DNA 结合蛋白预测-最终报告

贠龙飞，2018012381
清华大学软件学院
机器学习, 2021 年秋季学期

摘　要

DNA 结合蛋白 (DNABPs) 的鉴定是生物医学研究中的一个重要问题，因为 DNABPs 在各种细胞过程中起着至关重要的作用。目前，机器学习在生物医学等交叉方向上取得了 SOTA, 进一步提高这些方法性能的一个关键步骤是找到合适的蛋白质表示，结合生物、化学专业知识进行更好的特征工程。本文使用 PseACC 对氨基酸残基序列进行编码，使用 XGboost 对特征进行重要性排序，然后将降维后的特征作为 RBF kernel-SVM 和随机森林的输入。为了降低降维造成的信息丢失，借鉴了随机森林的思想，训练多个 SVM 做最大投票。降维显著降低了 PseACC 的冗余性，比传统方法具有更好的性能，并且对学到的结果进行分析证明了 XGBoost 进行降维用生化知识分析是可靠的。本文提出的方法在 PDB186 测试集上准确率最高达到了 85.41%，超过了调研文献中的方法。

1　介绍

1.1　选题背景及意义

DNA 结合蛋白 (DNA-binding protein, DNABPs) 是与单链或双链 DNA 结合的蛋白质。如果结合具有序列特异性，通常发生在 B-DNA 的大沟 (major groove)。DNABPs 可以包含锌指结构域、α 螺旋-β 转角-α 螺旋结构域和亮氨酸拉链结构域以促进与核酸的结合。我们知道，蛋白质的一级结构决定高级结构，所以仅根据氨基酸序列来预测 DNA 结合蛋白是完全可行的。

DNABPs 可以与 DNA 相互作用，并在各种细胞过程中如转录调节、基因重组、基因重排、复制、修复和 DNA 修饰中发挥关键作用。

传统对于 DNA 结合蛋白的识别主要是通过实验获得的，包括层析、免疫共沉淀、X 射线晶体衍射等。这些方法不仅需要烦杂的实验步骤、昂贵的实验仪器、同时利用蛋白质的三级结构和一级结构而且最重要的是耗费大量的时间才能获得一个蛋白质的结果。所以仅根据蛋白质的一级结构预测该蛋白是否是 DNA 结合蛋白是非常有必要的。

但是如何把变长的氨基酸残基序列编码成为定长的特征表示，同时尽量保证序列的组成、序

列信息编码到特征张量中尚未得到解决，本文参考了 PseACC 的方法，并对其进行降维，提升了算法的性能。

2 相关工作

2.1 文献调研 (详细)

iDNA-Prot|dis [3] 不同排列的蛋白质序列非常多，而且它们的长度高度可变。如何把生物大分子的序列形式化表示同时保留序列的次序信息，是现今计算生物学面临的一个极具挑战性的问题。我们上课学过的分类方法如 KNN、SVM 等方法都需要样本的特征维度相同，所以不能直接迁移使用。Chou et al.[1] 提出了伪氨基酸组成方法 (Chou's PseAAC), 自从这种方法提出后，广泛活跃在计算生物学的各个领域如肽段的超二级结构预测 [7] 和抗癌肽段预测 [3] 等。

这篇文献的贡献可以归纳如下:

1. 通过理化、结构性质对 150 种合并氨基酸成氨基酸簇的方法进行实验，证明了将氨基酸合并确实既减轻了计算量同时也能提升预测性能。
2. 和四种已有方法进行对比，预测的准确率达到 77%, 优于其它方法。
3. 在一个独立的数据集上进行测试，同样也证明准确率优于其它方法。
4. 搭建了一个 Web Server，可以进行在线预测。

iDNA-Prot|dis[3] 使用的方法是使用高斯核的支持向量机。

DNA-Prot DNA-Prot 采用的方法是随机森林算法。它将每段氨基酸序列用 116 个特征进行描述，包括以下几个方面:

