Hans Journal of Computational Biology 计算生物学, 2020, 10(2), 21-30
Published Online June 2020 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2020.102003

Hans汉斯

DNA结合蛋白特征提取算法综述

陈鹏丞，高　雅，倪建威，张艳萍*

河北工程大学数理科学与工程学院，河北　邯郸
Email: *zhangyanping@hebeu.edu.cn

收稿日期：2020年7月21日；录用日期：2020年8月7日；发布日期：2020年8月14日

摘　要

DNA结合蛋白的识别与预测对于研究生物体的生命活动，理解生命活动内在机理具有十分重要的作用。随着蛋白质序列数目的快速增加，计算方法比传统实验方法具有更大的优势。本文从蛋白质的序列信息和结构信息入手，对目前DNA结合蛋白特征提取方法进行归纳总结。在PDB1075和PDB186数据集上，利用XGBoost算法对9种蛋白质序列特征提取方法进行对比分析。结果显示，不同的特征提取方法具有各自的优势与不足，其中，基于蛋白质序列进化信息的Local_DPP方法综合表现最好。

关键词

DNA结合蛋白，特征提取，序列信息，结构信息

An Overview of DNA-Binding Protein for Feature Extraction Algorithms

Pengcheng Chen, Ya Gao, Jianwei Ni, Yanping Zhang*

School of Mathematical Science and Engineering, Hebei University of Engineering, Handan Hebei
Email: *zhangyanping@hebeu.edu.cn

Received: Jul. 21st, 2020; accepted: Aug. 7th, 2020; published: Aug. 14th, 2020

Abstract

The recognition and prediction for DNA-binding proteins play a very important role in studying and understanding the internal mechanisms life activities. The huge numbers of protein sequences have been produced. Computational method has greater advantages than traditional experimental methods. In this paper, we summary the existed methods of DNA-binding protein for

*通讯作者。

文章引用: 陈鹏丞, 高雅, 倪建威, 张艳萍. DNA 结合蛋白特征提取算法综述[J]. 计算生物学, 2020, 10(2): 21-30.
DOI: 10.12677/hjcb.2020.102003

feature extraction based on the sequence information and structural information of the protein. The XGBoost algorithm is employed to compare and analyze the nine feature extraction methods of protein sequence on the PDB1075 and PDB186 datasets. The results demonstrate that different feature extraction methods have their own advantages and disadvantages. Among them, the Local_DPP method based on the evolution information of protein sequences has the best comprehensive prediction performance.

Keywords

DNA-Binding Protein, Feature Extraction, Sequence Information, Structure Information

1. 引言

DNA 结合蛋白是一类特殊的蛋白质，它能够与 DNA 相结合，并通过两者之间相互作用，实现 DNA 转录、复制等功能，进而对生物体的生命活动进行调控[1]，因此对 DNA 结合蛋白的识别研究，能够帮助人们更好地理解核酸和蛋白质之间相互作用的原理，进而帮助医学工作者确定疾病产生的原因及内在机理，并找到相应的靶点和基因片段，这对生物制药、精准医疗等相关领域的发展有着深远的意义[2]。

随着生物测序技术的快速发展，蛋白质序列数目急剧增加。传统的物化实验方法耗时极长且代价高昂，因此大量的 DNA 结合蛋白计算方法被提出和改进。这类方法主要从提取蛋白质序列相关信息出发，建立 DNA 结合蛋白的识别预测模型[3]。在过去十几年，出现了大量识别 DNA 结合蛋白的计算方法，本文将从特征提取的角度出发，对这些方法进行总结，并选择其中九种典型的计算方法进行实证分析，确定这些方法的实际结果，为 DNA 结合蛋白分类算法研究者提供新的思路和有价值的参考。

