文章编号: 1006-2475(2016) 01-0020-06

基于序列与结构特征结合的蛋白质与 DNA 绑定位点预测

杨 骥

(南京理工大学计算机科学与工程学院,江苏 南京 210094)

摘要: 目前国内外对于 DNA-蛋白质绑定位点预测的研究大多集中在仅以蛋白质序列信息或仅以蛋白质结构信息为基础进行计算,而二者结合所实现的预测效果较差。本文提出一种在蛋白质位置特异性得分矩阵序列特征的基础上,结合蛋白质残基的溶剂可及表面积、相对表面积、深度和突出指数这几个结合效果良好的结构特征的 DNA 与蛋白质绑定位点预测方法,并使用随机下采样方法解决训练集样本不平衡问题,最后使用支持向量机算法进行预测。实验结果表明,本文方法具有较好的预测能力。

关键词: 位置特异性得分矩阵; 可及表面积; 相对表面积; 深度与突出指数; 随机下采样; 支持向量机

中图分类号: TP181 文献标识码: A doi: 10.3969/j.issn.1006-2475.2016.01.005

Prediction of DNA-protein Binding Sites Based on Combining Sequence with Structure Information

YANG Ji

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Most of the research of DNA-protein binding sites are focusing on just computing protein sequence information or structure information, while the results are terrible if combing this two information, no matter what at home or abroad. To solve this problem, we combine protein structure information of accessible surface area, relative solvent accessibility, depth index and protrusion index with protein sequence information of position specific scoring matrix to predict DNA-Protein binding sites. Then we use under sampling to solve the unbalance problem of training dataset. Finally, we use support vector machine to make prediction. The result of experiment shows the method that we proposed can achieve better performance in prediction.

Key words: position specific scoring matrix; accessible surface area; relative solvent accessibility; depth index and protrusion index; under sampling; support vector machine

0 引 言

在诸多重要的生命活动当中,蛋白质与 DNA 之间的相互作用扮演着重要角色。比如 DNA 和转录因子之间的相互作用,在调控基因的复制与转录过程中起着重要的作用[1]。因此,能否正确识别蛋白质上的 DNA 结合位点,不仅仅关乎生命活动机制的理解,同时对蛋白质功能的注释也有帮助。

蛋白质和 DNA 之间的作用位点可以通过生物技术手段探测出来。然而,传统的生物学方法不仅费时费力,同时也难以应对日益庞大的蛋白质-DNA 复合物的数量。因此,寻找一种客观而有效的计算方法,来对 DNA 结合位点进行精确预测,已经成为全世界学者们研究的课题。

目前主流的 DNA 与蛋白质结合位点的预测方法,大部分为仅使用序列信息与仅使用结构信息[2-3],部分同时使用序列和结构信息的预测方法效果较为一般[4-5]。以目前日益增长的蛋白质与 DNA 复合物的数量来说,使用不同类型特征结合进行预测必将成为未来发展的重点,其比单一使用某一类型特征具有更高的效率与预测准确性。因此,本文在前人的基础上,提出一种以蛋白质位置特异性得分矩阵序列特征

收稿日期: 2015-10-19

作者简介: 杨骥(1990-) ,男,安徽合肥人,南京理工大学计算机科学与工程学院硕士研究生,研究方向: 模式识别与生物信息学。

为基础,融合蛋白质残基的溶剂可及与相对表面积,深度和突出指数这几个结构特征来对 DNA-蛋白质绑定位点进行预测的方法,旨在寻找一种比现有预测方法更好的预测方法; 并对训练集使用随机下采样方法,以及使用支持向量机算法来进行分类预测。

