^{M_{i,j}\lambda_u}p_i$$

其中 $M_{i,j}$ 即为 BLOSUM62 矩阵中两氨基酸对应得分,$\lambda\mu$ 即为 PSSM 格式化后的 α 文件中 standard ungapped Lambda value.

由于此方法得到大量得分结果,最理想的情况中,我们希望这些整理后的坐标呈强烈的线性关系,这样才能够进一步说明两个片段之间拥有强烈的相关性. 因此我们以单个蛋白质肽链前 $x\%$、整个训练集得分的第 $y\%$作为阈值,由于皮尔逊系数可以用来表示坐标点的相关性,因此我们使用皮尔逊相关性约束作为参数,同时调整窗口长度(n=7,9,11,……),建立模型进行训练,并使用训练结果最优的模型进行得分筛选,得到初步的待处理匹配位置,并将匹配位置整理成坐标形式. 筛选模型可分为四类:

type1:用单个蛋白质肽链前 $x\%$最为阈值,并进行皮尔逊相关系数的约束;

type2:用整体蛋白质肽链前 $y\%$作为阈值,并进行皮尔逊相关系数的约束;

type3:用单个蛋白质肽链前 $x\%$最为阈值,不进行皮尔逊相关系数的约束;

type4:用整体蛋白质肽链前 $y\%$作为阈值,不进行皮尔逊相关系数的约束.

最终我们将确定模型筛选出来的位置坐标进行线性匹配,以求找到更多的坐标点位于同一条直线. 那么这些位于同一直线的坐标点所代表的序列位点,即可视为拥有强烈的序列相似性和功能相关性.

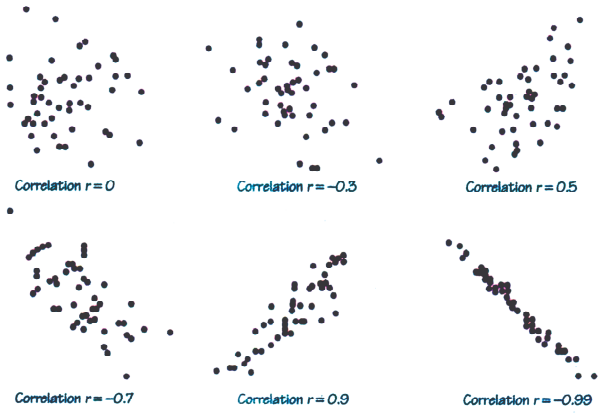

图 2 不同坐标点的皮尔逊系数[14]

Fig. 2 Pearson coefficient of different dataset[14]

2.4 对预测结果进行归一化整合

由于预测结果各异,且有着各自的优缺点,因此我们使用数学期望对两种预测器进行数据整合,根据数学期望公式,我们可整合支持向量机打分 $sdv(x)$ 与序列匹配打分 Score(α,β),具体函数如下:

$$E(x)=(1.0-p)\times\sum_{x=1}^{L}sdv(x)+Score(\alpha,\beta)\times p$$

最终,对长度为 L 的 α 序列,其于 β 序列的结合得分为 $E(x)$.

3 结果分析

PdDNA,即本论文所研究得出的蛋白质-DNA 相互作用位点位点预测方法,主要整合了 SVM 支持向量机打分和序列匹配度打分. 在 SVM 模型训练过程中,我们使用 PDNA_62 标准实验数据集进行测试,下表为五交叉检验结果随读码窗口长度变化所得到的预测信息. 我们的评估指标包括敏感度(Sensitivity, SN)、特异性(Specificity, SP)、强度(Strength)、准确度(ACC)和马太相关系数(MCC)和,其具体计算方法如下. 最终确定长度 11 为最佳窗口长度(表 1).

$$SN=TP/(TP+FN)$$

$$SP=TN/(TN+FP)$$

$$Strength=(SN+SP)/2$$

$$Acc=(TP+TN)/(TP+FP+TN+FN)$$

$$MCC=(TP\times TN-FP\times FN)/\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}$$

表 1 PDNA_62 数据集预测精度随 SVM 窗口长度变化情况

Tab. 1 The prediction result of the PDNA_62 dataset with different window sizes of the SVM

k	Sn/%	Sp/%	Acc/%	MCC	Str/%
3	76.38	76.64	76.61	0.38	76.51
5	77.21	77.85	77.77	0.40	77.53
7	76.75	78.59	78.36	0.40	77.67
9	76.81	78.34	78.16	0.40	77.59
11	76.84	79.71	79.36	0.42	78.28
13	76.28	78.33	78.07	0.40	77.31
15	76.94	78.21	78.05	0.40	77.57
17	76.28	79.44	79.05	0.41	77.86
19	75.44	80.27	79.67	0.41	77.86
21	74.22	80.82	80.00	0.41	77.53

对于数据集 PDNA_62，在基于 PdDNA 运算模型的测试中，其准确度（ACC）为 85.15%，马太相关系数（MCC）0.55 为，强度（Strength）为 85.89%，正集预测成功率为 84.91%，负集预测成功率为 86.87%. 与其他主流预测方法对比结果如下（表 2）. 同时，最终预测结果的 ROC 曲线中（图 3）也可明显看出，结合 SVM 与序列匹配的预测准确率比单独的 SVM 预测高出 5%～10%不等.

