《数据建模与计算案例集》

交互式融合特征表示与选择性集成的 DNA 结合蛋白质预测[1]

摘要：DNA 结合蛋白在各种细胞过程中发挥着极其重要的作用，在理解和解释蛋白质功能中，识别 DNA 结合蛋白是一个非常重要的任务。针对 DNA 结合蛋白的识别问题，本案例给出基于蛋白质序列数据的特征表示与选择性集成。首先给出具有交互效应的多信息融合的特征表示模型，它同时考虑了物化属性与进化信息之间的交互效应，以及非相邻残基的位置信息，能够充分挖掘隐藏在蛋白质序列背后的潜在的生物信息，生成具有强判别能力的特征。其次给出基于跳空距离的选择性集成算法，通过对特征表示算法的参数进行扰动，生成不同的输入特征空间。选择性集成算法通过选择(或修剪)得到具有差异性的基分类器，提升整体分类器的泛化能力。最后，本案例设计不同验证实验，在多个数据集，从不同层面进行评价分析，计算实验结果表明，具有交互效应的多信息融合的特征表示，在众多评价指标上均表现优异。在和相关文献的多个经典预测方法进行比较，本案例方法能够进一步提升识别性能。同时，本案例的交互融合特征表示有利于从交互作用的视角去理解 DNA 结合蛋白在细胞中的功能与作用。

关键词：特征表示、选择性集成、DNA 结合蛋白

1 背景介绍

在生物体的细胞中，与 DNA 相关的生命活动是在特定蛋白质的协助下发生的，它们又受到蛋白质-DNA 相互作用的调控[Ptashne, 2005]，这种调控是通过蛋白质与 DNA 链的特异性或者不太特异的结合而实现的。这类与 DNA 结合进而调控 DNA 相关生命活动的蛋白质称为 DNA 结合蛋白(DNA-binding proteins)。DNA 结合蛋白在生物细胞中属于功能蛋白，在各种重要的生命活动中起到至关重要的作用[Jones, 1987]。蛋白质-DNA 相互作用在生物体的遗传和进化机制中起着关键的作用，对蛋白质-DNA 相互作用的研究也是人类探索和理解生物的生长、发育、进化与疾病等生命活动机理的基础，它对蛋白质组的功能诠释和发现遗传病的潜在治疗都至关重要。

利用传统生物实验技术，能够准确识别 DNA 结合蛋白，这些技术包括：过

---

1. 本案例由福建技术师范学院游文杰撰写，案例的知识产权归属作者及所在单位所有。

2. 本案例授权新工科课题组内部研究交流用，不作为其他用途。

3. 本案例源自国家基金科研项目，不涉及企业保密。按照案例要求，在本案例中对有关名称、数据等做了必要的教学处理。

滤绑定位点测定(filter binding assays) [Cajone, 1989]，基因芯片上的染色质免疫沉淀(ChIP-chip) [Buck, 2004]和 X 射线衍射晶体分析法(X-ray crystallography) [Chou, 2003]。然而基于生物学方法的蛋白质结构与功能的测定，需要花费大量的物力和财力，费时又费力。随着蛋白质测序技术的飞速发展，蛋白质序列数据呈爆炸性增长，急需一个有效且可靠的基于生物序列的计算方法，这也是蛋白质组研究领域中重要的课题之一。

基于机器学习的 DNA 结合蛋白的预测方法，通常有两大类：基于蛋白质结构的预测[Zhang, 2010][Tjong, 2007]；基于蛋白质序列的预测[Robert, 2010] [Huang, 2011] [Kumar, 2007] [Shao, 2009] [Lin, 2011]。基于蛋白质结构预测 DNA 结合蛋白能得到较高的识别率，事实上，由于没有足够的蛋白质结构信息，这类方法无法被广泛应用在高通量序列的诠释中。相比较，目前更多的方法是基于氨基酸序列的蛋白质功能预测。大量实验已经表明，多肽或蛋白质一级结构(氨基酸残基排列顺序)相似，其折叠后的空间构象与其功能也很相似[Cai, 2004]。这类基于序列的蛋白质功能(DNA 结合蛋白)预测方法，包含两个主要过程：1)提取蛋白质序列中包含的生物信息，把蛋白质序列转化为相应的数值特征向量；2)利用得到的数值特征向量，使用机器学习中的算法，进行模型训练并对待测序列做预测。

特征向量表示，简称特征表示，就是从蛋白质序列中生成出数值型特征向量，也即将原始的序列数据转换成为能够用于分类的数值特征向量。在已过的几十年间，基于蛋白质序列的有效特征表示方法，主要包括有 1)基于氨基酸组成的方法[Szil ágyi, 2006] [Yu, 2006]，这类方法考虑了相邻的且连续的氨基酸残基间的信息；2)基于伪氨基酸组成的方法[Chou, 2001] [Chou, 2005]，这类方法考虑了非相邻(不连续)氨基酸残基间的信息；以及 3)基于蛋白质频率谱的方法[Ahmad, 2005]，这类方法考虑了蛋白质的进化信息。基于氨基酸组成方法(AAC)，使用序列的统计信息(Dong, 2015)，如常用的 k-mers 方法，这类方法简单，但所生成特征维数较高(20k)，存在维灾和过拟合问题。基于伪氨基酸组成方法，由 Kuo-chen Chou 提出并命名为 PseAAC[Chou, 2001]，它考虑了序列的局部顺序和全局顺序，能够较好的表达序列中的顺序与位置信息，该方法能将序列的位置信息映射到所生成特征向量中。基于蛋白质频率谱的方法，使用携带有进化信息的位置特异性得分矩阵(PSSM: Position Specific Scoring Matrix)，该矩阵表达了与其比对序列相关的同源物信息。大量关于 PSSM 的应用[Kumar, 2007][Ho, 2007][Liu, 2015b][Xu,

2015]表明使用携带进化信息的 PSSM 比序列自身所包含的信息更多也更重要。频率谱的方法通常具有更好的预测效果[Liu, 2015b]，被广泛应用于蛋白质预测中。

研究表明进化信息、物化属性以及序列的结构与位置等信息，对 DNA 结合蛋白的识别均具有一定的作用[Ahmad, 2005][Szilágyi, 2006][Yu, 2006]。如果仅仅采用氨基酸组成信息或者蛋白质频率谱等单个方法，所生成数值特征则都显得过于单一。目前在相关文献中主流的做法是，同时考虑不同的属性(如不同的蛋白质物化属性)和信息(如进化信息与结构信息等)，并对这些方法生成的特征向量进行组合[Zhang, 2016] [Li, 2014]，所生成的高维特征向量作为后继分类器的输入。本案例把这类显式的特征表示方法称为组合式融合特征表示(CFFR: Combined Fusion Feature Representation)。它们将氨基酸的物化属性、进化信息的频率谱以及序列信息(相邻和不相邻残基信息)进行融合，能够取得较好的预测性能[Wei, 2017] [Zhang, 2016]。

所谓集成学习[Zhou, 2009]，是指通过对训练样本的学习，构建多个具有差异性的学习模型(称为基分类器)，然后对这些基分类器使用某种方式进行组合，实现共同解决同一个学习任务。选择性集成[Zhou, 2002]，是指在集成学习的第一阶段(基分类器构建)和第二阶段(分类器组合)之间，增加一个对基分类器的修剪或选择的阶段，其目的是从众多基分类器中选取部分差异大且效果好的基分类器子集，并进行集成。目前，比较直观的选择性集成学习方法，是对基分类器进行排序来达到修剪集成分类器的目的[Tsoumakas, 2008] [Martínez, 2009] [Rokach, 2010] [Zhang, 2009]。