1　特征提取

1.1　序列特征提取

位置特异性得分矩阵(Position Specific Scoring Matrix,PSSM) 通过多序列对比的方法,能够反映出蛋白质序列的进化信息。

以包含 L 个氨基酸的蛋白质序列 P 为例。本文通过使用 PSI-BLAST[6] ,以 0.001 为其中参数 E 的值,通过搜索 Swiss-Prot 数据库执行蛋白质多序列对比,并进行 3 次迭代操作后,获得其位置特异性得分矩阵(共有 L 行,20 列) 。该矩阵如下:

$$p_{pssm}^{original}=\begin{pmatrix} o_{1,1} & o_{1,2} & \cdots & o_{1,20} \\ o_{2,1} & o_{2,2} & \cdots & o_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ o_{k,1} & o_{k,2} & \cdots & o_{k,20} \\ o_{L,1} & o_{L,2} & \cdots & o_{L,20} \end{pmatrix}_{L\times 20} \tag{1}$$

其中 $O_{k,j}$ 表示蛋白质序列 P 中的氨基酸 k 在进化过程中,突变为氨基酸 j 的得分。该得分为正,说明该突变发生的几率比预期的更高,若得分为负,则说明发生的几率比预期的要低。

在获得原始蛋白质的 PSSM 之后,需要对其按行进行归一化操作。这里用 1 ~ 20 表示 20 种天然氨基酸,并根据其首字母在字母表上的顺序进行排序。首先计算出原始 PSSM 矩阵中,每一行的均值和标准差,以第 k 行为例,μ_k 和 σ_k 分别表示第 k 行的均值和标准差,计算公式如下:

$$\mu_k=\frac{1}{20}\sum_{t=1}^{20}o_{k,t} \tag{2}$$

$$\sigma_k=\sqrt{\frac{1}{20}\sum_{t=1}^{20}(o_{k,t}-\mu_k)^2} \tag{3}$$

将执行归一化操作之后获得的 PSSM 矩阵表示为 P_{pssm},其中的第 k 行第 j 列元素值的计算方法如下:

$$p_{k,j}=\frac{o_{k,j}-\mu_k}{\sigma_k} \tag{4}$$

则最终归一化之后,包含 L 个氨基酸的蛋白质序列的 PSSM 矩阵表示如下:

$$p_{pssm}=\begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k,1} & p_{k,2} & \cdots & p_{k,20} \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{pmatrix}_{L\times 20} \tag{5}$$

在获得归一化的 PSSM 矩阵后,使用窗口大小为 W 的滑动窗口,来提取每个氨基酸所对应的 PSSM 特征向量。由于滑动窗口大小不同,对于预测器的性能有一定的影响,所以其大小的选择较为重要,具体选择过程将在后文中具体描述。本文经过试验后发现,滑动窗口 $W=11$ 为最佳选择,因此 PSSM 特征的维数为 $11\times 20=220$ 维。

1.2　结构特征提取

1.2.1　溶剂可及表面积

蛋白质残基的溶剂可及表面积(Accessible Surface Area,ASA) 的概念,最早是由 Lee 和 Richards[7] 所提出,并用于研究蛋白质分子的疏水性[8] 。随着研究的深入,已有研究表明其在蛋白质-DNA 相互作用的预测中起着重要的作用。

溶剂可及表面积是指当一个蛋白质或 DNA 分子的原子处在溶液当中时,溶剂分子可接触到的残基的表面积。其可操作定义为利用一个溶剂分子探球,沿蛋白质表面进行滚动,探球中心的所有可能轨迹点可勾勒出一个表面,即为溶剂可及表面,该表面的面积即为溶剂可及表面积。

溶剂可及表面积 A 计算公式如下:

$$D=\Delta Z/2+\Delta' Z \tag{6}$$

$$A=\sum(R/\sqrt{R^2-Z_i^2})\times D\times L_i \tag{7}$$

这里参数 R 是指被测分子中原子与溶剂分子的范德华半径之和。参数 L_i 指溶剂分子在指定区域 i 当中所滚过的弧长。参数 Z_i 是指从球的中心到第 i 个区域的垂直距离,参数 ΔZ 表示区域之间的间距,参数 $\Delta' Z$ 为 $R-Z_i$ 和 $\Delta Z/2$ 中二者中较小的那个值。并且对于给定的原子,需要计算其所有可滚动的弧之和。

