Hans Journal of Computational Biology 计算生物学, 2021, 11(1), 1-11
Published Online March 2021 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2021.111001

Hans汉斯

grDNA-Prot：基于氨基酸物理化学特性和支持向量机的DNA结合蛋白预测

张艳萍*，倪建威，高　雅，陈鹏丞，李旭涛

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：2021年2月11日；录用日期：2021年3月11日；发布日期：2021年3月23日

摘　要

DNA结合蛋白在细胞内外的各种活动中起着重要作用。本文提出一种新的DNA结合蛋白预测方法(grDNA-Prot)，使用20个氨基酸组成频率和基于AAindex数据库531个氨基酸物理化学性质的图形表示法描述蛋白质序列信息。此外，还采用三种特征选择方法来选择最优特征，并通过5折交叉验证，建立了基于支持向量机的DNA结合蛋白识别预测模型。为验证该方法的有效性，本文在独立测试数据集上与其他方法进行了比较。这些结果表明，Hydrophobicity (H)、Physicochemical properties (P)和Alpha and turn properties (A)是有效区分DNA结合蛋白和非DNA结合蛋白的主要氨基酸物理化学性质。

关键词

DNA结合蛋白，物理化学性质，图形表示法，特征选择，支持向量机

grDNA-Prot: The Prediction of DNA-Binding Proteins Based on Physicochemical Properties of Amino Acids and Support Vector Machine

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

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

Received: Feb. 11th, 2021; accepted: Mar. 11th, 2021; published: Mar. 23rd, 2021

*通讯作者。

文章引用: 张艳萍, 倪建威, 高雅, 陈鹏丞, 李旭涛. grDNA-Prot: 基于氨基酸物理化学特性和支持向量机的 DNA 结合蛋白预测[J]. 计算生物学, 2021, 11(1): 1-11. DOI: 10.12677/hjcb.2021.111001

Abstract

DNA-binding proteins played an important role in various intra- and extra-cellular activities. In this paper, a novel grDNA-Prot method of DNA-binding predictor is proposed, the protein sequence information is described with the probabilities of 20 amino acids and the 531 physicochemical properties indices of 20 amino acids in AAindex database based on the Cylindrical graphical representation. Furthermore, we employ three feature selection methods to select the optimal feature, which is used to establish the model for identify DNA-binding proteins basing on support machine vector with 5-fold cross-validation. In order to test the effectiveness of our method, we compare the accuracy performance with the other methods in independent test dataset. These results demonstrated that the physicochemical properties of hydrophobicity (H), Physicochemical properties (P) and the alpha and turn properties (A) are primarily responsible for distinguishing between DNA-binding proteins and non DNA-binding proteins.

Keywords

DNA-Binding Proteins, Physicochemical Properties, Graphical Representation, Feature Selection, SVM

1. 引言

DNA 结合蛋白在细胞内外各种生命活动中扮演着重要角色，例如：转录调控、DNA 复制、DNA 包装、DNA 修复与重组。这些独立折叠结构域中的蛋白质至少有一个结构基序，并且与 DNA 有亲和力[1]。DNA 结合蛋白或配体可以作为抗生素、药物、类固醇等多种生物化学物质应用于 DNA 的生物物理、生化和生物学研究中[2]。1986 年，Swiss-Prot 数据库只包含 3939 条蛋白质序列[3]，但是截止 2020 年 10 月 7 日 UniProtKB/Swiss-Prot 已经收录了 563,972 条蛋白质序列。基于实验的 DNA 结合蛋白检测方法虽然有很多[4] [5] [6] [7]，但存在价格昂贵、检测时间长、设备要求较高的问题[8]。因此，开发一种可以有效区分 DNA 结合蛋白和非 DNA 结合蛋白的计算方法具有十分重要的意义。到目前为止，有很多基于蛋白质结构和序列的计算方法来预测 DNA 结合蛋白。由于获取的限制，利用蛋白质结构预测 DNA 结合蛋白虽然准确但无法进行高通量注释[9] [10]。

近年来，研究人员开发出很多基于蛋白质序列预测 DNA 结合蛋白的计算方法，例如：iDNA-Prot [11]，DNA-Prot [12]，DNA-binder [13]，iDNA-Prot|dis [14]，PSFM-DBT [15]，DeepDRBP-2L [16]。为了从蛋白质序列中提取更多的生物信息，研究人员使用了 20 种氨基酸组成频率、进化保守信息、蛋白质二级结构和氨基酸物理化学性质等特征表示方法构造特征。同时，采用特征选择算法降低特征冗余和相关性，提高预测模型性能。此外，常用于预测 DNA 结合蛋白的机器学习分类算法包括随机森林(Random Forest) [11] [12] [17] [18]、支持向量机(support vector machine, SVM) [19]-[25]、Logistic 回归(Logistic Regression) [26]、朴素贝叶斯(Naive Bayes) [17]、人工神经网络(Artificial Neural Network, ANN) [27]。支持向量机具有较高的稳定性和精确性，被广泛应用于 DNA 结合蛋白预测识别领域。

