DNA 结合蛋白预测-选题报告

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

1 问题描述

1.1 选题动机

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。

1.2 数据集选择

我计划使用的数据集是 PDB186 和 PDB1075[7]。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。

2 文献调研

文献调研这一部分，我重点调研了 [3] 和 [4] 两篇文章，将会在 Sec. 2.1详细介绍。Sec. 2.2将会对其他文献进行简要介绍。

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 [8] 和抗癌肽段预测 [3] 等。

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。

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:

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

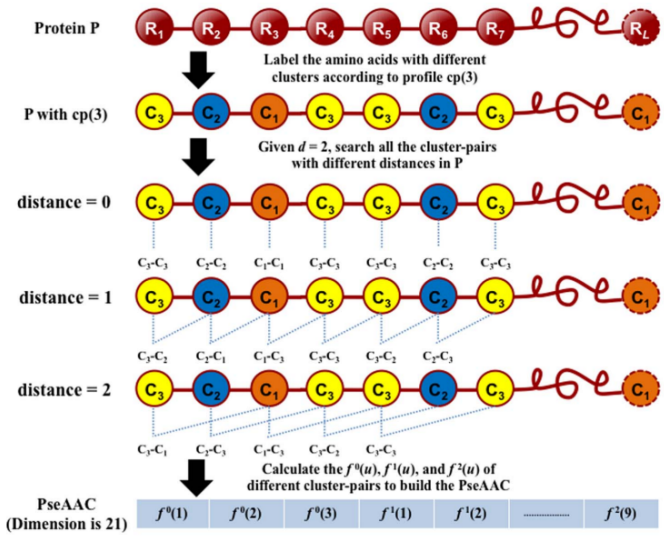

图 1: Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

iDNA-Prot|dis[3] 使用的方法是使用高斯核的支持向量机。使用的评价指标如下所示:

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

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。

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。

DNA-Prot 使用的评价指标和上一篇文献一样。

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.[5] 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.[6] 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。

3 初步拟定的解决方法

3.1 特征工程的改进

首先学习使用 PseAAC 表示氨基酸残基序列，我觉得这种特征工程方法并没有充分利用到化学生物的先验知识。而 PseAAC 只考虑到了氨基酸之间的距离信息，所以我们可以对氨基酸残基进一步细分对序列信息进行加权，比如对脂肪族氨基酸或者含羟基的氨基酸赋予更高的权重。

DNA-Prot 强调了理化性质的重要性，但是对氨基酸的序列信息包含的不够细致。我想把两种方法结合一下，可以考察第二个方法最后保留的重要特征，对第一个方法进行加权。

另外，也需要调研一下其它方法，比如把短序列的相似性作为一个特征等。

3.2 分类方法的改进

由于数据集的规模一般不是太大，XGBoost 加入了正则项可以防止过拟合，所以我考虑使用 XGBoost 对特征重要性排序进行特征提取。可以和 CFSS 做特征选择进行比较，选择较好的方法。之后选取重要的特征（去掉不使准确率发生显著变化的特征），使用支持向量机或随机森林进行分类预测。

3.3 评估方法

选择一个数据集进行 k-fold 训练，保留另一个数据集作为测试集。最后比较不同方法在测试集上的准确率来评价模型的泛化性。

4 工作计划

初步拟定的工作计划如下：

- 第 11 周：获得数据集，提取特征。
- 第 12 周：用 XGBoost 选择特征。
- 第 13 周：对结果进行分析，改进实验 pipeline。
- 第 14-15 周：撰写实验报告，改进实验 pipeline。

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] K Krishna Kumar, Ganesan Pugalenthi, and Ponnuthurai N Suganthan. Dna-prot: identification of dna binding proteins from protein sequence information using random forest. *Journal of Biomolecular Structure and Dynamics*, 26(6):679–686, 2009.

[5] 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.

[6] 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.

[7] 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.

[8] 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.