2. DNA 结合蛋白特征提取方法

目前根据研究人员使用信息的不同可以将方法大致分为两类：一类是基于蛋白质结构信息的方法[4]-[13]，另一类是基于蛋白质序列信息的方法[14]-[23]。

2.1. 基于结构信息的方法

基于结构信息的方法从蛋白质的二级结构和三维空间结构出发，将目标蛋白质与已有的蛋白质结构模板进行对比，提取相关信息，实现对蛋白质的识别与预测。2008 年，Kurgan 等人[4]首次基于蛋白质的二级结构信息构造了 α 螺旋(β 折叠)长度、平均 α 螺旋(β 折叠)长度等特征。Yang 等[5]提出基于蛋白质二级结构片段信息、定量递归分析和 k-字信息熵的特征表示方法。Dai 等[6]提出了基于蛋白质二级结构元素位置统计量的特征表示方法。Stawiski 等[7]提出使用保守残基、氢键电位的结构信息方法，该方法不依赖于序列或结构的同源性。Ahmad 等[8]开发了基于蛋白质的净电荷，电偶极矩和四极矩张量为特征的神经网络预测模型。Szilágyi 等[9]从氨基酸的组成、部分氨基酸空间分布的不对称性以及分子的偶极矩提取特征向量，该方法即使在结构信息不甚准确的蛋白质模型上也能提供较好的预测。Shanahan 等人[10]通过蛋白质结构比对将 DNA 结合蛋白分为 HTH 结构、HHH 结构和 HLH 结构，这项研究为 DNA 结合蛋白预测的后续发展奠定了基础。Gao 等人[11]在结构比对算法之上提出了 DBD-Hunter 方法，该方法结

合了结构对比和统计趋势的估计，在对 4000 种蛋白质的识别预测中，其准确率达到了 98%，性能优良，但此方法需要目标蛋白质的结构作为特征输入，为了克服这一障碍，次年该研究团队开发了一种基于线程的方法 DBD-Threader (DNA-Binding-Domain-Threader) [12]。DBD-Threader 仅需要目标蛋白质的序列信息，在对 179 种 DNA 结合蛋白和 3797 种非 DNA 结合蛋白的测试中，使用与靶序列同一性小于 30%的模板，DBD-Threader 方法的灵敏度和精确度分别达到 56%和 86%，性能优于标准序列比较方法 PSI-BLAST，可与需要目标结构作为输入的 DBD-Hunter 方法相媲美。之后，Zhao 等人[13]在 DFIRE 能量函数的基础上引入新的体积分数校正，从 DNA 结合蛋白的复杂结构中提取新的信息，并进一步提出蛋白质与 DNA 之间的结合亲和力。

这些基于结构信息的预测方法都取得了较高的准确率，但是只能对结构已知的蛋白质进行研究，而蛋白质的结构信息不容易获得，所以这种方法很难在后基因时代进行推广。另外在识别的过程中，蛋白质的分类目标常常与数据库中的蛋白质结构没有相似性，因此需要不依赖与其他已知蛋白质的序列或折叠相似性的新型预测方法。

2.2. 基于序列信息的方法

已有大量的实验结果表明，蛋白质的一级结构(序列排列顺序)相似，其功能也很相似，所以目前更多的方法使用序列信息预测蛋白质的功能，并在拥有海量序列数据的后基因时代得到了很好的发展。在 DNA 结合蛋白识别模型中，基于序列的特征提取方法大致可分为三类：基于氨基酸组成的方法(Amino acid composition)、基于氨基酸物化性质的方法(Physical and chemical properties)以及基于蛋白质序列进化信息的方法(Evolution information)。

2.2.1. 基于氨基酸组成的方法

氨基酸组成是生物信息学领域中广泛使用的特征。对于一个长度为 L 的蛋白质序列 $P = R_1R_2R_3\cdots R_L$，其中 R_1、R_1、…、R_L 分别代表蛋白质的第 1、第 2、…、第 L 个位置的氨基酸。

氨基酸组成(Amino acid composition)表示构成蛋白质的 20 种天然氨基酸在序列中出现的频率，可根据如下公式计算：

$$AAC_j = \frac{1}{L}\sum_{i=1}^{L} R(i,j), i \le 20 \tag{1}$$

$$R(i,j) = \begin{cases} 1, & \text{if } s_j = R_j \\ 0, & \text{else} \end{cases}$$

其中 R_j 表示序列中的氨基酸，s_j 是 20 种天然氨基酸中的一种，通过公式的转换可以得到 20 维特征。伪氨基酸组成。

基于氨基酸组成的方法在实现上相对比较简单，但这种方法单独使用的效果一般，需要和其它特征组合使用。如 Zhang 等[14]使用氨基酸组成和其它特征提出了 newDNA-prot 方法。为了体现氨基酸在序列中的位置信息，Chou 等[15]提出了伪氨基酸组成(Pseudo-amino acid Comiposition, PseAAC)思想，通过加入离散数反映蛋白质的序列顺序，进一步提高了分类效率，详细信息可参考该论文。自从伪氨基酸组成的概念出现后，其转换模型也被蛋白质领域的研究人员不断引用和改善。2011 年，PROFEAT 中已经增加了对于伪氨基酸组成特征的提取功能，特征预测的结果依赖于所选氨基酸的物理化学性质和具体参数取值[16]。