1.2.2　相对溶剂可及性

蛋白质残基的相对溶剂可及性(Relative Solvent Accessibility, RSA) [9] 为残基的溶剂可及表面积与其在三肽(ALA-X-ALA) 下的最大溶剂可及表面积之比。计算公式如下:

$$RSA_{relative}=\frac{ASA_i}{ASA_L} \tag{8}$$

其中 $RSA_{relative}$ 为相对溶剂可及性,ASA_i 为残基的溶剂可及表面积,ASA_L 为该残基在三肽下的最大溶剂可

及表面积。20 种氨基酸类型在其三肽下的最大溶剂可及表面积(MAX_{ASA}) 如表 1 所示。

表 1 20 种氨基酸的最大溶剂可及面积[10]

Amino Acid	Asp	Asn	Arg	Ala
MAX_{ASA}	163	157	248	106
Amino Acid	Cys	Glu	Gln	Gly
MAX_{ASA}	135	194	198	84
Amino Acid	His	Ile	Lys	Leu
MAX_{ASA}	184	169	205	164
Amino Acid	Met	Phe	Pro	Ser
MAX_{ASA}	188	197	136	130
Amino Acid	Thr	Trp	Tyr	Val
MAX_{ASA}	142	227	222	142

本文共构建了 5 对基于溶剂可及性表面积和相对溶剂可及性的特征: 分别是 1) AllASA 和 AllRSA (氨基酸上所有原子的 ASA 和 RSA 值) ; 2) MainASA 和 MainRSA (所有主链或骨架原子的 ASA 和 RSA 值) ; 3) SideASA 和 SideRSA(所有侧链原子的 ASA 和 RSA 值) ; 4) ApASA 和 ApRSA(所有极性侧链原子的 ASA 和 RSA 值,比如氧原子和氮原子) ; 5) NpASA 和 NpRSA(所有非极性侧链原子的 ASA 和 RSA 值) 。

1.2.3 突出指数

溶剂可及表面积可以很好地表征不同的表面残基,但对于某些内部残基,由于其溶剂可及表面积基本为零或接近零,该特征的区分度较弱。因此有研究者提出了另外 2 种衡量指标,分别是残基突出指数(Protrusion Index,PI) [11] 和残基深度指数(Depth Index,DI) [12] ,由此来更好地刻画与区分蛋白质内部空间结构的排列。其中深度指数将在后面详细介绍。

残基的突出指数最早用于蛋白质间相互作用,抗原决定簇以及蛋白水解裂解等方面的研究。随着近些年研究的深入,其在 DNA 和蛋白质之间相互作用上所起的重要作用开始显现。其定义与计算公式如图 1 与公式(9 ~ 11) 所示:

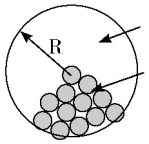

图 1 残基突出指数计算示意图[11]

$$V_{ext} = V_{sphere} - V_{int} \quad (9)$$

$$V_{int} = N_{atom} \times V_{atom} \quad (10)$$

$$PI = V_{ext} / V_{int} \quad (11)$$

这里 N_{atom}是指以一个蛋白质非氢原子为圆心,固定距离 R 为半径所形成的一个探测球,球内的所有非氢原子的数量,该探测球默认半径大小为 10 Å。V_{atom}是蛋白质中一个重原子的平均体积,这里的值为 20.1 $Å^3$。通过这些值,可以算出探测球中蛋白质所占的体积 V_{int}。同时探测球中所剩下部分的体积 V_{ext} 也可求出,则该蛋白质原子的突出指数 PI 即可求出。本文以构成各残基的所有原子的突出指数的平均值、最大值、最小值,以及所有侧链原子的均值为各残基的突出指数。

1.2.4 深度指数

残基深度指数的应用面很广,例如核磁共振(NMR) 的氨基氢/重氢交换中的速率分析,蛋白质核组装和排列分析等。此外,该特征也有助于预测蛋白质与 DNA 之间相互作用位点。