然而，如何从蛋白质序列中提取序列顺序信息或关键模式是最重要和最困难的问题。Huang 等人[19]从图形表示和氨基酸物理化学性质角度提取信息并使用 mRMR 算法建立蛋白质功能预测模型；Zou 等人

[20]根据组成信息、氨基酸理化性质、进化保守信息和结构功能信息预测 DNA 结合蛋白，从全局序列描述、非局部序列描述、局部序列描述三个角度构建 DNA 结合蛋白特征并建立基于支持向量机的预测模型；Kumar 等人[22]使用氨基酸组成和伪氨基酸组成作为输入特征建立 β-内酰胺酶蛋白质预测模型，取得了较好效果；Liu 等人[28]使用周氏伪氨基酸组成特征和基于距离变化的物理化学性质特征建立 microRNA 预测模型，并可以应用于计算生物学的许多领域。

蛋白质的结构和功能是由 20 种天然氨基酸的物理化学和生物化学性质定义的，这些氨基酸是蛋白质的组成成分。AAindex 数据[29]库包含 544 组氨基酸物理化学性质指数，每组指数由 20 种天然氨基酸数值构成。因此，从 AAindex 数据库中找出可以有效区分 DNA 结合蛋白与非 DNA 结合蛋白的氨基酸物理化学性质具有重要意义[30]。由于测量方法与设备限制，有 13 组氨基酸物理化学指数存在缺失值，无法进行分析。近年来，通过对 AAindex 数据库进行数据挖掘研究，研究者们提出了很多关于蛋白质功能的预测方法[31]。

本文通过柱状图表示法从 531 个氨基酸物理化学性质中提取了 531 个特征，同时还考虑氨基酸组成对蛋白质功能的影响。但是 551 个特征向量中存在冗余和多重相关性，这会增加预测器学习的难度和降低模型的精确度。为此，本文采用了三种特征选择方法来减少 551 个特征向量的冗余信息，即基于 LASSO 的方法[32]、基于过滤器的方法、基于包装器的方法，并结合支持向量机对 DNA 结合蛋白进行预测(5 折交叉验证)。并且所选取的特征通过了 t 检验，证明对于识别 DNA 结合蛋白和非 DNA 结合蛋白具有统计学意义。

为说明 grDNA-Prot 方法的有效性，本文在独立数据集 DNAiset 和 DNArset 上进行了验证分析，并与其他现有 DNA 结合蛋白预测方法 DNAbinder、iDAN-Port、DNA-port 进行对比。其中，DNArset 数据集中非 DNA 结合蛋白数量远大于 DNA 结合蛋白，与生物界中 DNA 结合蛋白与非 DNA 结合蛋白的分布情况相符。结果表明，grDNA-Port 方法优于现有基于序列的方法(DNAbinder、iDAN-Port、DNA-port)。本文 grDNA-Port 方法的分析框架如图 1 所示。