2.2.2. 基于氨基酸物化性质的方法

2018 年公布的 AAindex 数据库中，收录了 554 个关于 20 种天然氨基酸的物理化学性质指标值。基

于氨基酸物化性质的方法通过各种数学映射将蛋白质序列转化为特征向量。目前经常使用的变换方法包括基于组成转换分布的方法(CTD)、基于物化性质距离转换的方法(PDT)和基于自相关指数的变换方法等。Wang 等人[17] [18]选用了六种物理化学性质，即疏水性、氨基酸侧链体积、极性、极化率、溶剂可达比表面积和净电荷指数，并以此提出预测 DNA 结合蛋白的 NMBAC 模型。该方法首先对指标值进行中心标准化，然后利用以下公式得到蛋白质序列的 NMBAC 特征。

$$\mathrm{NMBAC}_{lag,j}=\frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j}*X_{i+lag,j}\right)\ \left(i=1,2,\cdots,n-lag;\ j=1,2,\cdots,6\right) \tag{2}$$

其中 j 代表物理化学性质，i 代表蛋白质序列 X 的位置，n 代表蛋白质序列的长度，lag 为两残基之间的顺序距离。

2.2.3. 基于蛋白质序列进化信息的方法

基于蛋白质序列进化信息的方法，主要是从氨基酸突变概率出发，对蛋白质序列进行比对分析。该方法首先将目标蛋白质序列在蛋白质数据库中进行比对，然后产生位置特异性打分矩阵(PSSM, position-specific scoring matrices)，进一步从打分矩阵中提取进化信息。目前比较权威的蛋白质数据库是 NCBI 构建的 NR 数据库，它整合了来自 GenPept，Swissprot，PIR，PDB 等数据库中的数据，可通过在线的 ftp 服务器 https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/访问。首先利用 PSI-BLAST 得到打分矩阵，之后通过数学公式将其转化为特征向量并结合分类算法进行使用。这类方法的优点是精确性比较高，但由于需要进行多序列比对，通常也有更高的时间和空间成本。

对于长度为 L 的蛋白质序列 S，可以表示为 $S_1S_2\cdots S_L$，其 PSSM 定义为：

$$P=\begin{bmatrix}p_{1,1} & p_{1,2} & \cdots & p_{1,20}\\ p_{2,1} & p_{2,2} & \cdots & p_{2,20}\\ \vdots & \vdots & \ddots & \vdots\\ p_{L,1} & p_{L,1} & \cdots & p_{L,20}\end{bmatrix}_{L\times 20}$$

PSSM 中的坐标 (i,j) 表示在替代过程中，位置i处的氨基酸被位置j处的氨基酸替代的对数似然得分。当矩阵中取值为正整数时，表示在替代的过程中蛋白质序列的相应位点上的氨基酸比较容易突变为对应横坐标上的 20 种氨基酸，其数值越大，表明此处替换的概率越大。当取值为负整数时，则与之相反，负值越大，表明此处越不易发生改变。位置特异性得分矩阵提供了一种直观衡量蛋白质序列上一个残基突变为其他类型氨基酸的倾向性。

在蛋白质序列进化信息的方法中，PSSM400 [1]是最常使用的特征之一，它需要对 PSSM 进行 Min-Max 标准化，然后分别计算序列中每一种氨基酸突变为 20 种氨基酸的概率，可得到 20 × 20 = 400 维特征。

PSSM400 只是简单计算突变概率累加得到特征，无法充分发掘 PSSM 中的其他隐含信息。因此有学者提出用离散数字表示蛋白质样本，如 Chou [19]等人提出的 Pse-PSSM，通过对 PSSM 离散化充分探索 PSSM 中嵌入的进化信息和序列信息，提高了膜蛋白预测分类的准确率。Local_DPP [20]在此基础上进行探索，把标准化后的 PSSM 进行分割，最大程度的提取局部保守信息。首先，它将 PSSM 划分为 n 个部分，前 $n-1$ 个部分的长度为 L/n，第 n 部分的长度为 $L-(n-1)\times(L/n)$，通过 $Part_1$ 和 $Part_2$ 的计算公式可以得到特征。

$$Part_1=\left\{F_j(k)=\frac{1}{length(k)}\sum_{i=0}^{length(k)-1}f_{i,j}\ \left(j=1,2,\cdots,20\right)\right\} \tag{3}$$