深度指数是指原子 i 与相近的溶剂可及原子 j (即 ASA 值 > 0 的原子) 的距离。计算公式如下:

$$DI = \min(d_1, d_2, d_3, \cdots, d_n) \quad (12)$$

其中 $d_1, d_2, d_3, \cdots, d_n$指原子 i 和所有溶剂可及原子的距离。因此,对于溶剂可及原子来说,其深度指数的值为 0。而对于内部原子来说,其深度指数的值与距离成正比。对残基来说,以所有构成该残基的原子的深度指数的平均值、最大值、最小值,以及所有侧链原子的均值为其残基深度指数。

2 分类器选择

2.1 支持向量机分类器

支持向量机(Support Vector Machine) [13] 是数据挖掘中一种基于统计学理论的分类方法,可以在有限样本条件下,获得较好的泛化能力,适用于模式识别、回归等诸多线性与非线性问题。本文选择使用 SVM 作为分类器。

对于一个线性可分的样本集来说,需要找到一个最优超平面,使其可以完全隔离开 2 类样本,并使得每类数据中离分类超平面最近的点与超平面的距离最大。

假设训练样本集为$\{(x_1, y_1) , \ldots , (x_i, y_j) , \ldots , (x_m, y_m) \}$,$x_i \in R^n$,其中 m 为样本数,n 为输入维数,分类标号 $y_i \in \{ +1, -1\}$,则 n 维空间中的线性判别函数为:

$$g(x) = w^T x + b \quad (13)$$

分类面方程为:

$$w \cdot x + b = 0 \quad (14)$$

其中 w 是超平面的法线方向,$w / \| w \|$ 是单位法向量,$\| w \|$ 是欧氏模函数。对其进行归一化,让所有样本满足$| g(x) | \geq 1$,即让距离分类面最近的样本的$| g(x) | = 1$,此时分类间隔为 $2 / \| w \|$,那么分类间隔

最大为 $\| w \|$，即:

$$\min\Phi(w) = \frac{1}{2} \| w \|^2 \tag{15}$$

对所有样本正确分类需要满足约束条件:

$$y_i [w^T x_i + b)] - 1 \geq 0, \quad i = 1,2,\cdots,n \tag{16}$$

满足式(16)中等号条件的少数样本称之为支持向量。则求解最优分类超平面的问题，转化为同时满足式(15)、式(16)的问题，使式(15)、式(16)同时成立的分类面即为最优分类面。定义如下拉格朗日函数:

$$L(w,b,\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i [y_i (w^T x_i + b) - 1] \tag{17}$$

其中 $\alpha_i \geq 0$ 为拉格朗日系数。联立式(15)～(17)，则问题变成对 w 和 b 求拉格朗日函数式(17)的最小值问题，并可求得最优解 α_i^*、w^* 和 b^*，则所求得的最优分类函数为:

$$f(x) = \mathrm{sgn}((w^*)^T x + b^*) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i x_i^* x + b\right) \tag{18}$$

其中 sgn(·)表示符号函数，其所确定的超平面即为最优分类超平面。根据此分类超平面对所有样本进行分类。

实际处理问题时，更多的是线性不可分数据集，这个时候就需要使用核函数来解决线性问题到非线性问题的推广。常见的核函数类型有线性核函数、多项式核函数、高斯径向基核函数、Sigmoid 核函数等。核函数的选择对于数据的正确预测相当重要。在这几种常见核函数中，高斯径向基核函数由于能准确地描绘出数据的分布情况，因此适用范围较广。本文选择此核函数作为聚类和分类工作所使用的核函数。其表达式为:

$$K(x_1, x_2) = \exp(-q \| x_1 - x_2 \|^2) \tag{19}$$

此时的最优分类函数即为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i K(x_1, x_2) + b\right) \tag{20}$$