针对 DNA 结合蛋白的识别问题，研究高效的特征表示方法，从序列中生成具有判别信息的特征，并对 DNA 结合蛋白进行准确的判别分类，其具有重要的信息学与生物学意义。针对不种蛋白质物化属性、进化信息和非相邻残基相互作用信息，本案例给出一种具有交互效应的多信息融合特征表示方法，算法能够生成携带有较强判别能力的特征，可以提高 DNA 结合蛋白的预测性能，并且这些特征也有助于从交互作用的视角去理解 DNA 结合蛋白。随后本案例对特征表示算法的参数进行扰动，生成多个基分类器，通过选择(或修剪)得到具有差异性的基分类器，进一步提升整体分类器的识别性能。计算实验表明，基于交互融合的特征表示，相比较于传统的组合式融合的特征表示，其识别效果有显著提高。同时，基于参数扰动的选择性集成，相比较于其它经典预测方法，在识别 DNA 结合蛋白的性能上也有显著提升。

本案例其后各节安排如下：第二节给出 DNA 结合蛋白质的预测方法，包括交互融合的特征表示和选择性集成算法；在多个蛋白质序列数据集上识别 DNA 结合蛋白，与多个经典预测方法的比较与分析放在了第三节；最后在第四节给出总结及未来可能的改进。

2 案例内容

在机器学习实际应用中，通常认为“数据和特征决定了机器学习的上限，而模型和算法能够逼近这个上限”。因此，本案例同时从这两方面着手：1)对多种信息进行有效融合，生成具有强判别能力的特征；2)对多个分类器进行选择集成，生成具有强泛化能力的分类模型。图 1 给出本案例的预测模型框架，包括有交互式融合的特征表示和选择性集成的分类学习。

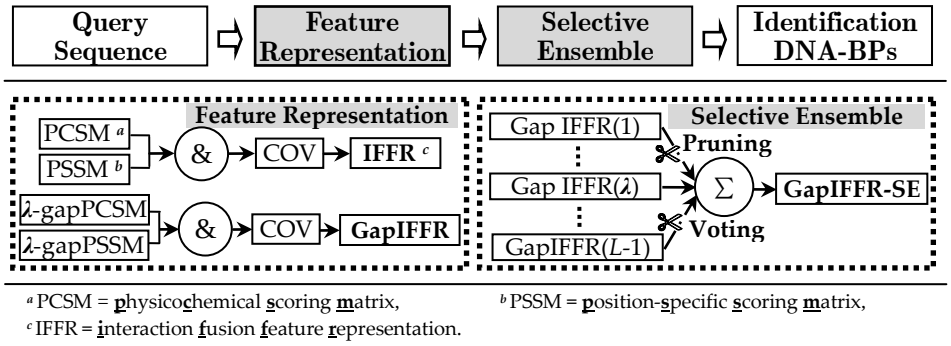


图 1. DNA 结合蛋白预测模型的框架图
左边(虚线框)是交互式融合器(特征表示)，右边(虚线框)是选择性集成器(分类学习)

2.1 假设

考虑蛋白质不同物理化学属性和进化信息，是从蛋白质序列识别 DNA 结合蛋白的关键。常见的组合式融合特征表示(CFFR)，在一定程度上，同时考虑蛋白质的物化属性、进化信息以及序列局部位置等信息，它能够提升识别 DNA 结合蛋白的能力。然而，CFFR 方法把物化属性与进化信息等均视为彼此独立的特征向量进行组合，忽略了它们之间还应该存在着交互效应。因此，本案例专注于具有交互效应的多信息融合的特征表示，考查不同的属性(蛋白质物化属性等)和信息(蛋白质进化信息等)之间是否存在交互效应？以及这种交互效应能否提升 DNA 结合蛋白的识别能力？因此，本案例假设：

假设 1: 不同的物化属性和进化信息之间存在交互效应。

在本案例，把考虑不同物化属性和进化信息之间的交互效应的特征表示，记为交互式融合(IFFR: Interaction Fusion Feature Representation)。另外，由于在蛋白质序列家族中，氨基酸的替换模式是高度特异的，并且同一蛋白质序列中不同距离的氨基酸残基之间存在的相互作用，本案例给出基于不同跳空距离的交互式融合特征表示，也即在二重信息(物化属性和进化信息)交互式融合基础上，实现不同距离的跳空操作(λ-gap)，给出三重信息融合特征表示算法λ-gapIFFR。本案例假设：

假设 2: 在交互式融合特征表示分析框架下，三信息融合λ-gapIFFR 优于二信息融合λ-gapPSSM。

多信息交互式融合特征表示λ-gapIFFR，其实质是对蛋白质序列的不同物化属性和进化信息进行具有交互效应的特征融合，并引入序列的跳空片段信息，因此，该算法同时考虑了蛋白质序列的不同物化属性、进化信息和序列局部顺序等信息。

2.2 模型

特征表示，是根据序列中的数学关系以及生物化学属性等指标，将由字符组成的序列，数值化成一个固定维数的特征向量，包含显性的特征和隐性的特征。针对蛋白质序列的特征表示，本节给出新的交互融合特征表示模型，该模型能够同时考虑不种信息自身内部的相关性(显性特征)，以及信息与信息之间的交互效应(隐性特征)。首先给出相关的概念描述，然后引出具有交互效应的多信息融合特征表示模型。

定义 1 (得分矩阵 Scoring Matrix)

给定任一(蛋白质)序列$S = R_1R_2\cdots R_L$，定义得分矩阵，

$$P = (p_i^{(j)})_{L\times M} = (p^{(1)}, p^{(2)}, \cdots, p^{(M)}) \tag{1}$$

其中$p_i^{(j)}$ $(i = 1,2,\cdots,L)$是序列中第i个氨基酸残基R_i在第j种指标上的得分，L为(蛋白质)序列S的长度，列数M为事先给定的指标个数。

考虑到蛋白质序列中不同距离的氨基酸残基之间存在着相互作用，借鉴伪氨基酸组成(非相邻残基)分析思想[Chou, 2001]，给出λ-gap 得分矩阵定义。

定义 2 (λ-gap 得分矩阵 λ-gap Score Matrix)

给定得分矩阵 $P=(p_{ij})_{L\times M}$ 和参数 λ，定义矩阵

$$G_\lambda = A_\lambda P = A_\lambda \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_L \end{pmatrix} = \begin{pmatrix} p_1 + p_{\lambda+1} \\ p_2 + p_{\lambda+2} \\ \vdots \\ p_{L-\lambda} + p_L \end{pmatrix} \tag{2}$$

为 λ-gap 得分矩阵，其中 $A_\lambda = (a_{ij})_{(L-\lambda)\times L}$ 为(0-1)矩阵，$a_{ij} \in \{0,1\}$，即

$$A_\lambda = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L-\lambda} \end{pmatrix} = \begin{pmatrix} \underbrace{1,0,\cdots,1}_{\lambda},0 & \cdots & 0,0,\cdots,0,0 \\ 0,\underbrace{1,\cdots,0,1}_{\lambda} & \cdots & 0,0,\cdots,0,0 \\ \vdots & \cdots & \vdots \\ 0,0,\cdots,0,0 & \cdots & 0,\underbrace{1,\cdots,0,1}_{\lambda} \end{pmatrix} \tag{3}$$