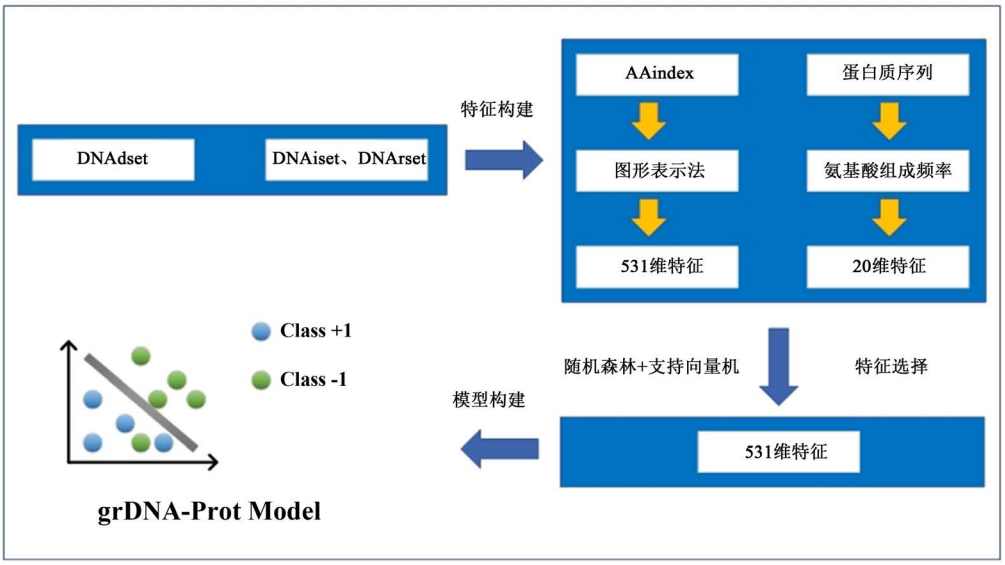

Figure 1. The research framework of this paper
图 1. 本文研究框架

2. 材料与方法

2.1. 数据集

本文将 DNAdset 作为训练集，独立数据集 DNAiset、DNArset 作为验证集，这三个数据集被使用于

多个 DNA 结合蛋白预测研究中[20] [33]。训练集 DNAdset 包含 231 个 DNA 结合蛋白和 231 个非 DNA 结合蛋白，使用 CD-HIT 程序[34]检验该数据集内蛋白质序列相似性低于 40%。Lin 等人[11]在 PDB (Protein Data Bank)数据库中检索关键词“DAN 结合蛋白”得到 2014 年 6 月 1 日以后发布的 97 条 DNA 结合蛋白，并与 199 条非 DNA 结合蛋白混合形成独立数据集 DNAiset。为模拟 DNA 结合蛋白在人体内分布情况，Kumar 等人[13]在 2007 年发布了包含 97 个 DNA 结合蛋白和 1500 个非 DNA 结合蛋白的 DNArset 数据集。数据集 DNAiset 和 DNArset 中序列相似性小于 30%。

2.2. 特征提取

特征包括氨基酸组成和氨基酸物理化学性质两类。许多氨基酸物理化学性质已经成功地应用于蛋白质长无序区、无序蛋白质结合残基的晶体结构注释、RNA 结合残基的晶体结构注释、DNA 结合残基的晶体结构注释等蛋白质结构和功能预测，例如氨基酸的疏水性、溶剂可及性、电荷和自由能。氨基酸的物理化学性质在蛋白质折叠和蛋白质与 DNA 相互作用中起着重要作用，本文使用 531 组氨基酸指数来表示氨基酸的各种物理化学性质。图形表示法[35]通过可视化方法提取蛋白质序列信息，是一种计算成本较低、无需对齐的方法，常常使用氨基酸的物理化学性质。本文在 Yu 等人[36]的工作基础上，用柱形表示法来表示蛋白质序列，将所建立的协方差矩阵特征值的最大值作为蛋白质序列的数值特征。因此，本文通过 531 组氨基酸指数得到 531 个数值特征来表示一条蛋白质序列。

下面将定义柱面图形表示法。20 个氨基酸分布在圆柱体的底圆上，每个氨基酸在圆柱表面形成一条线。这种几何结构显示了氨基酸残基在蛋白质序列中的组成和分布。根据 531 个理化性质指标值对 20 个氨基酸进行排序。我们使用柱坐标来显示单位圆柱表面的蛋白质序列。

柱面坐标和笛卡尔坐标之间的转换如公式 1 所示：

$$x_n = \cos\left(\frac{2\pi}{20} i_n\right),\quad y_n = \sin\left(\frac{2\pi}{20} i_n\right),\quad z_n = \frac{n}{N} \tag{1}$$

其中，N 表示蛋白质序列长度，$n = 1,2,3,\cdots,N$ 表示蛋白质序列中第 n 个氨基酸，$i_n = 0,1,2,\cdots,19$ 表示第 n 个氨基酸具体指标值。以 AAindex 数据库中氨基酸指数 BHAR880101 为例(若两个氨基酸指标值相同则按字母顺序排列)，20 个氨基酸的排序为：M < W < F < H < C < A < L < V < Y < T < I < N < K < Q < E < S < P < D < R < G。假设蛋白质序列为：

MKRRIRRERNKMAAAKSRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKL

则蛋白质序列的柱状图表示如图 2 所示：

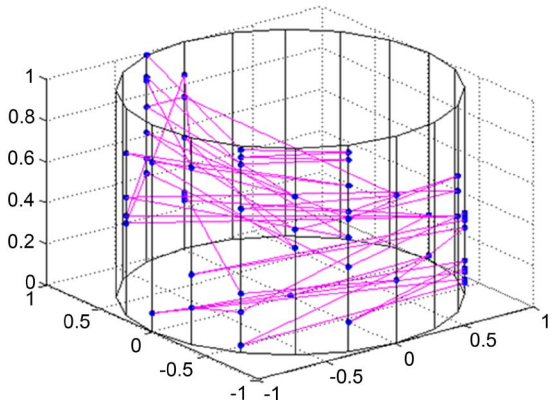

Figure 2. The cylindrical graphical representation of a protein sequence
图 2. 蛋白质序列的柱状图表示

为了更直观地了解蛋白质序列中隐含的生物学特性，并分析它们的相似性与相异性，对蛋白质序列的圆柱形序列求协方差矩阵的最大特征值。协方差矩阵如公式 2 所示：

$$P=\begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{bmatrix} \tag{2}$$

其中：

$$S_{xx}=\sum_n (x_n-\bar{x})^2, S_{xy}=\sum_n (x_n-\bar{x})(y_n-\bar{y})$$

$$S_{yy}=\sum_n (y_n-\bar{y})^2, S_{xz}=\sum_n (x_n-\bar{x})(z_n-\bar{z})$$

$$S_{zz}=\sum_n (z_n-\bar{z})^2, S_{yz}=\sum_n (y_n-\bar{y})(z_n-\bar{z})$$

同时，定义 $F=\max\{\lambda_i\}$，λ_i 为协方差矩阵的特征值 P。根据 20 个氨基酸的 531 个物理化学性质指标值，一条蛋白质序列通过上述方法可以转换成 531 个数值特征。另外，将蛋白质序列中 20 个氨基酸频率作为组成特征。