再根据其所确定的超平面对所有非线性样本进行分类。

2.2　交叉验证

本文使用五重交叉验证来验证所提方法的有效性: 将整个样本数据集平均分成 5 个子集，选取其中 1 个子集数据作为测试集，剩下的 4 个子集作为训练集。随后，将这 5 个子集轮流作为测试集，重复上述步骤。这样既可以避免过度学习，也可以避免欠学习状态的发生，从而使得最终的结果具有较强的说服力。

3　采样方法和实验数据集

3.1　实验数据集

本文使用的数据集是由 Li[14] 等人收集的，记为 PDNA-224。该数据集由 224 个非冗余蛋白质链，共 57348 个氨基酸所组成。根据 Sarai[2] 等人的研究所得，以 3.5 Å 为界，若数据集中某一残基其原子与在 DNA 中相匹配的原子之间的距离小于 3.5 Å，则该残基即认为是绑定残基，否则为非绑定残基。最终该数据集中包含 3778 个绑定残基和 53570 个非绑定残基。

3.2　采样方法

使用 SVM 分类器针对不平衡数据进行分类时，若训练集中 2 类样本数目相差较大的话，其训练后的分类面会向少数类样本偏移，从而使得 SVM 过度拟合多数类样本，而低估了少数类样本点，导致在分类过程中的少数类样本错分，并使得算法漏检率大大增强。

由于本文是个典型的不平衡问题，非绑定残基数远远多于绑定残基数，因此采用随机下采样[15-17]方法，来处理不平衡问题: 首先选取全部绑定残基，随后从所有非绑定残基中，无重复地选择与绑定残基数相同的非绑定残基，由这二者组成新的训练集。如此重复 N 次，从而最终得到平衡样本集 S。

4　实验结果与讨论

4.1　结果评估指标

对于二分类问题来说，所有样本可分为正类和负类。本文通过阈值依赖(Threshold-dependent)方法来对 SVM 分类结果进行评估。这里主要用到 4 个标量，分别是 TP(True Positive)、TN(True Negative)、FP(False Positive)和 FN(False Negative)。TP 指一个正确预测的正类样本，TN 指一个正确预测的负类样本，FP 指一个错误预测的正类样本，FN 指一个错误预测的负类样本。通过这 4 个标量，使用特异性(Specificity，Spe)、灵敏度(Sensitivity，Sen)、准确度(Accuracy，Acc)和马修斯相关系数(Matthews correlation coefficient，MCC)来评估本文方法的性能，其计算公式如下:

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{21}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{22}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{23}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} + \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}} \tag{24}$$

同时，由于 ROC(Receiver Operating Characteris-

tic) 曲线可以表现出特异性和灵敏度之间连续变化的关系,所以将其也作为衡量本文所提出方法的标准之一。

4.2 滑动窗口选择

滑动窗口的大小对于预测器的表现有较大影响。为使本文所提方法获得最佳的预测结果,分别测试了滑动窗口大小为 3,...,21 时,预测器的表现。表 2 为不同大小的滑动窗口在 PDNA-226 数据集上的表现。从表中可以看出,滑动窗口大小从 3 开始,当滑动窗口大小不断增大时,预测器性能处于不断上升中,当滑动窗口大小为 11 时,预测器的表现最好,而当滑动窗口大小大于 11 时,随着滑动窗口大小的增大,预测器性能处于不断下降状态。因此本文所选择的滑动窗口大小为 11。

表 2 PDNA-224 数据集上基于 PSSM 特征的滑动窗口结果对比

Size	Sen/%	Spe/%	Acc/%	MCC
3	65.7348	74.8108	74.1727	0.23076
5	66.6401	76.5039	75.8104	0.24983
7	65.8413	76.7957	76.0257	0.24788
9	67.6518	76.4133	75.7973	0.25469
11	**67.8115**	**76.9729**	**76.3289**	**0.26045**
13	67.8913	76.9205	76.2858	0.26043
15	66.7199	76.5562	75.8647	0.25073
17	66.2141	77.0152	76.2558	0.25188
19	65.4952	76.9145	76.1117	0.24697
21	67.27902	76.3669	75.7281	0.25222