其中参数 λ $(1 \le \lambda \le L\text{-}1)$ 表示矩阵 A_λ 中任一行向量 a_i 中两个非零元 1 之间的距离(λ-gap)。特别地，当 $\lambda = 0$ 时，A_0 退化为单位矩阵 I_L，也即 0-gap 得分矩阵

$$G_0 = A_0 P = IP = P \tag{4}$$

λ-gapSM 间接刻画了序列中不相邻残基之间(跳空距离为 λ)的位置信息。

定义 3 (得分协方差矩阵 Score Covariance Matrix)

给定 λ-gap 得分矩阵 $G_\lambda = (g_{ij})_{(L-\lambda)\times M}$，定义协方差矩阵

$$\Sigma = Cov(G_\lambda) = G_\lambda^T G_\lambda = (\sigma_{ij})_{M\times M} \tag{5}$$

为 λ-gap 得分协差阵；显然，矩阵 Σ 为对称方阵。

定理 1 (矩阵向量化 Matrix Vectorization)

设对称方阵 $\Sigma = (\sigma_{ij})_{M\times M}$ 的上三角矩阵为 U，即

$$U = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1,M} \\ & \sigma_{22} & \cdots & \sigma_{2,M} \\ & & & \vdots \\ & & & \sigma_{M,M} \end{pmatrix} \tag{6}$$

对 U 按列拉直运算(matrix vec operator)，并保留元素 σ_{ij} 满足 $i \le j$，可得，

$$v = vec(U) = (\sigma_{1,1},\sigma_{1,2},\sigma_{2,2},\cdots,\sigma_{1,M},\sigma_{2,M},\cdots,\sigma_{M,M}) \tag{7}$$

则所得向量 v 的维数仅与 M 有关，而与 L(序列长度)和 λ(跳空距离)无关。

证明：由于v中任一元素σ_{ij}必须满足$i \le j$，也就是，矩阵(6)的上三角元素构成向量v，所以向量v的维数等于$1+2+\cdots+M$。也即，向量v的维数$M(M+1)/2$，仅与M有关，而与L(序列长度)和参数λ均无关。□

在蛋白质序列分析中常用的得分矩阵，如位置特异性得分矩阵(PSSM：position-specific score matrix)，它是一个行数为 L(L 为序列长度)列数为 20(20 类标准氨基酸)的矩阵。蛋白质数据搜索程序 PSI-BLAST，能够通过多次迭代寻找最优结果(Altschul, 1997) [Schäffer, 2001]，对于寻找蛋白家族的新成员或者发现远亲物种的相似蛋白非常有效，使用它能够生成一个**位置特异得分矩阵 PSSM：**

$$P=\begin{bmatrix} p_1^{(1)} & p_1^{(2)} & \cdots & p_1^{(20)} \\ p_2^{(1)} & p_2^{(2)} & \cdots & p_2^{(20)} \\ \vdots & \vdots & & \vdots \\ p_L^{(1)} & p_L^{(2)} & \cdots & p_L^{(20)} \end{bmatrix}_{L\times 20} \tag{8}$$

元素$p_i^{(j)}$表示蛋白质进化过程中蛋白质序列第i个位置($1 \le i \le L$)的氨基酸残基R_i突变为第j类($1 \le j \le 20$)氨基酸的概率(对数似然得分)，取值越大说明替换的可能性越大。该矩阵表达了序列的进化信息(Liu, 2015b)。关于 PSSM 的详细计算步骤见附录 1。

同时，本案例还给出氨基酸物化属性得分矩阵(PCSM)。在对 DNA 结合蛋白的识别过程中，本案例假设不同氨基酸物化属性对预测结果将产生不同的贡献，因此，在蛋白质的特征表示过程中，应该考虑合适的氨基酸物化属性。AAindex[Kawashima, 2008]是一个包含多个氨基酸物理化学属性的氨基酸指数表，其中 AAindex1 部分的每一项表示氨基酸的某种物理化学属性量化后的数据，含有 20 个数值。对于第j种物化属性$Q^{(j)}$，任一条蛋白质序列S可表示为$q_1^{(j)},q_2^{(j)},\cdots,q_L^{(j)}$，其中$L$是序列长度，$q_i^{(j)}$ $(1 \le i \le L)$是序列中第i个氨基酸残基R_i的第j种物化属性指数。假设考虑有M种物化属性，则有氨基酸**物化属性得分矩阵 PCSM**

$$Q=\begin{bmatrix} q_1^{(1)} & q_1^{(2)} & \cdots & q_1^{(M)} \\ q_2^{(1)} & q_2^{(2)} & \cdots & q_2^{(M)} \\ \vdots & \vdots & & \vdots \\ q_L^{(1)} & q_L^{(2)} & \cdots & q_L^{(M)} \end{bmatrix}_{L\times M} \tag{9}$$

在本案例的实验部分，本案例仅使用文献[Shen, 2008]的 6 个物化属性进行分析，

它们分别是：(1) Hydrophobicity; (2)　Hydrophilicity; (3) Mass; (4) pK1(a-CO2H); (5) pK2(NH3); (6) pI(25°C)。详细的氨基酸物化指数见附录 2。

显然，对于得分矩阵 PSSM 和 PCSM，由定义 2 可分别得到对应的得分矩阵 λ-gapPSSM 和λ-gapPCSM。给定长度为 L 的蛋白质序列，有 PSSM 矩阵 P 和 PCSM 矩阵 Q，水平拼接得到矩阵 $W=(P,Q)=(w_{ij})_{L\times(M+20)}$，由定义 2，可得$\lambda$-gap 得分矩阵($\lambda$-gapSM)，即

$$G_\lambda = A_\lambda W = A_\lambda(P,Q) = A_\lambda P, A_\lambda Q \tag{10}$$

由定义 3 和分块矩阵运算，容易得到，

$$\begin{aligned}\Sigma &= Cov(G_\lambda) = (A_\lambda P, A_\lambda Q)^T (A_\lambda P, A_\lambda Q) \\ &= \begin{pmatrix} P^T A_\lambda^T \\ Q^T A_\lambda^T \end{pmatrix} (A_\lambda P, A_\lambda Q) = \begin{pmatrix} P^T A_\lambda^T A_\lambda P & P^T A_\lambda^T A_\lambda Q \\ Q^T A_\lambda^T A_\lambda P & Q^T A_\lambda^T A_\lambda Q \end{pmatrix}_{(M+20)\times(M+20)}\end{aligned} \tag{11}$$

由定理 1，上式所对应的特征向量的维数也仅与 M 有关，与序列长度 L 和参数λ 均无关。

上面所给特征表示模型中，本案例分别利用了物化属性Q和进化信息P各自本身所蕴含的相关性信息$Q^T A_\lambda^T A_\lambda Q$和$P^T A_\lambda^T A_\lambda P$，生成显性特征。同时，还考虑了物化属性和进化信息之间的**交互效应项**$Q^T A_\lambda^T A_\lambda P$（或$P^T A_\lambda^T A_\lambda Q$），生成隐性特征。其中$A_\lambda^T A_\lambda$刻画了非相邻残基(距离为$\lambda$)的位置信息。特别地，当跳空距离 $\lambda = 0$（$A_0 = I$）时，(11)式退化为

$$\Sigma = (P,Q)^T (P,Q) = \begin{pmatrix} P^T \\ Q^T \end{pmatrix} \begin{matrix} P & Q \end{matrix} = \begin{pmatrix} P^T P & P^T Q \\ Q^T P & Q^T Q \end{pmatrix}_{(M+20)\times(M+20)} \tag{12}$$

Table 1
Feature Representation Methods Developed within the Framework of the Proposed Model.