1. 将二十氨基酸按照侧链的类型分成十种，分别计算每类的频率。
2. 计算疏水性、亲水性和中性氨基酸出现的频率，并统计了 27 种三肽的频率。
3. 使用长度为 10 的肽段滑动窗口在每个序列上滑动，定义出现超过 6 个疏水性氨基酸为富疏水性肽段 (hydrophobic rich peptides)，统计它们出现的频率。并类似计算富亲水性氨基酸肽段、富中性氨基酸肽段的频率。
4. 使用 PSIPRED 仅从序列预测超二级结构作为分类特征。
5. 从 UMBCAAIndex Database 提取了 14 个理化性质作为分类特征。

然后作者使用了 correlation-based feature subset selection method(CFSS) 做特征选择。CFSS 方法已成功地应用于各种生物信息学问题，以降低特征维数，提高预测精度。CFSS 方法不是对单个特征进行评分和排名，而是对特征子集的价值进行评分和排名。

这篇文章的贡献可以归纳如下:

1. DNA-Prot 对所有特征的准确率为 84.37%，对 20 个特征的准确率为 79.17%, 这说明 CFSS 进行特征选择是有效的。作者也对留下来的特征进行了分析。比如 DNA 带有负电而正电氨基酸出现的频率被保留，甘氨酸作为侧链最小的氨基酸可以为蛋白质的高级结构提供柔韧性所以甘氨酸的频率也被保存下来。

2. 在一个大型的独立数据集上进行测试准确率达到了 81.83%, 证明的模型的有效性。
3. DNA-Prot 也和 SVM 的 counterpart 进行了对比，准确率大概提升了十个点。

2.2 其它论文调研

Kumar et al.[4] 采用的编码方式是单个氨基酸和二肽出现的概率，使用 SVM 对含有部分 DNA 结合功能的蛋白质和全序列 DNA 结合功能的蛋白质分别进行训练和测试，得出结论是训练结果不具有迁移性。Lin et al.[5] 提出了 iDNA-Prot 方法，将伪氨基酸组成方法和 Grey Model 相结合后使用随机森林进行预测。[2] 提出了一种新的特征定长编码方法 Auto-Cross Covariance transformation (ACC)，然后他们将这种方法和支持向量机结合区分 DNABPs。

2.3 与已有方法的差别

我使用了 PseAAC 来从蛋白质的氨基酸序列中提取特征，但是考虑到用这种方法提取得到的特征维数相对较大（1220 维，d=3），这甚至超过了数据集的大小。因为 XGBoost 加入了正则项可以防止过拟合，所以我首先把原来的训练集划分成训练集和验证集。使用 XGBoost 去拟合训练集，对特征的重要性进行排序，通过比较在验证集上的准确率筛选出比较重要的特征以实现降维和特征选择的目的。

但是仅使用 PseAAC 并不是特别充分，因为它的主要目的还是将变长的氨基酸序列表示成定长的特征，如果要实现更好的准确率和泛化能力必然要求充分考虑到化学生物的先验知识。我对氨基酸残基进一步细分然后对序列信息进行加权。比如 DNA 的核酸通常带有负电，所以对于带有正电荷的三种氨基酸残基或者小的侧链的脂肪族氨基酸和含有羟基（容易形成氢键）的氨基酸可以适当增加它们的权重。相比于 DNA-Prot，这种方法并不需要太多的特征工程 (DNA-Prot 手工制造了 116 个特征) 而且对特征的选择也更容易获得，可以达到更高的性能和准确率。

最后要注意的一点是 PseAAC 并不是一个校准模型，随着 d 的增大必然会造成两个或多个氨基酸共同出现的概率会降低，而某一个短肽段恰恰可能是一个蛋白质中二级结构或者超二级结构的片段。也就是说，在某种情况下短的肽段含有的信息应当对于结合生物大分子具有更重要的作用，但是使用 PseAAC 方法会固有的遮蔽这种信息。所以我把对应的特征乘以指数的 d 次方。

3 方法

3.1 预处理

读取数据集中每个蛋白质的氨基酸序列，去除蛋白质名字、空白符等无用信息，去除未确定的氨基酸和不属于二十种标准氨基酸的衍生氨基酸，按照是否结合 DNA 归到正例和负例。