本文共得到 551 维特征用于 DNA 结合蛋白的预测，表示为 $X=(X_1,X_2,\cdots,X_{549})$。并对特征进行标准化处理，标准化公式如式所示

$$Y_i=\frac{X_i-\bar{X}_i}{\sqrt{Var(X_i)}} \tag{3}$$

其中，$\bar{X}_i$ 和 $Var(X_i)$ 分别为 X_i 的平均值与标准差。

2.3. 支持向量机(Support Vector Machine, SVM)

支持向量机是 DNA 结合蛋白预测领域中应用最广泛的机器学习算法。机器学习算法的原理是在高维特征空间中构造一个超平面，将数据点分为两类或者多类。基于径向基核函数的支持向量机算法已经广泛应用于蛋白质 ATP 结合位点预测等研究[37]。径向基核函数如下所示：

$$K(x_i,y_j)=\exp\left(-\gamma\|x_i-y_i\|^2\right) \tag{4}$$

其中 γ 为径向基核函数的宽度。本文使用网格搜索和交叉验证确定参数 C 与 γ，两个参数的取值范围为 $2^i\left(i\in\{-10,-9,-8,\cdots,8,9,10\}\right)$。

2.4. 性能评估

本文选择支持向量机算法作为分类器，使用五折交叉验证避免过拟合出现，并获得低均方误差的可靠结果。使用的评价指标为 Accuracy (ACC)、Sensitivity, Precision, Specificity, F-measure and Matthews correlation coefficient (MCC) [38]。同时使用 ROC 曲线及 ROC 曲线下面积 AUC 评估分类器预测性能[39]。这些指标由如下公式给出[40]：

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

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

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\%$$

$$\text{F-measure} = \frac{2\times\text{TP}}{2\times\text{TP}+\text{FN}+\text{FP}} \times 100\%$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TN}+\text{FN})(\text{TN}+\text{FP})(\text{TP}+\text{FN})(\text{TP}+\text{FP})}} \times 100\%$$

其中，TP 和 TN 分别代表准确预测为 DNA 结合蛋白和非 DNA 结合蛋白的例数，FP 代表预测为 DNA 结合蛋白的非 DNA 结合蛋白例数，FN 代表预测为非 DNA 结合蛋白的 DNA 结合蛋白例数。TP、TN 越高，FP、FN 越低，代表模型预测效果更好。

2.5. 特征选择

为降低 551 维特征向量之间的冗余、提高预测模型精度，本文在训练数据集上(DNAdset)进行特征选择，得到一组非冗余的特征子集。本文比较了三种特征选择方法：LASSO 方法、基于 Filter 的方法和基于 Wrapper 的方法，使用网格搜索和 5 折交叉验证确定每种方法中的最优参数，从而实现最优化预测性能(以 AUC 为评价指标)。第一种方法是利用 LASSO 回归方法得到特征间无线性相关性的特征子集。第二种方法先使用最大相关度最小冗余度(Maximum Relevance Minimum Redundancy, mRMR)方法对特征进行排序[41]，然后选择排名前 200 维特征。第三种方法是在基于 Filter 的方法基础上，根据随机森林(Random Forest, RF)算法得出特征重要性得分并对特征进行排序，然后通过支持向量机(Support Vector Machine, SVM)分类器得出的 AUC 值选择特征。具体方法为将重要性从大到小排序后特征依次放入当前模型中，若 AUC 值提高则特征被选进最优特征集，否则被排除最有特征集之外。

3. 实证分析

3.1. 特征选择结果

在这三种特征选择方法的基础上，采用基于 AUC 性能指标值的支持向量机分类器的 5 次交叉验证，得到了基于 DNAdset 的最优特征选择方法。详细结果见表 1。

Table 1. The AUC of three feature selection methods based on 5-fold cross-validation for SVM on DNAdset
表 1. 三种特征选择方法在数据集 DNAdset 上基于 5 折交叉验证和支持向量机的 AUC 结果

Feature selection methods	Machine learning Classifier	AUC
LASSO	SVM ($C = 6, \gamma = -5$)	0.9523
Filter-based (mrmr)	SVM ($C = 9, \gamma = -8$)	0.9384
Wrapper-based (RF-SVM)	SVM ($C = 10, \gamma = -9$)	0.9546

在表 1 中，结果表明，与其他两种基于支持向量机(SVM)分类器的 5 次交叉验证的特征选择方法相比，Wrapper-based (RF-SVM)特征选择方法具有更高的 AUC 值。在独立数据集(DNAiset 和 DNArset)上，采用支持向量机的最优参数和阈值 0.49 (基于 AUC 值)来测试方法的性能。此外，利用 Wrapper-based (RF-SVM)算法对选取的 33 个特征，在阈值 0.44 范围内得到 ACC、MCC、精密度、灵敏度、特异度和 F-测度值。具体评价结果见表 2。此外，DNAdset、DNAiset 和 DNArset 的 ROC 曲线如图 3 所示。