Scoring Matrix	Matrix Dimension	Cov (**Def.3**)	Vector Dimension **(Theorem1)**	Feature Representation
SM(**Def.1**)	L×M	M×M	M(1+M)/2	
PCSM	L×6	6×6	21	CovPCSM
PSSM	L×20	20×20	210	CovPSSM
			231	CFFR [a]
	L×26	26×26	351	IFFR [b]
λ-gapSM(**Def.2**)	(L-λ)×M	M×M	M(1+M)/2	
λ-gapPCSM	(L-λ)×6	6×6	21	GapPCSM
λ-gapPSSM	(L-λ)×20	20×20	210	GapPSSM
			231	GapCFFR
	(L-λ)×26	26×26	351	GapIFFR

[a] CFFR = **C**ombined **F**usion **F**eature **R**epresentation with PCSM and PSSM.
[b] IFFR = **I**nteraction **F**usion **F**eature **R**epresentation with PCSM and PSSM.

表 1 汇总了在所提特征表示模型框架下，不同的特征表示方法的相关信息，包括得分矩阵的维数，所生成的特征向量的维数，以及特征表示方法的简称与缩略词等。其中，CovPCSM 考虑了 6 个不同物化属性及其这些物化属性自身内部的相关性，生成的特征维数 dimension=21。CovPSSM 考虑了序列在 20 个氨基酸上的进化信息及其自身内部的相关性，生成的特征维数 dimension =210。而 CFFR 方法是对它们二者进行简单串联组合，生成的特征维数为两者之和 dimension=231。IFFR 不仅考虑了 6 个物化属性自身内部和进化信息自身内部的相关性，并且更进一步考虑了物化属性和进化信息之间的交互效应项，生成的特征维数 dimension=351。在本案例中还考虑了序列中不相邻残基的相互作用信息，给出考虑跳空距离为λ的多信息融合特征表示方法：GapPSSM、GapCFFR 和 GapIFFR。

2.3 算法

针对 DNA 结合蛋白预测问题，本节分别给出交互式融合的特征表示算法(算法 1)和选择性集成的分类学习算法(算法 2)。

1)**交互式融合特征表示算法：**

基于所提特征表示模型，给出新的特征表示算法，即多重信息交互式融合的特征表示算法 GapIFFR。该算法考虑了不同物化属性和进化信息的交互效应，同时还考虑了序列中不相邻氨基酸残基间的作用信息。详细算法如下：

Algorithm 1 Gap-based Interaction Fusion Feature Representation (**GapIFFR**)

Input: $\boldsymbol{seq\_FASTA}$ // Query protein sequence
$\boldsymbol{\lambda}$ // Distance of gaps
Output: $\boldsymbol{v}$ // Numeric vector

1: **Initialization**: $\boldsymbol{L}$ = length of sequence $\boldsymbol{seq\_FASTA}$, $\boldsymbol{\lambda} \leqslant \boldsymbol{L}$-1
2: Obtain PSSM matrix $\boldsymbol{P}$ by calling **PSI-BLAST** (Set *evalue*=0.001, *num_iterations*=3): $\boldsymbol{P} = (p_i^{(j)})_{L\times 20}$
3: Obtain PCSM matrix $\boldsymbol{Q}$ from **AAindex** dataset: $\boldsymbol{Q} = (q_i^{(j)})_{L\times M}$
4: Horizontally concatenate $\boldsymbol{P}$ and $\boldsymbol{Q}$: $\boldsymbol{W} = [\boldsymbol{P}\ \&\ \boldsymbol{Q}] = (w_{ij})_{L\times(20+M)}$
5: Computing matrix $\boldsymbol{G_\lambda}$ in term of **Definition2**:
$$\boldsymbol{G_\lambda} = A_\lambda W = (g_{ij})_{(L-\lambda)\times(20+M)}$$
6: Compuing matrix $\boldsymbol{C}$ in term of **Definition3**:
$$\boldsymbol{C} = \mathrm{cov}(G_\lambda) = G_\lambda^T G_\lambda = (c_{ij})_{(20+M)\times(20+M)}$$
7: **Return** a row vector $\boldsymbol{v}$ in term of **Theorem1**:
$$\boldsymbol{v} = (c_{1,1}, c_{2,1}, \cdots, c_{20+M,1}, c_{1,2}, c_{2,2}, \cdots, c_{20+M,2}, \cdots, c_{1,20+M}, c_{2,20+M}, \cdots, c_{20+M,20+M})$$

算法 1 的输入参数 λ，也即序列残基之间的跳空距离，当λ=0 时，以上特征表示算法仅考虑了序列的不同物化属性和进化信息，算法 1 实现的是二信息交互式融合 IFFR。特别地，算法 1 的第 4 行 $\boldsymbol{W}=\boldsymbol{P}$ 时(即忽略 $\boldsymbol{Q}$)，实现 GapPSSM 算法；算法 1 的第 4 行 $\boldsymbol{W}=\boldsymbol{Q}$ 时(即忽略 $\boldsymbol{P}$)，实现 GapPCSM 算法；而 GapCFFR 算法返回的特征向量也就是这两算法所生成特征向量的合并(详见表 1)。

2)选择性集成学习算法：

给定蛋白质序列集，随机划分训练集S_{trn}，验证集S_{val}和测试集S_{tst}。假设$D_{trn}^{(\lambda)} = \{(\mathrm{x}_i^{(\lambda)}, y_i)\}$为对应于$S_{trn}$的训练集，其中任一训练样本$(\mathrm{x}_i^{(\lambda)}, y_i)$的输入变量$\mathrm{x}_i^{(\lambda)} = (x_{i1}^{(\lambda)}, x_{i2}^{(\lambda)}, \cdots, x_{ip}^{(\lambda)}) \in \mathbb{R}^p$是由算法 1 得到的跳空距离为$\lambda$的$p$维特征向量，输出变量为$y_i \in Y = \{+1, -1\}$。同理可得验证集$D_{val}^{(\lambda)}$和测试集$D_{tst}^{(\lambda)}$。在$D_{trn}^{(\lambda)}$ $(1 \le \lambda \le L-1)$上训练基分类器C_λ，构成集合$T = \{C_1, C_2, \cdots, C_{L-1}\}$，$\tilde{T}$为 T的任一子集，计算子集$\tilde{T}$对应的集成基分类器在相应的验证集$D_{val}^{(\lambda)}$上的泛化误差$\varepsilon(\tilde{\mathrm{T}})$，选取泛化误差最小的子集$T^* = \arg\min_{\tilde{\mathrm{T}}\subset\mathrm{T}} \varepsilon(\tilde{\mathrm{T}})$。

理论上，最优基分类器子集T^*可通过穷举法得到。然而，当 L 较大时，穷举法的计算量太大。一种简单直观的选择策略是：对基分类器C_i按性能指标 M 进行排序，选取前 k(奇数)个基分类器构成的子集T^*做为对集成分类器T的修剪，并对子集$T^* \subset T$采用投票(Max-Wins Voting, MWV)策略[Moreira M, 1998]进行表决。以下给出基于 GapIFFR 的选择性集成算法：

Algorithm 2 GapIFFR-based Selective Ensemble (**GapIFFR-SE**)

Input: $\boldsymbol{S}_{trn}$, $\boldsymbol{S}_{val}$, $\boldsymbol{S}_{tst}$, $\boldsymbol{C}$, $\boldsymbol{M}$, $\boldsymbol{k}$ // $\boldsymbol{C}$ is a base classifier algorithm, // $\boldsymbol{M}$ is the evaluation criteria (such as Accuracy, MCC, etc.)