3.2 PseAAC 提取定长特征

假设有一个氨基酸残基序列为 $P = R_1R_2R_3...R_L$, PseACC 将其转化成 $P = [\psi_1\psi_2...\psi_\Omega]^T$, 长度 Ω 和组分 Ψ_u 依赖于从蛋白质序列中提取信息的方法。

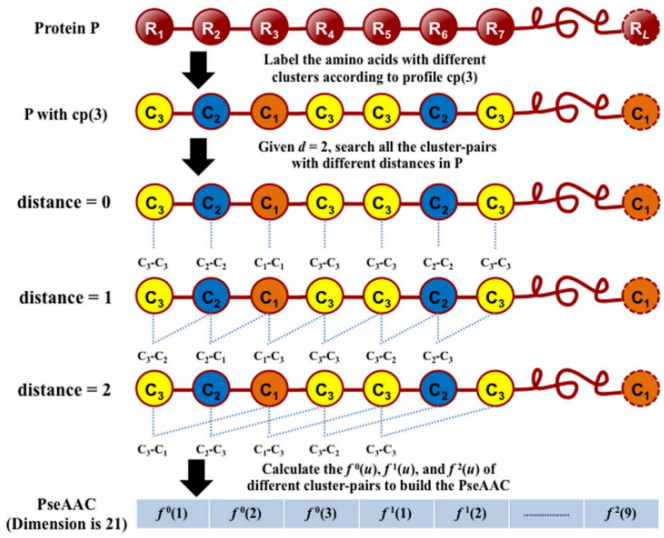

图 1: Illustration of PseACC.The characters C1, C2, and C3 represent the three different clusters and are coloured with orange, blue, and yellow, respectively. When the maximum pairwise distance d = 2.

定义氨基酸残基对之间的出现频率为

$$f(R_i, R_j|d) \tag{1}$$

R_i 和 R_j 可以是二十种基本氨基酸的任何一种，d 是 R_i 和 R_j 之间的氨基酸残基数。当 d=0 时，有

$$f(R_i, R_j|0) = f^0(u_0), (1 \le u_0 \le 20) \tag{2}$$

当 d=1 时，有

$$f(R_i, R_j|1) = f^1(u_1), (1 \le u_1 \le 400) \tag{3}$$

所以 Ψ_u 可以被定义为一个 $\Omega = 20 + 400d$ 的向量:

$$\Psi_u = \begin{cases} f^0(u), & 1 \le u \le 20 \\ f^1(u), & 21 \le u \le 420 \\ f^2(u), & 421 \le u \le 820 \\ \quad ... & \\ f^d(u), & 21 + 400(d-1) \le u \le 20 + 400d \end{cases} \tag{4}$$

但是当 d 较大比如 d=100 时，维数 Ω 也将变得很大，这将导致严重的维数灾问题。Liu et al.[3] 提出了一种压缩特征维数的方法。

$$cp(20) = \{A; C; D; E; F; G; H; I; K; L; M; N; P; Q; R; S; T; V; W; Y\} \tag{5}$$

是二十种标准氨基酸的单字母表示。他们在对大量的 DNABPs 的氨基酸序列进行分析后发现可以把氨基酸分成氨基酸簇，这些氨基酸同时出现时蛋白质是 DNA 结合蛋白的可能性较高。其中一种氨基酸簇分类方法如下:

$$cp(13) = \{MF; IL; V; A; C; WYQHP; G; T; S; N; RK; D; E\} \tag{6}$$

本文使用 XGBoost 对获得的向量表示进行排序后挑选出重要性较高且对验证集的准确率影响小于一定阈值的特征，成功将 d=3 时的 1220 维特征降到 290 维左右。

但是降维后效果反而可能不如单用 XGBoost 进行分类预测因为考虑到有一些信息丢失了，所以我对二十种氨基酸侧链的性质进行分析，并参考了 DNA-Prot 的富疏水性、富亲水性、富中性氨基酸的概念加入了新的特征，并赋予它们较大的权重以便提高分类准确率。