Table 2. Performance of prediction model on DNAdset, DNAiset and DNArset
表 2. 预测模型在 DNAdset、DNAiset 和 DNArset 上的性能表现

Datasets	threshold value	AUC	ACC	MCC	F-measure	Sensitivity	Specificity	Precision
DNAdset	0.44	0.9546	0.8896	0.7819	0.8940	0.9307	0.8485	0.8600
DNAiset	0.44	0.8698	0.8277	0.6169	0.7463	0.7732	0.8543	0.7212
DNArset	0.44	0.7383	0.6055	0.1776	0.1923	0.7732	0.5947	0.1098

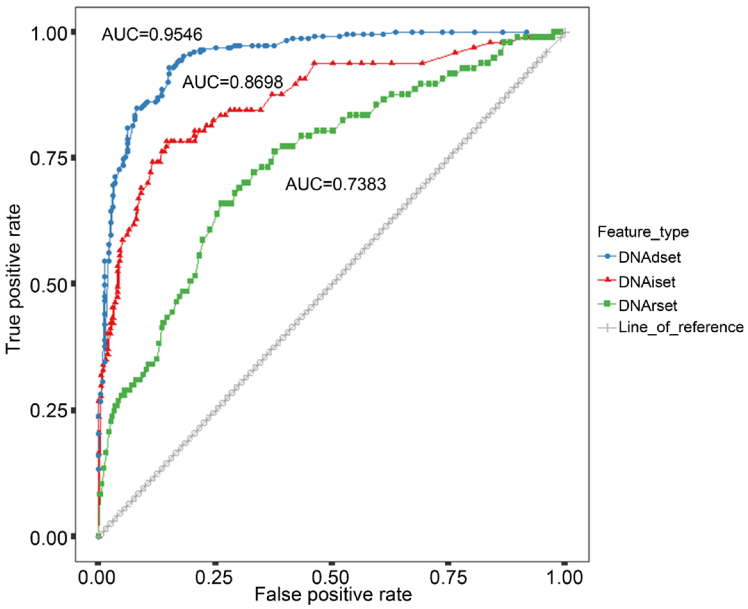

Figure 3. The ROC curves of DNAdset, DNAiset and DNArset
图 3. 在数据集 DNAdset、DNAiset 和 DNArset 上的 ROC 曲线

3.2. 两类特征预测性能分析

为了研究两类特征对 DNA 结合蛋白和非 DNA 结合蛋白的区分能力是否具有统计学意义，本文采用双侧 *t* 检验，显著性水平为 0.05，33 个特征中有 24 个(24/33 = 72.7%)具有显著性。因此，所选特征对区分 DNA 结合蛋白和非 DNA 结合蛋白具有统计学意义。

在选取的 33 个特征中，包含两种特征类型。其中属于氨基酸的物理化学性质的特征较多，为 19 个；属于氨基酸组成的较少，为 14 个。结果表明，融合后的 33 个特征会提高对 DNA 结合蛋白的预测性能(如图 4 所示)，两类特征均对 DNA 结合蛋白的预测有重要影响。具体评价结果见表 3。

在 AAindex 数据库(Tomii 和 Kanehisa，1996)中，理化性质指标基于最小生成树方法可以分为六类：Alpha and turn properties (A)、Beta propensity (B)、Composition (C)、Hydrophobicity (H)、Physicochemical properties (P)和 Other properties (O)。将所选择的 19 个基于氨基酸物理化学性质的特征与六类属性对比表明(如表 4 所示)，蛋白质序列中氨基酸残基的 Hydrophobicity (H)、Physicochemical properties (P)和 Alpha and turn properties (A)，是区分 DNA 结合蛋白和非 DNA 结合蛋白的主要原因。

3.3. 与其他方法对比分析

为了验证 grDNA-Prot 方法的有效性，将 grDNA-Prot 方法与现有的三个 web 服务器(DNAbinder、iDNA-Prot 和 DNA-Prot)在两个独立测试数据集 DNAiset 和 DNArset 上进行了性能比较。详细结果见表 5。对于 DNAiset，grDNA-Prot 对 ACC、AUC、MCC 和 F-measure 的性能评价高于 DNAbinder 和 DNA Prot。