Output: $\boldsymbol{Y}$ // class label of the test dataset $\boldsymbol{S}_{tst}$.

(1) Initialization process:

—Set $\boldsymbol{T}=\Phi$, $\boldsymbol{L}$=minimum sequence length of $\boldsymbol{S}_{trn}$, $\boldsymbol{S}_{val}$ and $\boldsymbol{S}_{tst}$, calculate $\boldsymbol{D}_{trn}(\boldsymbol{\lambda})$, $\boldsymbol{D}_{val}(\boldsymbol{\lambda})$ and $\boldsymbol{D}_{tst}(\boldsymbol{\lambda})$ by calling **GapIFFR** with $\boldsymbol{\lambda}$={1,2,⋯,L-1}.

(2) Training base classifiers process:

—**For** i=1 to L-1 **do**

——Update $\boldsymbol{T}=\boldsymbol{T}\cup\boldsymbol{C}_i$, where the base classifier $\boldsymbol{C}_i$ is trained on the training dataset $\boldsymbol{D}_{trn}(i)$ using the given classifier $\boldsymbol{C}$.

—**EndFor**

(3) Selection (Pruning) process:

—**For** j=1 to L-1 **do**

——Calculate $\boldsymbol{M}_j$ for each base classifier $\boldsymbol{C}_j\in\boldsymbol{T}$ on the validation dataset $\boldsymbol{D}_{val}(j)$ using the evaluation criteria $\boldsymbol{M}$.

—**EndFor**

—Sort $\boldsymbol{M}_j$ in descending order, and select $\boldsymbol{T}^*=\{\boldsymbol{C}_{\lambda 1},\boldsymbol{C}_{\lambda 2},\cdots,\boldsymbol{C}_{\lambda k}\}\subset\boldsymbol{T}$, where $\boldsymbol{C}_{\lambda 1},\boldsymbol{C}_{\lambda 2},\cdots,\boldsymbol{C}_{\lambda k}$ correspond to the top $\boldsymbol{k}$ of the $\boldsymbol{M}_j$ values.

(4) Ensemble (Voting) process:

—Predict the class label of the test dataset $\boldsymbol{S}_{tst}$,

$$\boldsymbol{Y}=\boldsymbol{sign}\{\sum_{t=1}^{k}\boldsymbol{C}_{\lambda_t}(\mathbf{X})\}$$

where $\boldsymbol{C}_{\lambda t}$ is the $\boldsymbol{\lambda}_t$-th base predictor on the dataset $\mathbf{X}\in\boldsymbol{D}_{tst}(\boldsymbol{\lambda}_t)$, $\{\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\cdots,\boldsymbol{\lambda}_k\}\subset\{1,2,\cdots,L\text{-}1\}$.

Return $\boldsymbol{Y}$

算法 2 选择性集成 GapIFFR-SE，其实质是对参数λ进行扰动，生成不同的输入特征空间，并通过选择(或修剪)得到具有差异性的基分类器子集，达到提升集成分类器的性能。

3 实验

3.1 实验数据与评价指标

为了验证所提方法的有效性，选取 6 个 DNA 结合蛋白序列数据(包含 1 组独立测试集)进行分析，它们的样本容量相对较充足(≥300)，同时它们又都是序列

同源性小于 40%的数据集，这些能保证实验结果的相对可信性。表 2 给出数据的汇总信息与数据来源[2]。

Table 2
Datasets Used for Training, Testing and Benchmarking Study in This Paper.

Dataset	Number of Proteins			Min Length	Similarity
	DNA-BP	non-DNA-BP	Total		
Alternate Dataset	1153	1153	2306	51	≤25%
PDB1075 Dataset (Liu 2014)	525	550	1075	50	≤25%
Independent1 Dataset (Kumar 2009)	823	823	1646	35	≤40%
Independent2 Dataset (Kumar 2009)	88	233	321	30	≤40%
Training Dataset (Kumar 2007)	146	250	396	26	≤25%
Testing Dataset (Wang and Brown 2006)	92	100	192	45	≤25%

为了客观系统评估所提方法的预测性能，本案例分别采用 Jackknife 校验法、k-fold 交叉校验法(k-foldCV)和独立校验法(HoldOut)对算法进行比较和评估。其中 k-foldCV 能够有效降低由于数据不充分而造成的过学习和欠学习状态的发生，在实践中，10-foldCV 被认为是标准方法；Jackknife 校验法被认为是较客观的统计校验方法，它能够避免由于训练和测试数据的随机划分而造成的随机性，保证实验结果的可复制性；而独立校验法(HoldOut)则能够反映算法对新鲜样本(独立测试数据集)的预测能力。

对算法性能的评估指标有：预测准确率(ACC: Accuracy)、敏感性(SE：Sensitivity)、特异性(SP：Specificity)和综合评价预测结果的相关性系数 Mathews 相关系数(MCC：Mathews Correlation Coefficient)，详细定义如下：

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

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

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

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

[2] http://www.imtech.res.in/raghava/dnabinder/download.html
http://server.malab.cn/Local-DPP/Datasets.html
http://www3.ntu.edu.sg/home/EPNSugan/index_files/dnaprot.htm

其中，*TP* (真阳性)表示 DNA 结合蛋白被预测为 DNA 结合蛋白的个数，*TN* (真阴性)表示非 DNA 结合蛋白被预测为非 DNA 结合蛋白的个数， *FP* (假阳性)表示非DNA 结合蛋白被错误预测为DNA 结合蛋白的个数，*FN* (假阴性)表示 DNA 结合蛋白被错误预测为非 DNA 结合蛋白的个数。

ACC 表示预测结果中真阳性与真阴性之和在总测试实例中的百分比；SE 表示真阳性在所有预测为阳性测试数据中的百分比；SP 表示真阴性在所有预测为阴性测试数据中的百分比。对于完美的预测系统，这三指标都应该达到 100%。然而，对于非平衡数据集，若 SE 增加时，则 SP 必然下降，反之亦然，这些指标不能很好的评估预测结果，相比较 MCC 是个更平衡的评估标准，其取值范围在[-1,+1]之间，值为 1 表示预测结果与真实类别完全相关，值为 0 表示是完全随机的预测，值为-1 表示完全相反的相关性。另外，ROC 曲线图中曲线下面积(area under the curve, AUC)可以作为更加客观的分类性能评估标准。ROC 曲线图是一个单位平方，两坐标轴(真阳性率和假阳性率)的数值从 0 到 1，AUC 最大值为 1，对应于完美分类器。

必须指出的是，本案例中用于比较的实验结果均是使用基分类器：线性核 SVM(参数默认)，由于本案例更多专注于蛋白质序列的特征表示方法，文中不对分类器做任何的优化。事实上，可以通过调整分类器及其参数，以及选用更为有效的物化属性子集，可以得到更高的预测结果。