我设置了一个滑动窗口，分别对每个蛋白质的一级结构用一个长度为十个氨基酸残基的滑动窗口扫描:

- 当疏水性氨基酸 (Gly, Ala, Val, Ile, Pro, Phe) 的数目在窗口中超过了 6 个，标记该蛋白质为富疏水性蛋白质 (hydrophobic rich protein) 或富非极性蛋白质。
- 当亲水性氨基酸 (Ser, Thr, Tyr, Met, Cys, Asn, Gln) 的数目在窗口中超过了 6 个，标记该蛋白质为富亲水性蛋白质 (hydrophilic rich protein) 或富极性蛋白质。
- 当带有负电的氨基酸 (Glu, Asp) 在窗口中超过了 4 个，标记该蛋白质为富负电蛋白质或富酸性蛋白质。
- 当带有正电的氨基酸 (Lys, Arg, His) 的数目在窗口中超过了 4 个，标记该蛋白质为富正电蛋白质或者富碱性蛋白质。

3.3 模型选择

表 Tab. 2是 XGBoost 对于拟合的训练集给出二十种标准氨基酸的特征重要性。从中我们可以看出:

- 赖氨酸 (Lys/L)、精氨酸 (Arg/A) 等带正电的氨基酸重要性较高，组氨酸因为咪唑基的存在重要性相对较低。因为核酸往往带有负电荷，所以带有正电的氨基酸有利于蛋白质与 DNA 的结合。
- 部分含有羟基的氨基酸如苏氨酸 (Thr/T) 的重要性也较高，因为羟基有利于和碱基、脱氧核糖之间形成氢键从而利于与 DNA 结合。
- 甘氨酸 (Gly/G) 因为具有较小的侧链基团 (只有一个氢原子) 也常常出现在高级构象的卷曲折叠中从而有利于降低能量促进折叠，所以重要性也较高。

从上述分析中可以得知，使用 XGboost 进行特征提取确实合理且学到了有用的信息。但是同时我们也注意到，有大量的特征的重要性是 0(692 个)，我首先将这部分特征去除掉。然后将剩余的特征分成五十个为一组把并把最高的重要性作为剩余特征筛选的阈值，并通过在验证集上测量去除每五十个特征后的准确率变化来进一步对剩下的近 600 个特征进行降维。

接下来我对比使用了 kernel SVM 和 Random Forest 对降维后的特征进一步进行二分类的效果。

考虑到使用 XGBoost 降维也有可能会丢失掉一部分信息，所以我借鉴了 random forest 的思路，用五个 SVM 设置不同的超参数，其中使用网格搜索调好的参数的 SVM 的权重设为 1.5，然后做加权最大投票，大于 0 的 label 置为正例，小于 0 的 label 置为负例。下面用消融实验证明最大投票的有效性：我随机进行了五次训练，然后预测得到的准确率如 Tab. 1所示，使用单 SVM 预测的平均准确率是 81.18%, 使用多 SVM 预测得到的平均准确率达到了 83.33%，提升了两个点，

index	single SVM	multiple SVM
1	81.72%	82.26%
2	81.72%	84.95%
3	80.11%	84.95%
4	82.80%	83.33%
5	79.57%	81.18%

表 1: The above ablation study shows the effectievness of using multiple SVM to conduct a majority voting.

hydrophilic amino acids								
Phe	**Tyr**	**Trp**	**Ser**	**Thr**	**Cys**	**Met**	**Asn**	**Gln**
0.00023744	0.00136127	0.00032807	0.00162829	0.00401071	0.00098587	0.00204659	0.00141568	0.0013664
hydrophobic amino acids						**Alkaline**		
Gly	**Ala**	**Val**	**Leu**	**Ile**	**Pro**	**Lys**	**Arg**	**His**
0.0028982	0.00098825	0.00275268	0.00098952	0.00144329	0.00237099	0.00435894	0.00416522	0.00134643

表 2: The importance of single amino acids obtained by using XGboost.