4.3 特征与特征组合的预测结果

表 3 表示当使用不同特征与特征组合时,其对预测器性能的影响。从表中可以看出,当单独使用 PI 和 CI 特征时,预测器的准确度最低,为 60.5%,相关系数为 0.134。单独使用 ASA 特征时,准确度为 63.6%,相关系数为 0.167,当单独使用 RSA 特征时,准确度为 64.6%,相关系数为 0.16。单独使用 PSSM 特征时,准确度为 76.3%,相关系数为 0.26。当将不同特征组合使用时,预测器的性能在不断提高,并在最终将 PSSM 特征、PI 特征、CI 特征、ASA 特征和 RSA 特征同时使用时,使得预测器的性能得到最大幅度提升,其准确度为 83.8%,相关系数为 0.40。其对应 ROC 曲线图,如图 2 所示。

表 3 PDNA-226 数据集上不同特征与组合结果对比

Feature	Sen/%	Spe/%	Acc/%	MCC
P	67.8115	76.9729	76.3289	0.26045
P – C	65.6549	60.1687	60.5543	0.13391
ASA	68.8498	63.1502	63.5509	0.16765
RSA	65.9212	64.4749	64.5766	0.16038
P + P – C + ASA	76.9169	81.3275	81.0174	0.35509
P + P – C + ASA + RSA	**79.1533**	**84.1234**	**83.7744**	**0.40199**

注: P 表示 PSSM 特征,P – C 表示 PI 和 CI 特征

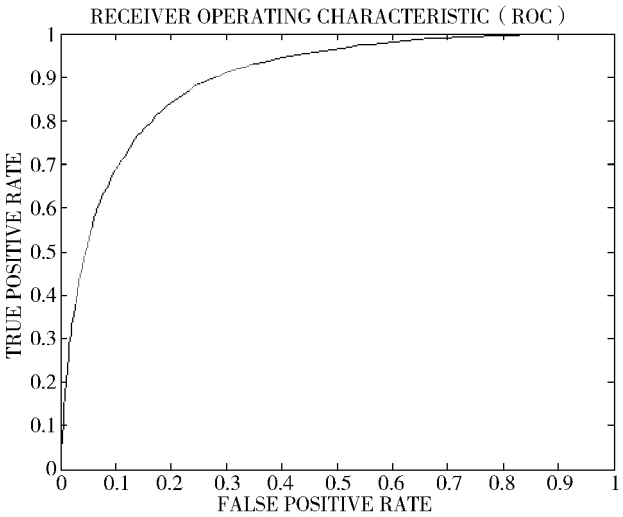

图 2 PDNA-226 数据集上最佳特征组合对应 ROC 曲线图

4.4 不同分类方法的结果对比

由上文可知,本文最终选择的特征组合为 PSSM 特征、PI 特征、CI 特征、ASA 特征和 RSA 特征。本文在 PDNA-224 数据集上,使用该特征组合,分别使用 SVM 分类方法和 RBF(Radial Basis Function) 分类方法这 2 种常用分类方法进行预测。通过比较这 2 种常见分类方法的预测结果,选择本文最终所使用的分类方法,从而使本文所建立的预测模型更具有说服力。其预测结果如表 4 所示。

表 4 在 PDNA-226 数据集上与其他预测方法的结果对比

Classifier	Sen/%	Spe/%	Acc/%	MCC
RBF	78.4345	82.2938	82.0225	0.37494
SVM	**79.1533**	**84.1234**	**83.7744**	**0.40199**

从表 4 中可知。SVM 分类方法在 PDNA-226 数据集上,无论 Acc、MCC 评价指标还是从 Sen 与 Spe 评价指标来看,SVM 分类方法都比 RBF 分类器预测效果要好。

同时,SVM 方法引入了结构风险,并采用核映射的思想,因此其和传统分类方法相比,克服了传统方法的大样本需求,同时有效地避免了维数灾难与局部极小的问题,在解决非线性问题上具有卓越的优越