3.2 二信息交互融合特征表示的评估

本节实验主要讨论特征表示的模型选择问题，针对 DNA 结合蛋白的预测问题，比较并评估基于物化属性与进化信息的二重信息交互式融合特征表示 IFFR 的性能。

首先，在 4 个基准数据集上利用 Jackknife 验证比较 CovPCSM，CovPSSM，CFFR 和 IFFR 四个算法性能，结果如表 3。这里 CovPCSM 方法只单一的考虑物化属性自身，识别效果一般。同理，CovPSSM 方法也只单一的考虑进化信息自身，但识别效果较好。而 CFFR 方法是对它们二者进行简单串联组合，所生成特征向量同时考虑物化信息和进化信息，识别效果略优于 CovPSSM 的结果。IFFR 方法不仅考虑了物化属性内部和进化信息内部的相关性，并且更进一步考虑了物化属性和进化信息之间的交互效应项，也因此取得更好的识别性能。

Table 3

Prediction performance comparison on four independent datasets using IFFR (CFFR) models which integrates the evolutionary- and physicochemical-based information (Jackknife validation test)

DataSet	Measurements	Feature Representation Method			
		CovPCSM	CovPSSM	CFFR [a]	IFFR [b]
Alternate Dataset	MCC	0.3015	0.4701	0.4735	**0.4824**
	ACC(%)	63.62	73.11	73.29	**73.76**
	SE(%)	85.08	82.22	**82.31**	**82.31**
	SP(%)	42.15	64.01	64.27	**65.22**
PDB1075 Dataset	MCC	0.3882	0.5266	0.5504	**0.5533**
	ACC(%)	68.65	76.00	77.21	**77.40**
	SE(%)	57.27	69.64	71.09	**71.82**
	SP(%)	80.57	82.67	**83.62**	83.24
Independent1 Dataset	MCC	0.6881	0.9612	0.9612	**0.9624**
	ACC(%)	84.14	98.06	98.06	**98.12**
	SE(%)	78.01	97.57	97.45	**97.57**
	SP(%)	90.28	98.54	**98.66**	**98.66**
Independent2 Dataset	MCC	NaN	0.6826	0.6761	**0.6937**
	ACC(%)	72.59	87.23	86.92	**87.85**
	SE(%)	0.00	**78.41**	**78.41**	77.27
	SP(%)	100.00	90.56	90.13	**91.85**
Training Dataset	MCC	0.4050	0.6922	0.7099	**0.7197**
	ACC(%)	72.22	85.61	86.36	**86.87**
	SE(%)	77.60	88.00	88.00	**88.80**
	SP(%)	63.01	81.51	**83.56**	**83.56**

[a] CFFR = Combined Fusion Feature Representation;
[b] IFFR = Interaction Fusion Feature Representation.
Note: Bold values are the best prediction results.

然后，以下在 4 个独立数据集上，进一步考查所提特征表示算法 IFFR 与三个经典的特征表示算法(PsePSSM, PseAAC 和 AAC)[文献]的性能比较，为使比较的结果更加客观可信，实验使用 30 次的 10-fold CV 校验结果进行分析。

从图 2 知，在数据集 Alternate Dataset, PDB1075 Dataset 和 Independent2 Dataset 中，基于 IFFR 特征表示算法具有卓越的性能，其平均性能均优于其它算法(PsePSSM, PseAAC 和 AAC)。在全部数据集中，IFFR 特征表示通常有较小的标准误差，这在某种程度上说明 IFFR 特征表示对训练样本集的随机构成不敏感，鲁棒性更好。数据集 Independent1 Dataset 中，基于 PsePSSM 特征表示算法也有很好的表现，明显优于 PseAAC 和 AAC 的结果。这是因为 IFFR 与 PsePSSM 都使用了 PSSM 进化信息，也就是 PSSM 所携带的进化信息比序列自身所包含的信息更为丰富也更加重要，因此，考虑进化信息能够达到提升预测性能的目的。总之，相比较于经典算法(PsePSSM, PseAAC 和 AAC)，在 4 个独立数据集中本案例的 IFFR 特征表示更具优势。

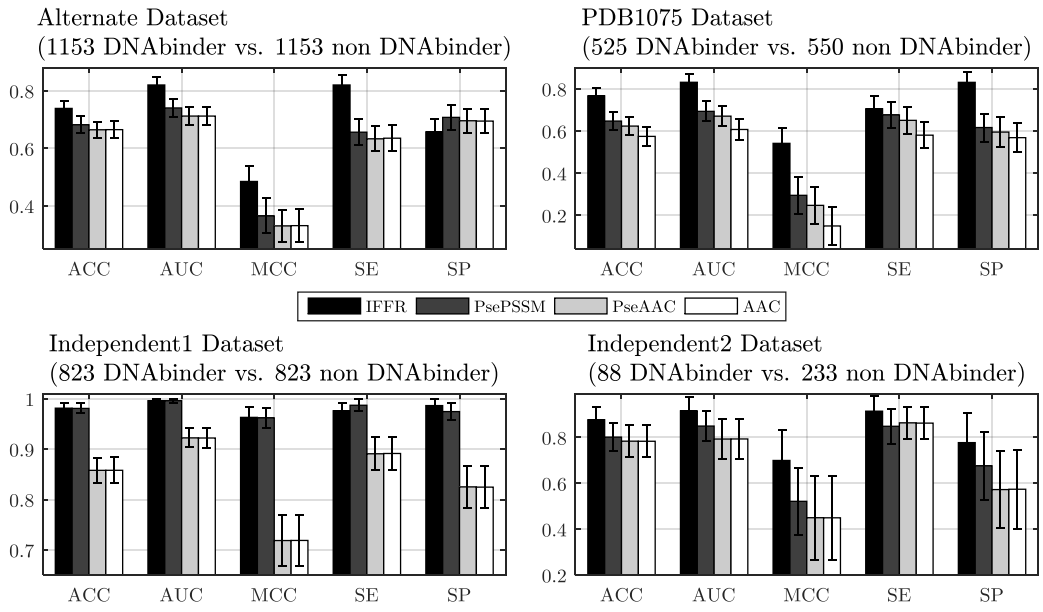


图 2 不同特征方法的性能指标(Accuracy, AUC, MCC, Sensitivity and Specificity)比较(30 次的 10-fold CV).