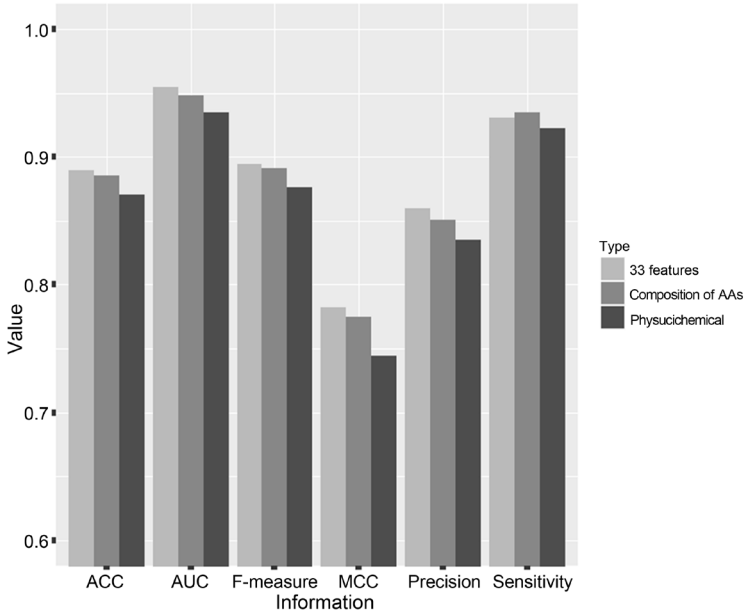

Figure 4. Comparison of the three feature types in DNAdset
图 4. 在 DNAdset 中三种类型特征的比较

Table 3. The performance of the three feature types in DNAdset
表 3. 在 DNAdset 中三种类型特征的性能

Feature information	AUC	ACC	MCC	Sensitivity	Precision	F-measure
Composition of AAs	0.9477	0.8853	0.7744	0.9351	0.8504	0.8907
Physicochemical	0.9351	0.8701	0.7443	0.9221	0.8353	0.8765
33 features	0.9546	0.8896	0.7819	0.9307	0.8600	0.8940

Table 4. The distribution of 19 features based on physicochemical properties of amino acids in six groups
表 4. 19 个基于氨基酸物理化学性质的特征在六类属性中分布情况

	A	B	C	H	P	O
BUNA790101	1	0	0	0	0	0
CHAM820102	0	0	0	1	0	0
EISD860102	0	0	0	1	0	0
FASG760103	1	0	0	0	0	0
FAUJ880103	0	0	0	0	1	0
FUKS010104	0	0	1	0	0	0
HOPT810101	0	0	0	1	0	0
LEVM760101	0	0	0	1	0	0
MAXF760104	0	0	0	0	0	1
MEEJ800101	0	0	0	1	0	0
MITS020101	0	0	0	0	1	0
MIYS990101	0	0	0	1	0	0
OOBM770105	0	0	0	0	1	0
RACS770103	0	0	0	1	0	0

Continued

RADA880103	0	0	0	0	1	0
ROSG850101	0	0	0	0	1	0
TANS770101	1	0	0	0	0	0
WOLR790101	0	0	0	1	0	0
ZIMJ680104	0	0	0	1	0	0
TOTAL	3	0	1	9	5	1

Table 5. Comparison of the predicted results by grDNA-Prot and other methods on DNAiset and DNArset
表 5. grDNA-Prot 在独立测试集 DNAiset 和 DNArset 上与其他方法的预测结果比较

Datasets	Methods	ACC	AUC	MCC	Sensitivity	Specificity	Precision	F-measure
DNAiset	DNAbinder	0.709	0.809	0.459	0.845	0.643	0.536	0.656
	iDNA-Prot	**0.889**	**-**	**0.752**	**0.659**	**1.000**	**1.000**	**0.795**
	DNA-Prot	0.824	0.732	0.589	0.526	0.969	0.894	0.662
	grDNA-Prot	**0.828**	**0.870**	**0.617**	**0.773**	**0.854**	**0.721**	**0.746**
DNArset	DNAbinder	0.3845	-	0.1007	-	-	-	0.143
	iDNA-Prot	0.614	-	0.132	-	-	-	0.172
	DNA-Prot	**0.735**	-	**0.152**	-	-	-	**0.197**
	grDNA-Prot	**0.606**	**0.738**	**0.178**	-	-	-	**0.192**

此外，在 DNArset 的真实环境中，grDNA-Prot 比 DNAbinder、iDNA-Prot、DNA-Port 获得更高的 MCC，比 DNAbinder、iDNA-Prot 获得更高的 F-measure。本文提出方法的性能接近 iDNA-Port 和 DNA-Prot，但在两个数据集上的综合效果最优。这些结果表明 grDNA-Prot 方法可以有效地鉴定 DNA 结合蛋白。

4. 结论

DNA 结合蛋白在细胞内外各种生命活动中起着重要的作用，现今已经研究出多种预测 DNA 结合蛋白的计算方法。本文提出的方法包含 20 维氨基酸组成频率特征和 531 维基于柱形图表示法的氨基酸物理化学性质特征，使用基于 Wrapper 的方法对融合后特征进行特征选择，选择出包含这两种类型的 33 维特征，最后建立了基于支持向量机的预测模型。同时发现，Hydrophobicity (H)、Physicochemical properties (P) 和 Alpha and turn properties (A)是区分 DNA 结合蛋白和非 DNA 结合蛋白的主要理化性质。因此，研究结果表明所选取的特征可以更好地解释绑定机制。

此外，通过在两个独立测试数据集(DNAiset 和 DNArset)上与其他方法(DNA-prot、iDNA-prot 和 DNAbinder)的比较，证明了 grDNA-Prot 方法的有效性。因此，grDNA-Prot 可以相对准确地预测 DNA 结合蛋白。

基金项目

本文得到了河北省自然科学基金项目(F2019402078)、河北省高等学校科学技术研究项目(QN2018235)和河北省研究生创新资助项目(CXZZSS2021092)的支持，在此表示感谢。

参考文献

[1] Lilley, D.M.J (1995) DNA Protein Structural Interactions. Oxford University Press, Oxford.