在公式中，k 代表第 k 个子矩阵，$length(k)$ 代表第 k 个子矩阵的长度，$f_{i,j}$ 代表 PSSM 中的归一化值。$F_j(k)$ 表示进化过程中分段序列中每个位置第 j 种氨基酸出现的平均概率。

$$Part_2=\left\{\phi_j^{\xi}(k)=\frac{1}{length(k)-\xi}\sum_{i=0}^{length(k)-1-\xi}\left(f_{i,j}-f_{i+\xi,j}\right)^2\left(\xi=1,2,\cdots,\lambda;1<\lambda<length(k)\right)\right\} \tag{4}$$

$\varphi_j^{\xi}=(k)$ 表示被 ξ 分隔的两个残基之间第 j 类氨基酸的平均相关性。例如，对于第 k 个子矩阵中的氨基酸类型j，$\varphi_j^1=(k)$ 是通过沿着蛋白质链最近邻残基相关因子。$Part_1$ 包含局部进化信息，$Part_2$ 包含序列顺序信息。前 $n-1$ 个子矩阵的特征空间表示由下式给出：

$$FV(n-1)=(Part_1,Part_2) \tag{5}$$

最后一个子矩阵特征如下所示：

$$FV(n)=\left(F_1(n),\cdots,F_{20}(n),\Phi_1^1(n),\cdots,\Phi_{20}^1(n),\cdots\Phi_1^{\lambda}(n),\cdots,\Phi_{20}^{\lambda}(n)\right) \tag{6}$$

其中，$F_j(n)$ 和 Φ_j^{ξ} 的计算方法与前 $n-1$ 个子矩阵的计算方法相同，最终的特征向量为：

$$FV(n)=\left(FV(n-1),FV(n)\right) \tag{7}$$

另外一种基于 PSSM 构造的 PSSMDCT [18]特征提取方法，通过离散余弦变换(DCT)对 PSSM 进行处理，并将部分压缩后的 PSSM 保留为特征向量。离散余弦变换(DCT for Discrete Cosine Transform)类似于离散傅里叶变换(DFT for Discrete Fourier Transform)，但是只使用实数，它可以将信息密度的分布从均匀分布改为不均匀分布。在此条件下，对 PSSM 进行压缩，保留 PSSM 的低频部分作为最终特征(因为低频部分比高频部分包含更多信息)。当输入原始的 PSSM 之后，相应的转换公式如下：

$$DCT(i,j)=\alpha_i,\alpha_j\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}Mat(m,n)\cos\frac{\pi(2m+1)i}{2M}*\cos\frac{\pi(2n+1)j}{2n} \tag{8}$$

$$\alpha_i=\begin{cases}\sqrt{1/M}, & i=0\\ \sqrt{2/M}, & 1\le i\le M-1\end{cases}$$

$$\alpha_j=\begin{cases}\sqrt{1/n}, & j=0\\ \sqrt{2/n}, & 1\le j\le N-1\end{cases}$$

根据上述压缩公式，包含大部分信息的部分(低频部分)被压缩在 PSSM 的左上角。保留前 100 个系数作为 PSSM-DCT 的最终特征。与此相似的方法还有 PSSM-DWT [18] (基于离散小波变换的方法)，PSSM-DBT 和 PSSM-AB [24]。

而 Sliding window [21]在原始 PSSM 的第一列和最后一列加入零向量，通过选择合适的长度 w 对原始 PSSM 进行滑动拼接，形成标准的 PSSM，并通过选择合适的窗口长度 ws 形成平滑的 PSSM，计算每列的均值、标准差、极差和上下四分位点，构成最终特征。滑动窗口的思想在 DNA 结合残基的识别预测中应用广泛，考虑了氨基酸残基间的影响，认可度较高。

3. 实证分析

3.1. 数据集

本文选用 PDB1075 [22]和 PDB186 [23]作为实证分析的数据集，具体见表 1。数据集 PDB1075 和 PDB186 中的蛋白质序列均来自 PDB 数据库(http://www.rcsb.org/pdb/home/home.do)，PDB1075 含有 525 个 DNA 结合蛋白和 550 个非 DNA 结合蛋白，PDB186 包含 93 条 DNA 结合蛋白序列和 93 条非 DNA 结