表 2　对于 PDNA_62 数据集，PdDNA 预测结果

Tab. 2　The prediction result of the PDNA_62 dataset of PdDNA algorithm

Methods	ACC/%	MCC	SN/%	SP/%	Str/%
BindN−RF[14]	78.2	-	78.1	78.2	78.15
BindN+[15]	79	0.44	77.3	79.3	78.3
PDNAsite[16]	85.11	0.582	86.27	84.91	85.59
PdDNA	85.15	0.55	86.87	84.91	85.89

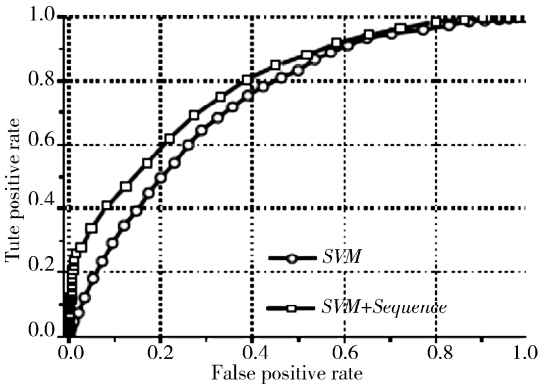

图 3　结合 SVM 预测器和基于序列匹配预测器，通过 ROC 曲线展示数据集 PDNA_62 的最终的预测结果

Fig. 3　ROC curves for the DNA-binding sites prediction in PDNA_62 dataset by combining SVM predictor with sequence-based predictor

同时，对于自建库 PDNA_224 也拥有更好的预测情况（表 3）. 从 ROC 曲线（图 4）中可以看出，其 SVM 与序列匹配整合结果也比单独的 SVM 预测略有提高，但相比 PDNA_62 数据集提升幅度较小，这可能是由于我们的正负集样本数量差距进一步扩大，进而造成了 SVM 分类器出现偏移造成的.

表 3　对于 PDNA_224 数据集，PdDNA 预测结果

Tab. 3　The prediction result of the PDNA_224 dataset of PdDNA algorithm

Methods	ACC/%	MCC	SN/%	SP/%	Str/%
BindN-RF[15]	76.5	-	75.1	76.7	75.9
BindN+[16]	74	0.37	74.1	76.1	75.1
PDNAsite[17]	82.25	0.405	83.17	82.34	82.67
PdDNA	83.07	0.42	83.08	83.03	83.05

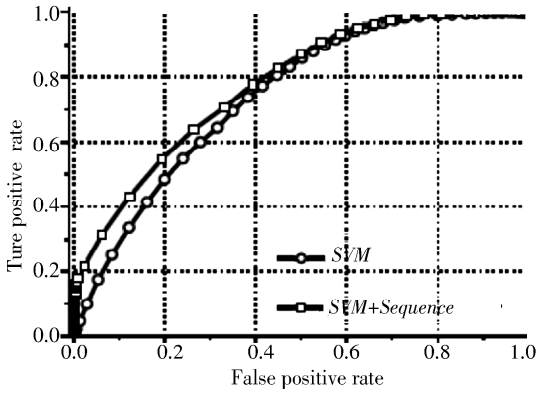

图 4　结合 SVM 预测器和基于序列匹配预测器，通过 ROC 曲线展示数据集 PDNA_224 的最终的预测结果

Fig. 4　ROC curves for the DNA-binding sites prediction in PDNA_224 dataset by combining SVM predictor with sequence-based predictor

4　讨　论

在本论文中，我们通过 SVM 支持向量机方法、序列匹配算法的融合，得到了一种全新的用来预测蛋白质－DNA 相互作用位点的实验方法—PdDNA. 并根据这个方法对标准实验数据库 PDNA_62 进行了数据分析，结果表明我们的 PdDNA 程序所得到的预测结果为：ACC，85.15%；MCC，0.55；Strength，85.89%；SP，84.91%；SN：86.87%. 其预测准确率高于目前蛋白质预测领域任何其他算法. 其高达 86%以上的预测准确度也证明了我们的预测方法具有高度的可用性和实用性，能够给蛋白质功能组学、分子生物学等其他领域的科学研究带来重要的预测参考，进一步提高科研工作的效率

同时，我们还建立了具有参考意义的 PDNA_224 蛋白质-DNA 相互作用位点数据库. 对于自建数据库的预测效果也令人满意，其最好的 ACC 为 83.07，MCC 为 0.42，Str 为 83.05. 以期对他人研究和算法设计提供有价值的数据来源.