[2] Zimmer, C. and Wähnert, U. (1986) Nonintercalating DNA-Binding Ligands: Specificity of the Interaction and Their

Use as Tools in Biophysical, Biochemical and Biological Investigations of the Genetic Material. *Progress in Biophysics and Molecular Biology*, **47**, 31-112. https://doi.org/10.1016/0079-6107(86)90005-2

[3] Boute, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot. In: Edwards, D., Ed., *Plant Bioinformatics*, Vol. 406, Humana Press, Totowa, 89-112. https://doi.org/10.1007/978-1-59745-535-0_4

[4] Helwa, R. and Hoheisel, J.D. (2010) Analysis of DNA-Protein Interactions: From Nitrocellulose Filter Binding Assays to Microarray Studies. *Analytical and Bioanalytical Chemistry*, **398**, 2551-2561. https://doi.org/10.1007/s00216-010-4096-7

[5] Freeman, K., Gwadz, M. and Shore, D. (1995) Molecular and Genetic Analysis of the Toxic Effect of Rap1 Overexpression in Yeast. *Genetic*, **141**, 1253-1262. https://doi.org/10.1093/genetics/141.4.1253

[6] Jaiswal, R., Singh, S.K., Bastia, D. and Escalante, C.R. (2015) Crystallization and Preliminary X-Ray Characterization of the Eukaryotic Replication Terminator Reb1-Ter DNA Complex. *Acta Crystallographica Section F*: *Structural Biology Communications*, **71**, 414-418. https://doi.org/10.1107/S2053230X15004112

[7] Buck, M.J. and Lieb, J.D. (2004) Chip-Chip: Considerations for the Design, Analysis, and Application of Genome-Wide Chromatin Immunoprecipitation Experiments. *Genomics*, **83**, 349-360. https://doi.org/10.1016/j.ygeno.2003.11.004

[8] Langlois, R.E. and Lu, H. (2010) Boosting the Prediction and Understanding of DNA-Binding Domains from Sequence. *Nucleic Acids Research*, **38**, 3149-3158. https://doi.org/10.1093/nar/gkq061

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

[10] 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

[11] Lin, W.Z., Fang, J.A., Xiao, X.K. and Chou, K.C. (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE*, **6**, e24756. https://doi.org/10.1371/journal.pone.0024756

[12] Kumar, K.K., Pugalenthi, G. and Suganthan, P.N. (2009) DNA-Prot: Identification of DNA Binding Proteins from Protein Sequence Information Using Random Forest. *Journal of Biomolecular Structure and Dynamics*, **26**, 679-686. https://doi.org/10.1080/07391102.2009.10507281

[13] Kumar, M., Gromiha, M.M. and Raghava, G.P. (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

[14] 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

[15] Zhang, J. and Liu, B. (2017) PSFM-DBT: Identifying DNA-Binding Proteins by Combing Position Specific Frequency Matrix and Distance-Bigram Transformation. *International Journal of Molecular Sciences*, **18**, Article No. 1856. https://doi.org/10.3390/ijms18091856

[16] Zhang, J., Chen, Q.C. and Liu, B. (2019) DeepDRBP-2L: A New Genome Annotation Predictor for Identifying DNA Binding Proteins and RNA Binding Proteins Using Convolutional Neural Network and Long Short-Term Memory. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1. https://doi.org/10.1109/TCBB.2019.2952338

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

[18] Wei, L.Y., Tang, J.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

[19] Huang, T., Chen, L., Cai, Y.D. and Chou, K.C. (2011) Classification and Analysis of Regulatory Pathways Using Graph Property, Biochemical and Physicochemical Property, and Functional Property. *PLoS ONE*, **6**, e25297. https://doi.org/10.1371/journal.pone.0025297

[20] Zou, C., Gong, J. and Li, H. (2013) An Improved Sequence Based Prediction Protocol for DNA-Binding Proteins Using SVM and Comprehensive Feature Analysis. *BMC Bioinformatics*, **14**, Article No. 90. https://doi.org/10.1186/1471-2105-14-90

[21] Li, S., Li, D.P., Zeng, X.X., Wu, Y.F., Guo, L. and Zou, Q. (2014) nDNA-Prot: Identification of DNA-Binding Proteins Based on Unbalanced Classification. *BMC Bioinformatics*, **15**, Article No. 298. https://doi.org/10.1186/1471-2105-15-298

[22] Kumar, R., Srivastava, A., Kumari, B. and Kumar M. (2015) Prediction of Beta-Lactamase and Its Class by Chou's Pseudo-Amino Acid Composition and Support Vector Machine. *Journal of Theoretical Biology*, **365**, 96-103. https://doi.org/10.1016/j.jtbi.2014.10.008

[23] Shahana, Y.C., Swakkhar, S. and Abdollah, D. (2017) iDNAProt-ES: Identification of DNA-Binding Proteins Using Evolutionary and Structural Features. *Scientific Reports*, **7**, Article No. 14938. https://doi.org/10.1038/s41598-017-14945-1

[24] Hu, J., Zhou, X.G., Zhu, Y.H., Yu, D.J. and Zhang, G.J. (2020) TargetDBP: Accurate DNA-Binding Protein Prediction via Sequence-Based Multi-View Feature Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **17**, 1419-1429.