合蛋白序列。为了保证分类结果的准确性，以上两个数据集使用 PISCES 减少蛋白质序列间的冗余度，相似性低于 25%。

Table 1. PDB1075 and PDB186 datasets
表 1. PDB1075 和 PDB186 数据集

Data	DNA-binding proteins	Non-DNA-binding proteins	Total
PDB1075	525	550	1075
PDB186	93	93	186

3.2. 分类器

XGBoost 算法是以 CART 为基分类器的集成学习方法之一，由于其出色的运算效率和预测准确率在数据建模比赛中得到广泛的应用。与随机森林赋予每一颗决策树相同的投票权重不同，XGBoost 算法中下一棵决策树的生成和前一棵决策树的训练和预测相关(通过对上一轮决策树训练准确率较低的样本赋予更高的学习权重来提高模型准确率)。相比于其他集成学习算法，XGBoost 一方面通过引入正则项和列抽样的方法提高了模型稳健性，另一方面又在每棵树选择分裂点时采取并行化策略从而极大提高了模型运行的速度。

3.3. 评价指标

为了防止过拟合，本文在训练数据集 PDB1075 中采用五折交叉验证方法建立预测模型。选取以下 5 个性能指标对模型进行评价，包括准确率(ACC)，灵敏度(SN)，特异性(SP)，马修斯相关系数(MCC)，AUC，具体计算公式如下：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\times 100\%$$

$$\mathrm{SN}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\%$$

$$\mathrm{SP}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\times 100\%$$

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

其中，TP和TN分别是正确预测 DNA 结合蛋白和非结合蛋白的数目，FP表示预测为 DNA 结合蛋白的非结合蛋白，FN表示预测为非结合蛋白的 DNA 结合蛋白，这些指标值越高意味着更好的预测结果。

上述指标大多数依赖于阈值的选择，通过确定阈值将预测的概率分为正类和负类，而AUC(ROC 曲线下的面积)不受阈值影响。因此，与其它指标相比，AUC指数可以为预测提供更全面的视角。

3.4. 对比结果分析

本文主要选取基于氨基酸组成的频率特征 AAC、基于氨基酸物化性质的特征 NMBAC、伪氨基酸组成特征 PseAAC 及基于蛋白质进化信息的 PSSM400、Local_DPP、PSSM-DCT、PSSM-DWT、PSSM-AB、Sliding window 进行对比分析。根据上述 9 种特征提取方法将蛋白质序列转化为特征向量，在训练集 PDB1075 上建立预测模型，并在 PDB186 上进行测试，几种方法的性能比较如表 2、表 3 所示。

在训练集上，蛋白质进化信息的 Local_DPP 方法综合表现最优，其预测准确率达到 77.2%，灵敏度

为 78.5%，MCC 指标为 0.544，AUC 值为 0.82。其余方法中，Sliding window 灵敏度最好，为 81.7%。PSSM-DWT 特异性和 AUC 最高，分别为 77.2%和 0.84。综合比较，可以发现蛋白质进化信息的方法要优于氨基酸组成和氨基酸物理化学性质的方法。

从图 1 的 ROC 曲线可以看出，各类特征提取方法的 AUC 值均大于 0.7，表明这些方法都有着较好的分类结果。其中 PSSM-DWT 的 AUC 最高，达到了 0.84。Local_DPP、PSSM-AB 和 PseAAC 方法得到的 AUC 值为 0.82，并列第二位。AUC 指标表现最差的是氨基酸物化性质的方法 NMBAC，仅有 0.76。

Table 2. Classification results of PDB1075 data set
表 2. 训练数据集的分类结果

类别	特征	ACC	SN	SP	MCC	AUC
氨基酸组成	AAC	73.9	72.5	69.8	0.468	0.80
伪氨基酸组成	PseAAC	73.56	76.3	73.2	0.482	0.82
氨基酸物化性质	NMBAC	72.6	73.6	68.6	0.457	0.76
蛋白质进化信息	PSSM-AB	75.3	77.4	76.5	0.526	0.82
	Local_DPP	77.2	78.5	75.2	0.554	0.82
	PSSM-DWT	75.4	76.5	77.2	0.517	0.84
	PSSM-DCT	69.3	73.6	69.1	0.418	0.78
	Sliding window	72.5	81.7	59.8	0.458	0.81
	PSSM400	76.2	75.3	76.8	0.508	0.79