Fig. 2. Comparison of IFFR with existing feature representation methods on four independent data sets. (30 random results of 10-fold CV, base classifier: linear SVM)

综上，由表 3 和图 2 可以观察到：基于物化属性和进化信息的(双信息)交互融合 IFFR 的预测成功率要高于其它单信息特征表示算法 CovPCSM, CovPSSM 和 AAC，也要高于其它双信息(或组合式)融合算法 CFFR, PseAAC, PsePSSM 等。这些实验结果也表明：在 DNA 结合蛋白中存在着物化属性与进化信息之间的交互效应，并且这种交互效应的隐式特征能够提高识别率。因此，交互融合特征表示同时刻画了显式特征和隐式特征，它可以更全面的描述 DNA 结合蛋白的信息。这就验证了假设 1 的结论。

3.3 参数敏感性分析与模型比较

本节实验主要讨论多重信息融合特征表示模型中的参数选择问题，考查算法的参数λ的敏感性，也即在所提模型框架下，不同的跳空距离λ对结果的影响。为保证实验结果的可重复性以及后继的可比较性，本案例仍使用 Jackknife 校验法进行分析。

以下在 4 个独立数据集上进行 Jackknife 校验比较，其中算法参数(跳空距离)λ分别从 1 连续变化至 L-1(L 为蛋白质序列的长度)，观测三个不同算法 GapPSSM，GapCFFR 和 GapIFFR 在 4 个指标(MCC, ACC, SE 和 SP)上的变

化。结果如下图。

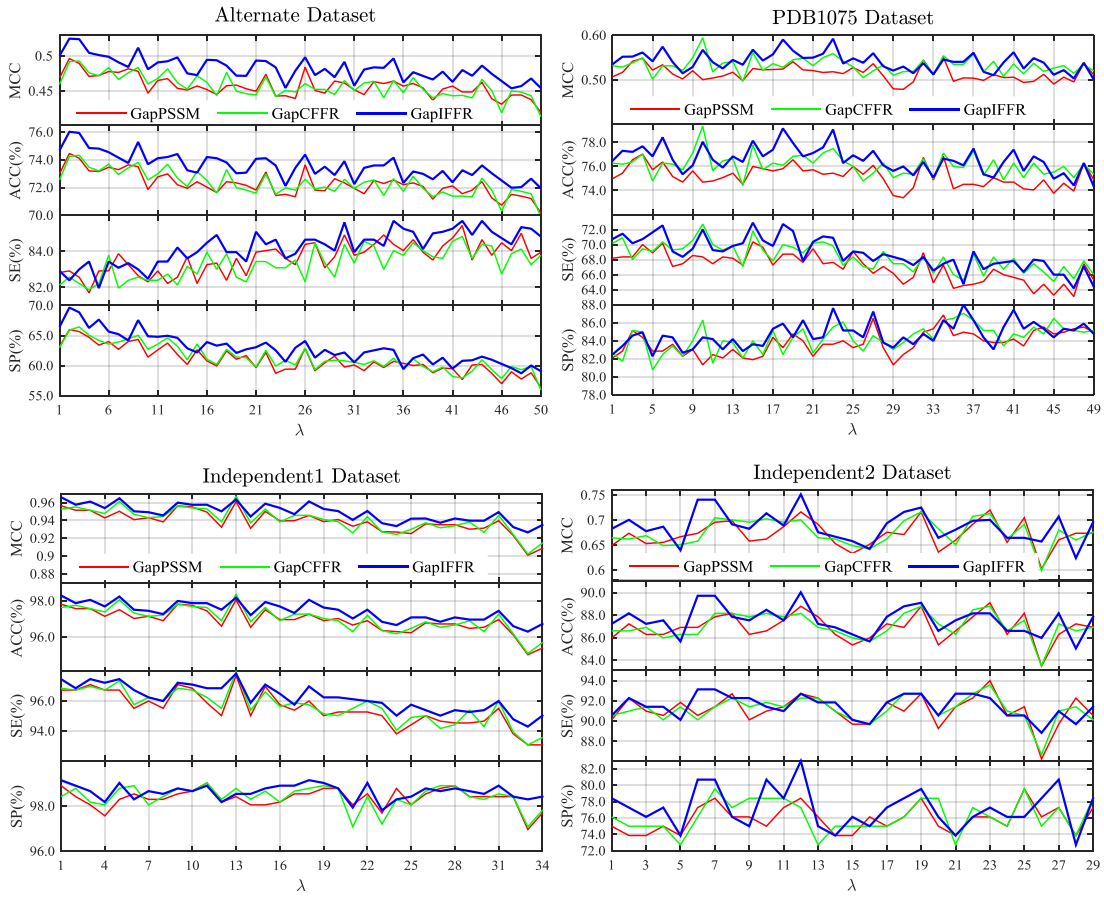


图 3 多信息融合特征表示 GapPSSM, GapCFFR 和 GapIFFR 的性能(MCC, Accuracy, Sensitivity 和 Specificity)比较以及参数 λ 对结果的影响
(Jackknife validation test, base classifier: linear SVM)

Note: FRDIF(λ-gapPSSM) models with evolutionary- and sequence- information; FRTIF(λ-gapPPSM) models with evolutionary-, physicochemical- and sequence- information;

从图3容易看出，性能曲线的波动幅度较大，也即不同跳空距离(参数λ)对分类器性能指标的影响较明显，说明识别结果对参数λ较敏感。对于不同的数据集，参数λ对预测结果的影响各不相同，其中预测性能最优的参数λ值也都不相同。例如，对于性能指标MCC而言，在数据集Alternate Dataset上最优参数λ取值为2，在数据集PDB1075 Dataset上最优参数λ取值为23，在数据集Independent1 Dataset上最优参数λ取值为1，在数据集Independent2 Dataset上最优参数λ取值为12。这也许是因为在序列中两个氨基酸残基的跳空距离为λ，与它们在三维空间上的实际距离并不存在必然联系。

从图3也容易看出，三信息交互式融合GapIFFR的性能要优于组合式融合GapCFFR，并且也优于两信息交互式融合算法GapPSSM。更进一步，利用上图的Jackknife校验结果做统计显著性检验，为使结果更为可信，同时选取配对参

数T检验和配对非参数符号秩检验，验证不同特征表示方法对预测性能是否存在显著性差异。因为这些方法的预测性能是在相同的训练集和测试集上进行，因此在配对统计检验中，任何的差异均可认为来源于假设(算法)间的差异，而不存在样本集的随机组成差异。

Table 4
Statistical Significance Test (Parametric and Non-parametric Statistical Test)

DataSet	Evaluation indice	GapPSSM vs. GapIFFR		GapCFFR vs. GapIFFR	
		Paired T-test	signed-rank test	Paired T-test	signed-rank test
Alternate Dataset	MCC	(-) 1.353×10^{-20}	(-) 7.557×10^{-10}	(-) 1.018×10^{-19}	(-) 7.557×10^{-10}
	ACC	(-) 2.756×10^{-20}	(-) 7.513×10^{-10}	(-) 5.259×10^{-19}	(-) 7.977×10^{-10}
	SE	(-) 2.502×10^{-8}	(-) 6.470×10^{-7}	(-) 4.528×10^{-14}	(-) 2.710×10^{-9}
	SP	(-) 1.624×10^{-15}	(-) 2.159×10^{-9}	(-) 3.492×10^{-12}	(-) 2.056×10^{-8}
PDB1075 Dataset	MCC	(-) 2.753×10^{-13}	(-) 3.775×10^{-9}	(-) 0.0016	(-) 0.0026
	ACC	(-) 1.207×10^{-13}	(-) 4.657×10^{-9}	(-) 0.0013	(-) 0.0023
	SE	(-) 3.765×10^{-13}	(-) 5.556×10^{-9}	(-) 0.0037	(-) 0.0096
	SP	(-) 6.848×10^{-7}	(-) 2.244×10^{-6}	(-) 0.0248	(-) 0.0342
Independent1 Dataset	MCC	(-) 3.390×10^{-12}	(-) 3.653×10^{-7}	(-) 2.585×10^{-9}	(-) 7.443×10^{-7}
	ACC	(-) 2.768×10^{-12}	(-) 3.444×10^{-7}	(-) 2.325×10^{-9}	(-) 6.871×10^{-7}
	SE	(-) 5.013×10^{-11}	(-) 3.495×10^{-7}	(-) 2.159×10^{-8}	(-) 4.131×10^{-6}
	SP	(-) 5.993×10^{-5}	(-) 1.278×10^{-4}	(-) 6.478×10^{-4}	(-) 8.413×10^{-4}
Independent2 Dataset	MCC	(-) 0.0045	(-) 0.0064	(-) 0.0170	(-) 0.0264
	ACC	(-) 0.0067	(-) 0.0092	(-) 0.0202	(-) 0.0322
	SE	(=) 0.0994	(=) 0.0810	(=) 0.1160	(=) 0.1247
	SP	(-) 0.0011	(-) 0.0018	(-) 0.0202	(-) 0.0232

Here, (-) implies that the second algorithm is statistically better than the first one, (=) means that the two algorithms have no significant differences between them, and the p-value are given.