[25] Wang, Y.B., Ding, Y.J., Guo, F., Wei, L.Y. and Tang, J.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

[26] Liu, X.J., Gong, X.J., Yu, H. and Xu, J.H. (2018) A Model Stacking Framework for Identifying DNA Binding Proteins by Orchestrating Multi-View Features and Classifiers. *Genes*, **9**, Article No. 394. https://doi.org/10.3390/genes9080394

[27] Ahmad, S., Gromiha, M.M. and Sarai, A. (2004) Analysis and Prediction of DNA-Binding Proteins and Their Binding Residues Based on Composition, Sequence and Structural Information. *Bioinformatics*, **20**, 477-486. https://doi.org/10.1093/bioinformatics/btg432

[28] Liu, B., Fang, L.Y., Wang, S.Y., Wang, X.L., Li, H.T. and Chou K.C. (2015) Identification of MicroRNA Precursor with the Degenerate K-Tuple or Kmer Strategy. *Journal of Theoretical Biology*, **385**, 153-159. https://doi.org/10.1016/j.jtbi.2015.08.025

[29] Kawashima, S., Pokarowski, P., Pokarowska, M., Mkolinski, A., Katayama, T. and Kanehisa, M. (2008) AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Research*, **36**, D202-D205. https://doi.org/10.1093/nar/gkm998

[30] Huang, H.L., Lin, I.C., Liou, Y.F., Tsai, C.T., Hsu, K.T., Huang, W.L., Ho, J. and Ho, S.Y. (2011) Predicting and Analyzing DNA-Binding Domains Using a Systematic Approach to Identifying a Set of Informative Physicochemical and Biochemical Properties. *BMC Bioinformatics*, **12**, Article No. S47. https://doi.org/10.1186/1471-2105-12-S1-S47

[31] Tung, C.W. and Ho, S.Y. (2008) Computational Identification of Ubiquitylation Sites from Protein Sequences. *BMC Bioinformatics*, **9**, Article No. 310. https://doi.org/10.1186/1471-2105-9-310

[32] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **73**, 273-282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

[33] Fang, Y., Guo, Y., Feng, Y. and Li, M. (2008) Predicting DNA-Binding Proteins: Approached from Chou's Pseudo Amino Acid Composition and Other Specific Sequence Features. *Amino Acids*, **24**, 103-109. https://doi.org/10.1007/s00726-007-0568-2

[34] Huang, Y., Niu, B.F., Gao, Y., Fu, L. and Li, W.Z. (2010) CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics*, **26**, 680-682. https://doi.org/10.1093/bioinformatics/btq003

[35] Randic, M., Zupan, J., Balaban, A.T., Vikic-Topic, D. and Plavšić, D. (2011) Graphical Representation of Proteins. *Chemical Reviews*, **111**, 790-862. https://doi.org/10.1021/cr800198j

[36] Yu, J.F., Dou, X.H., Wang, H.B., Sun, X., Zhao, H.Y. and Wang, J.H. (2015) A Novel Cylindrical Representation for Characterizing Intrinsic Properties of Protein Sequences. *Journal of Chemical Information and Modeling*, **55**, 1261-1270. https://doi.org/10.1021/ci500577m

[37] Zhang, Y.N., Yu, D.J., Li, S.S., Fan, Y.X., Huang, Y. and Shen, H.B. (2012) Prediction Protein-ATP Binding Sites from Primary Sequence through Fusing Bi-Profile Sampling of Multi-View Features. *BMC Bioinformatics*, **13**, Article No. 118. https://doi.org/10.1186/1471-2105-13-118

[38] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. and Nielsen, H. (2000) Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*, **16**, 412-424. https://doi.org/10.1093/bioinformatics/16.5.412

[39] Sonego, P., Kocsor, A. and Pongor, S. (2008) ROC Analysis: Applications to the Classification of Biological Sequences and 3D Structures. *Briefings in Bioinformatics*, **9**, 198-209. https://doi.org/10.1093/bib/bbm064

[40] Deng, L., Pan, J., Xu, X., Yang, W., Liu, C. and Liu, H. (2018) PDRLGB: Precise DNA-Binding Residue Prediction Using a Light Gradient Boosting Machine. *BMC Bioinformatics*, **19**, Article No. 522. https://doi.org/10.1186/s12859-018-2527-1

[41] Peng, H., Long, F.H. and Ding, C. (2015) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159