可以弥补失去一部分特征造成的损失。

4 实验

4.1 数据集选择

我使用的数据集是 PDB186 和 PDB1075[6]。PDB 是目前最主要的收集生物大分子 (e.g. 蛋白质、核酸和糖) 的结构的数据库，是通过 X 射线单晶衍射、核磁共振、电子衍射等实验手段确定的三维结构数据库。这两个数据集是从 PDB 中采样得到的包括 DNA 结合蛋白和非 DNA 结合蛋白两类，同时具有以下优点：

1. 氨基酸残基数小于 50 的蛋白质被手动滤除，因为他们可能是某个蛋白质的碎片。
2. 由于实验手段的缺陷含有未知氨基酸的蛋白质被去除。
3. 在一类 (e.g. 正例或负例) 中的两个蛋白质的相似度小于 25% 以获得足够的多样性。

4.2 特征选择

首先要确定 PseAAC 的 d 的大小，我通过比对 XGBoost 得到的准确率，最后选择 $d = 3$ 作为 PseACC 的最大氨基酸对的距离。

表 Tab. 3所示的是设定不同的重要性阈值后得到的特征维数和验证集上准确率的变化。我选择了 0.00022 作为最后的阈值对 PseAAC 得到的特征进行裁剪，既保证了特征维数不至于太多同时也保证了不会出现欠拟合的问题。

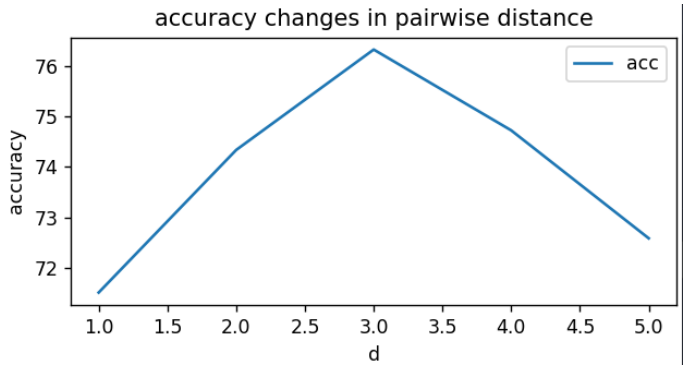

图 2: The overall Acc values achieved for cp(20) with different d values

Threshold	feature_number	Accuracy
1.17e-6	535	75.80%
0.00011	485	77.42%
0.00022	435	76.88%
0.00035	385	73.66%
0.054	335	75.268%
0.00077	285	76.34%

表 3: Threshold of feature importance and Accuracy changes

4.3 实验结果

使用的评价指标如下所示:

$$\begin{cases} Sn = & \dfrac{TP}{TP+FN} \\ Sp = & \dfrac{TN}{TN+FP} \\ Acc = & \dfrac{TP+TN}{TP+TN+FP+FN} \\ MCC = & \dfrac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \tag{7}$$

和其它方法的比较见表 Tab. 4, 数据来源于 [3], 选择 PDB186 作为测试集。可以看到，我们的方法从各个指标上都超过了调研文献的方法。

ROC 曲线有个很好的特性：当测试集中的正负样本的分布变换的时候，ROC 曲线能够保持不变。在实际的数据集中经常会出现样本类不平衡，即正负样本比例差距较大，而且测试数据中的正负样本也可能随着时间变化。ROC 图见图 Fig. 1,AUC 面积达到了 0.88，超过了之前的方法 (iDNA-Prot|dis 0.834, DNAbinder 0.814 etc.)。AUC 表示随机给定一正一负两个样本，将正样本排在负样本前面的概率，因此 AUC 越大，说明正样本越有可能被排在负样本之前，即分类额越好。

Methods	ACC(%)	MCC	Sn(%)	Sp(%)	AUC(%)
OURS	**84.41**	**0.724**	**68.8**	**64.50**	**88**
iDNA-Prot|dis	72.00	0.445	79.50	64.50	78.60
iDNA-Prot	67.20	0.344	67.70	66.70	N/A
DNA-prot	61.80	0.240	69.90	53.80	N/A
DNAbinder	60.80	0.216	57.00	64.50	60.70
DNABIND	67.70	0.355	66.70	68.80	69.40
DNA-Threader	59.70	0.279	23.70	95.70	N/A
DBPPred	76.90	0.538	79.60	74.20	79.10

表 4: Results. Ours is better.