Table 3. Classification results of PDB186 data set
表 3. 测试数据集的分类结果

类别	特征	ACC	SN	SP	MCC	AUC
氨基酸组成	AAC	63.2	67.8	65.3	0.26	0.68
伪氨基酸组成	PseAAC	73.12	75.12	73.2	0.468	0.74
氨基酸物化性质	NMBAC	74.2	77.2	73.6	0.504	0.75
蛋白质进化信息	PSSM-AB	74.2	79.3	74.2	0.531	0.76
	Local_DPP	78.3	88.3	67.2	0.584	0.82
	PSSM-DWT	76.8	81.2	77.2	0.569	0.79
	PSSM-DCT	65.3	70.2	56.9	0.297	0.70
	Sliding window	67.3	72.6	68.2	0.315	0.72
	PSSM400	61.5	69.9	53.8	0.232	0.65

在测试集上，Local_DPP 方法的预测结果最好，准确率、灵敏度、MCC、AUC 值分别为 78.3%，88.3%，0.584，0.82。PSSM-DWT 方法的准确率为 76.8%。另外几种蛋白质进化信息的方法准确率较低，如 PSSM-DWT 方法的准确率为 65.3%，PSSM400 的准确率为 61.5%，Sliding window 方法的准确率为 67.3%。物理化学性质 NMBAC 方法和伪氨基酸组成 PseAAC 方法预测准确率分别为 74.3%、73.12%，高于部分进化信息的方法。综上所述，不同的特征计算方法对于 DNA 结合蛋白的识别预测具有十分显