性。综上所述,本文最终选择 SVM 方法作为最终的预测方法。

4.5　与其他预测方法结果对比

将本文所使用的方法和目前预测能力较为强大的两大预测方法,在 PDNA-226 数据集上对 DNA 绑定残基进行预测,其结果对比如表 5 所示。

表 5　在 PDNA-226 数据集上与其他预测方法的结果对比

Classifier	Sen/%	Spe/%	Acc/%	MCC
BindN + [4]	66.7	77.1	76.1	0.26
PreDNA [14]	76.1	82.2	81.8	0.35
本文方法	**79.1**	**84.1**	**83.8**	**0.40**

从表 5 中可以发现,本文所使用的方法,与目前预测能力较为强大的两大预测方法相比,效果要高出不少。与表 5 中预测能力最强,同时也是结合序列与结构信息的 PreDNA 方法相比,本文的方法在特异性、灵敏度、准确性和相关系数方面,分别从 76.1%,82.2%,81.8%,0.35 提升到 79.1%,84.1%,83.8%,0.401,分别提升了 3,1.9,2 和 5 个百分点。对比结果说明,本文所采用的方法在预测 DNA 绑定残基上具有较好的性能。

5　结束语

本文主要探讨了一种新的 DNA 与蛋白质的结合位点预测方法。与一般的预测方法不同,不仅仅使用序列或结构特征,而是通过实验,选择数个组合效果良好的序列与结构特征,将其组合使用。并对训练集使用随机下采样方法进行处理,最后使用支持向量机来进行预测。通过对实验结果的分析以及和其他目前常用的预测器的结果对比,可以说明本文所提出的方法的有效性。

参考文献:

[1] Ptashne M. Regulation of transcription: From lambda to eukaryotes [J]. Trends in Biochemical Sciences, 2005,30(6): 275-279.

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

[3] Hwang S, Gou Z, Kuznetsov I B. DP-Bind: A Web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins [J]. Bioinformatics, 2007,23(5): 634-636.

[4] Wang L, Huang C, Yang M Q. BindN + for accurate prediction of DNA and RNA-binding residues from protein sequence features [J]. BMC Systems Biology, 2010,4(Suppl1): S3.

[5] Wang Liangjiang, Yang M Q, Yang J Y. Prediction of DNA-binding residues from protein sequence information using random forests [J]. BMC Genomics, 2009,10(Suppl 1): S1.

[6] Yu Dong-jun, Hu Jun, Wu Xiao-Wei, et al. Learning protein multi-view features in complex space [J]. Amino Acids, 2013,44(5): 1365-1379.

[7] Lee B, Richards F M. The interpretation of protein structures: Estimation of static accessibility [J]. Journal of Molecular Biology, 1971,55(3): 379-400,IN3-IN4.

[8] Ahmad S, Gromiha M M, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information [J]. Bioinformatics, 2004,20(4): 477-486.

[9] Meshkin A, Ghafuri H. Prediction of relative solvent accessibility by support vector regression and best-first method [J]. EXCLI Journal, 2010,9: 29-38.

[10] Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families [J]. Proteins, 1994,20(3): 216-226.

[11] Pintar A, Carugo O, Pongor S C X. An algorithm that identifies protruding atoms in proteins [J]. Bioinformatics, 2002,18(7): 980-984.

[12] Pintar A, Carugo O, Pongor S C X. DPX: For the analysis of the protein core [J]. Bioinformatics, 2003,19(2): 313-314.

[13] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995,20(3): 273-297.

[14] Li Tao, Li Qianzhong, Shuai Liu, et al. PreDNA: Accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information [J]. Bioinformatics, 2013,29(6): 678-685.

[15] Laurikkala J. Improving identification of difficult small classes by balancing class distribution [C] // Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine. 2001: 63-66.

[16] Weiss G M, Provost F. The effect of class distribution on classifier learning: An empirical study [EB/OL]. http://www.inf.ed.ac.uk/teaching/courses/dme/studpres/leegon.pdf, 2001-09-10.

[17] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets [J]. Computational Intelligence, 2004,20(1): 18-36.