但我们的数据集中仍存在正负样本不平衡的问题，由于客观实验限制我们很难在数据库中找到足量的正集数据. 正集样本过少的情况下，其样本无法广泛分布，因此 SVM 预测的分类器确会朝向正集偏移，导致其敏感度下降. 根据 LibSVM 的官方文档[18]中的说明，在特征数远小于样本数的情况下建议使用 RBF 内核. 在 RBF 内核中 LibSVM 会启用超平面分类器，并启用了随机过抽样方法，以此对抗样本数量不平衡的问题，这也是本文中采

用的解决方法.但这种方法对计算量需求极大.例如在 PDNA_224 数据集中,负集样本为正集样本的 10 倍,此时负集样本被随机抽样为 10 份并分别与正集样本进行建模,最终对模型进行拟合.虽精度较高但计算时间会陡增十倍.若在计算资源不足的情况下,也可以考虑手动设置惩罚因子 C,对分类结果进行加权.仍以 PDNA_224 数据集举例,正负数据集样本比例为 1∶10,则惩罚因子 C 比例可定位 10∶1,在 LibSVM 中的参数则应设置为-w1 10;-w-1 1,则可在节省计算资源的情况下得到大致相同的分类结果.缺点则是会偏离原始数据的概率分布,因而本文中没有采用.

参考文献:

[1] Ptashne M. Regulation of transcription: from lambda to eukaryotes[J]. Trends Biochem Sci, 2005, 30: 275.

[2] Kornberg R D. Chromatin structure: a repeating unit of histones and DNA [J]. Science, 1974, 184: 868.

[3] Luscombe N M, Austin S E, Berman H M, *et al* An overview of the structures of protein-DNA complexes [J]. Genome Biol, 2000, 1: 1.

[4] 张松,夏学峰,沈金城, 基于序列保守型和蛋白质相互作用的真核细胞亚细胞定位预测[J].生物化学与生物物理进展, 2008, 35: 531.

[5] 高苏娟, 胡秀珍. 基于支持向量机的整体分类器算法 预测酶蛋白质中四类简单超二级结构[J]. 计算生物学, 2014, 4: 1.

[6] 赵凌志, 刘颖, 覃征. Weighted SVM 在蛋白质磷酸化位点预测中的应用[J]. 计算机工程与应用, 2006 (3): 155.

[7] 须文波, 陆克中. 神经网络在蛋白质二级结构预测中的应用[J]. 生物信息学, 2006, 4: 26.

[8] 王玲. 基于知识发现的生物信息学[J]. 生物工程进展, 2000, 20: 27.

[9] Berman H, Henrick K, Nakamura H, *et al*. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data[J]. Nucleic Acids Res, 2007, 35, S1: D301.

[10] Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins[J]. BMC bioinformatics, 2005, 6: 33.

[11] Wang G L, Dunbrack Jr R L. PISCES: a protein sequence culling server[J]. Bioinformatics, 2003, 19: 1589.

[12] Chang C C, Lin C J. LIBSVM : a library for support vector machines[J]. ACM Trans Intell Syst Technol, 2011, 2: 27.

[13] Shu N J, Zhou T P, Hovmöller S. Prediction of zinc-binding sites in proteins from sequence[J]. Bioinformatics, 2008, 24: 775.

[14] Mindrila D, Balentyne P. Scatterplots and correlation [EB/OL]. [2020-02-25]. https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf, 2017.

[15] Wang L J, Yang M Q, Yang J Y. Prediction of DNA-binding residues from protein sequence information using random forests[J]. BMC Genomics, 2009, 10: S1.

[16] Wang L J, Huang C Y, Yang M Q, *et al*. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features[J]. BMC Syst Biol, 2010, 4: S3.

[17] Zhou J Y, Xu R F, He Y L, *et al*. PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context [J]. Sci Rep, 2016, 6: 27653.

[18] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification [EB/OL]. [2020-02-25]. https://www.researchgate.net/profile/Chenghai_Yang/publication/272039161_Evaluating_unsupervised_and_supervised_image_classification_methods_for_mapping_cotton_root_rot/links/55f2c57408ae0960a3897985/Evaluating-unsupervised-and-supervised-image-classification-methods-for-mapping-cotton-root-rot.pdf, 2003.

引用本文格式:

中 文: 王皆恒, 李校. 一种预测判断蛋白质 DNA 相互作用位点的新方法[J]. 四川大学学报: 自然科学版, 2020, 57: 1009.

英 文: Wang J H, Li X. A new method to predict Protein-DNA binding sites [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 1009.