著的影响。

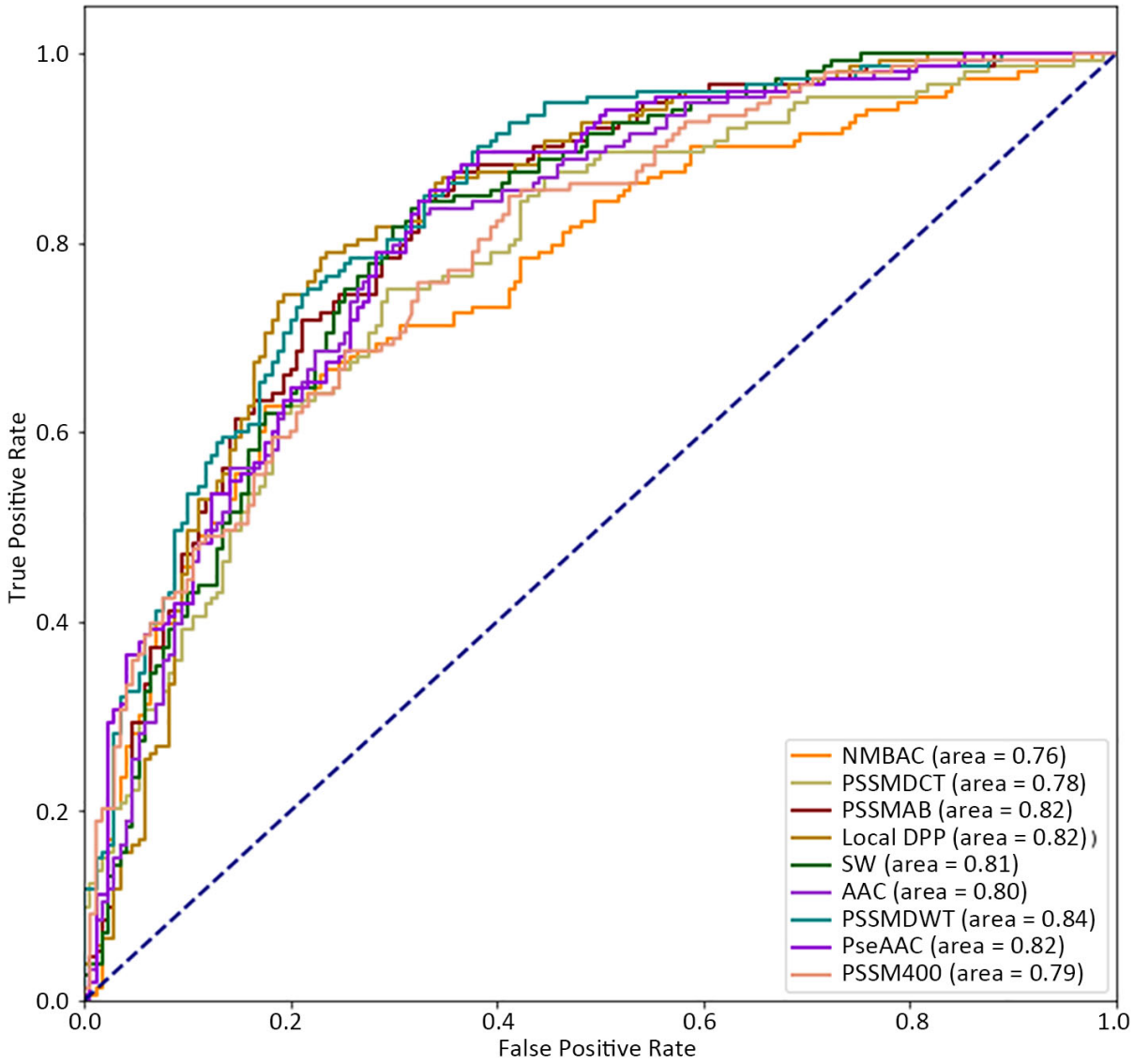

Figure 1. ROC curves of nine methods
图 1. 九种方法的 ROC 曲线图

4. 结论

基于结构信息的特征提取方法由于应用条件的苛刻性，很难在后基因时代推广，所以现阶段的方法更多注重从蛋白质序列中挖掘内在的生物信息。在三类基于蛋白质序列信息的提取方法中，基于蛋白质序列进化信息的方法综合性能优于氨基酸组成和氨基酸理化性质的方法，Local_DPP 方法在训练集和测试集上的综合性能最好。PseAAC 方法组合了氨基酸组成信息和理化性质，其特征仅有 25 维，但在分类准确率上超过了 70%，所以在组合特征时可以优先考虑使用。

总的来说，虽然目前已提出很多基于序列信息的 DNA 结合蛋白预测方法，但其预测性能仍不能令人满意。蛋白质序列特征提取方法是蛋白质结构和功能识别方法性能提升的瓶颈，如何有效地表示蛋白质序列仍需要进一步的研究。

基金项目

本文得到了河北省自然科学基金项目(F2019402078)和河北省高等学校科学技术研究项目(QN2018235)的支持，在此表示感谢。

参考文献

[1] Kumar, M., Gromiha, M.M. and Raghava, G.P.S. (2007) Identification of DNA-Binding Proteins Using Support Vector Machines and Evolutionary Profiles. *BMC Bioinformatics*, **8**, Article No. 463. https://doi.org/10.1186/1471-2105-8-463

[2] 汤希玮. 蛋白质复合物识别算法综述[J]. 长沙大学学报, 2017, 31(5): 19-23.

[3] 张军. 基于序列信息的 DNA/RNA 结合蛋白识别[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2018.

[4] Kurgan, L.A., Cios, K.J. and Chen, K. (2008) SCPRED: Accurate Prediction of Protein Structural Class for Sequences of Twilight-Zone Similarity with Predicting Sequences. *BMC Bioinformatics*, **9**, Article No. 226. https://doi.org/10.1186/1471-2105-9-226

[5] Yang, J.-Y., Peng, Z.-L. and Chen, X. (2010) Prediction of Protein Structural Classes for Low-Homology Sequences Based on Predicted Secondary Structure. *BMC Bioinformatics*, **11**, Article No. S9. https://doi.org/10.1186/1471-2105-11-S1-S9

[6] Dai, Q., Li, Y., Liu, X., Yao, Y., Cao, Y. and He, P. (2013) Comparison Study on Statistical Features of Predicted Secondary Structures for Protein Structural Class Prediction: From Content to Position. *BMC Bioinformatics*, **14**, Article No. 152. https://doi.org/10.1186/1471-2105-14-152

[7] Szilágyi, A. and Skolnick, J. (2006) Efficient Prediction of Nucleic Acid Binding Function from Low-Resolution Protein Structures. *Journal of Molecular Biology*, **358**, 922-933. https://doi.org/10.1016/j.jmb.2006.02.053

[8] Stawiski, E.W., Gregoret, L.M. and Mandel-Gutfreund, Y. (2003) Annotating Nucleic Acid-Binding Function Based on Protein Structure. *Journal of Molecular Biology*, **326**, 1065-1079. https://doi.org/10.1016/S0022-2836(03)00031-7