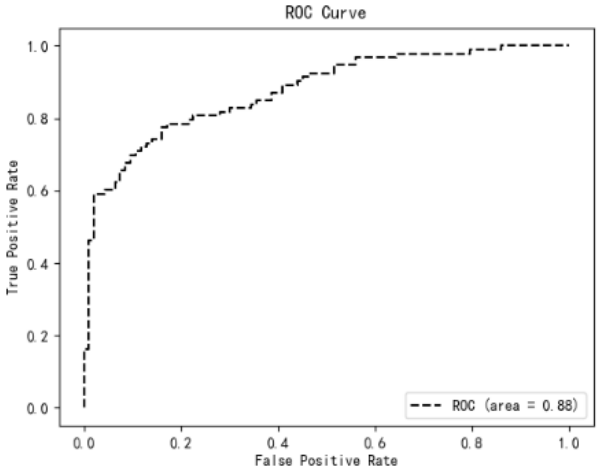

图 3: The ROC (receiver operating characteristic) curves obtained by different methods on the benchmark dataset. The areas under the ROC curves or AUC are 0.834, 0.826, 0.814, 0.815, 0.789 and 0.761 for iDNA-Prot|dis (cp(20)), iDNA-Prot|dis (cp(14)), DNAbinder (dimension 21), DNAbinder(dimension 400), DNA-Prot and iDNA-Prot, respectively.

5 结论

本文用 PseACC 对氨基酸序列进行编码，用 XGBoost 防止过拟合同时对特征进行降维。降维后的结果再加上手工加入的特征作为 SVM 的输入，最后为了防止降维导致信息丢失使用 5 个 SVM 做最大投票，从指标上全方位超过了调研的方法。

其实这只是一个简化的问题，后面可以进一步研究一下具体结合到 DNA 的是什么样的氨基酸序列，或者说蛋白质是整个具有 DNA 结合性还是只有部分结合性?

总的来说，做这个大作业的体验不错，既锻炼了自己的码力将课上学过的知识用到了实际的问题种，也温习了生物知识。我对用学过的机器学习解决交叉问题非常感兴趣，希望以后还有机会可以做交叉的研究。也希望以后能有机会多读文献，思考机器学习的发展。

参考文献

[1] Kuo-Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

[2] Qiwen Dong, Shanyi Wang, Kai Wang, Xuan Liu, and Bin Liu. Identification of dna-binding proteins by auto-cross covariance transformation. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 470–475. IEEE, 2015.

[3] Zohre Hajisharifi, Moien Piryaiee, Majid Mohammad Beigi, Mandana Behbahani, and Hassan Mohabatkar. Predicting anticancer peptides with chou s pseudo amino acid composition and investigating their mutagenicity via ames test. *Journal of Theoretical Biology*, 341:34–40, 2014.

[4] Manish Kumar, Michael M Gromiha, and Gajendra PS Raghava. Identification of dna-binding proteins using support vector machines and evolutionary profiles. *BMC bioinformatics*, 8(1):1–10, 2007.

[5] Wei-Zhong Lin, Jian-An Fang, Xuan Xiao, and Kuo-Chen Chou. idna-prot: identification of dna binding proteins using random forest with grey model. *PloS one*, 6(9):e24756, 2011.

[6] Bin Liu, Jinghao Xu, Xun Lan, Ruifeng Xu, Jiyun Zhou, Xiaolong Wang, and Kuo-Chen Chou. 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(9):e106691, 2014.

[7] Dongsheng Zou, Zhongshi He, Jingyuan He, and Yuxian Xia. Supersecondary structure prediction using chou's pseudo amino acid composition. *Journal of Computational Chemistry*, 32(2):271–278, 2011.