[9] Ahmad, S. and Sarai, A. (2004) Moment-Based Prediction of DNA-Binding Proteins. *Journal of Molecular Biology*, **341**, 65-71. https://doi.org/10.1016/j.jmb.2004.05.058

[10] Shanahan, H.P., Garcia, M.A., Jones, S. and Thornton, J.M. (2004) Identifying DNA-Binding Proteins Using Structural Motifs and the Electrostatic Potential. *Nucleic Acids Research*, **32**, 4732-4741. https://doi.org/10.1093/nar/gkh803

[11] Gao, M. and Skolnick, J. (2008) DBD-Hunter: A Knowledge-Based Method for the Prediction of DNA-Protein Interactions. *Nucleic Acids Research*, **36**, 3978-3992. https://doi.org/10.1093/nar/gkn332

[12] Gao, M. and Skolnick, J. (2009) A Threading-Based Method for the Prediction of DNA-Binding Proteins with Application to the Human Genome. *PLoS Computational Biology*, **5**, e1000567. https://doi.org/10.1371/journal.pcbi.1000567

[13] Zhao, H., Yang, Y. and Zhou, Y. (2010) Structure-Based Prediction of DNA-Binding Proteins by Structural Alignment and a Volume-Fraction Corrected DFIRE-Based Energy Function. *Bioinformatics*, **26**, 1857-1863. https://doi.org/10.1093/bioinformatics/btq295

[14] Zhang, Y., Xu, J., Zheng, W., Zhang, C., Qiu, X., Chen, K. and Ruan, J. (2014) newDNA-Prot: Prediction of DNA-Binding Proteins by Employing Support Vector Machine and a Comprehensive Sequence Representation. *Computational Biology and Chemistry*, **52**, 51-59. https://doi.org/10.1016/j.compbiolchem.2014.09.002

[15] Chou, K.-C. (2001) Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. *Proteins*: *Structure, Function, and Bioinformatics*, **43**, 246-255. https://doi.org/10.1002/prot.1035

[16] Zhang, P., *et al.* (2016) A Protein Network Descriptor Server and Its Use in Studying Protein, Disease, Metabolic and Drug Targeted Networks. *Briefings in Bioinformatics*, **18**, 1057-1070.

[17] Feng, Z.-P. and Zhang, C.-T. (2000) Prediction of Membrane Protein Types Based on the Hydrophobic Index of Amino Acids. *Journal of Protein Chemistry*, **19**, 269-275. https://doi.org/10.1023/A:1007091128394

[18] Wang, Y., Ding, Y., Guo, F., Wei, L. and Tang, J. (2017) Improved Detection of DNA-Binding Proteins via Compression Technology on PSSM Information. *PLoS ONE*, **12**, e0185587. https://doi.org/10.1371/journal.pone.0185587

[19] Chou, K.-C. and Shen, H.-B. (2007) MemType-2L: A Web Server for Predicting Membrane Proteins and Their Types by Incorporating Evolution Information through Pse-PSSM. *Biochemical and Biophysical Research Communications*, **360**, 339-345. https://doi.org/10.1016/j.bbrc.2007.06.027

[20] Wei, L., Tang, J. and Zou, Q. (2017) Local-DPP: An Improved DNA-Binding Protein Prediction Method by Exploring Local Evolutionary Information. *Information Sciences*, **384**, 135-144. https://doi.org/10.1016/j.ins.2016.06.026

[21] Wang, C., Fang, Y., Xiao, J. and Li, M. (2011) Identification of RNA-Binding Sites in Proteins by Integrating Various Sequence Information. *Amino Acids*, **40**, 239-248. https://doi.org/10.1007/s00726-010-0639-7

[22] Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X. and Chou, K.-C. (2014) iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. *PLoS ONE*, **9**, e106691. https://doi.org/10.1371/journal.pone.0106691

[23] Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B. and Zhang, H. (2014) Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. *PLoS ONE*, **9**, e86703. https://doi.org/10.1371/journal.pone.0086703

[24] Zou, Y., Ding, Y., Tang, J., Guo, F. and Peng, L. (2019) FKRR-MVSF: A Fuzzy Kernel Ridge Regression Model for Identifying DNA-Binding Proteins by Multi-View Sequence Features via Chou's Five-Step Rule. *International Journal of Molecular Sciences*, **20**, 4175. https://doi.org/10.3390/ijms20174175