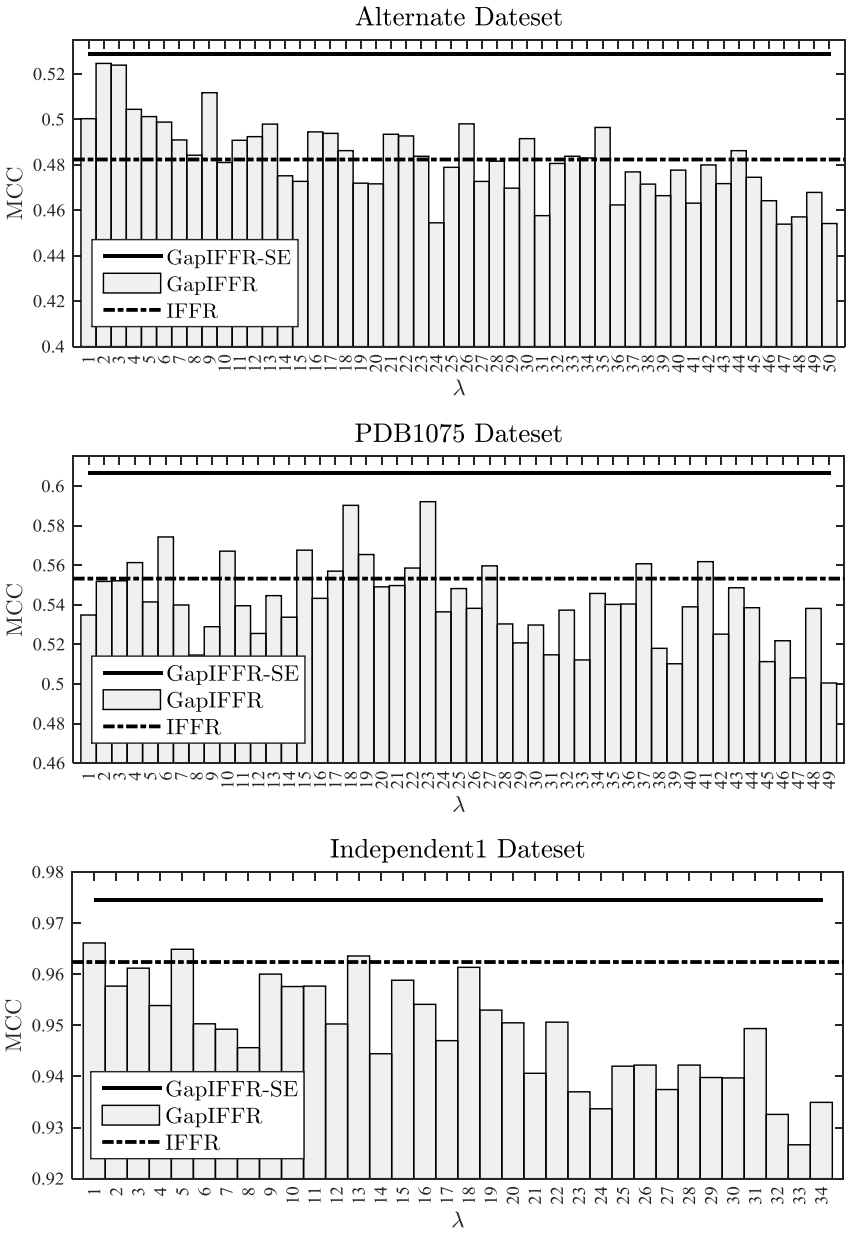
表 4 给出比较结果，对于算法 GapPSSM(GapCFFR)和 GapIFFR，除在数据集 Independent2 Dataset 的 SE 指标以外，全部数据集的 4 个性能指标的配对检验的 p-value 均小于显著水平 0.05，说明两算法之间的预测性能存在统计意义上的差异，也即在 4 个数据中，特征表示算法 GapIFFR 的预测性能显著地高于 GapPSSM 和 GapCFFR 的预测性能；在数据集 Independent2 Dataset 的 SE 指标上两算法不存在统计意义上的差异。因此，在实验中基于 GapIFFR 的特征表示显著优于其它两算法。可以认为在交互融合特征表示模型框架下，三信息融合 GapIFFR 显著地优于二信息融合 GapPSSM。同时，交互式融合特征表示也显著地优于组合式融合特征表示。以上实验也就验证了假设 2 的论断。

3.4 基于参数扰动的选择性集成的评估

本节实验主要讨论基于不同跳空距离λ的选择性集成问题，也即对参数λ进行扰动，生成不同的输入特征空间，以构建具有差异性的基分类器，从而提升整

体学习器的泛化能力。同样为保证实验结果的可比较性，本案例仍使用 Jackknife 校验法，假设蛋白质数据集有 N 条序列，把其中每条序列依次作为待测样本，剩余 N-1 条蛋白质序列采用 K-fold 交叉校验法(本案例 K=10)，其中(K-1)拆做为训练集((K-1)/K×(N−1)个样本)用于训练模型，1 拆做为验证集(1/K×(N−1)个样本)用于选择(剪枝)以确定集成学习器的结构。

为节省篇幅，以下仅选取更能反映学习器泛化性能的 MCC 指标进行比较，其它指标可做类似分析。在 4 个数据集上进行实验，考查所提选择性集成算法 GapIFFR-SE(k=3)与算法 IFFR(当做基准算法)以及算法 GapIFFR 的性能比较。结果如下图。



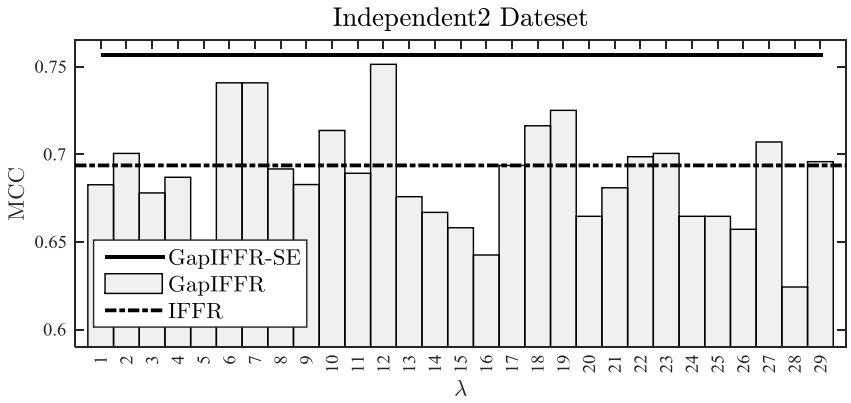


图 4 三信息融合特征表示 GapIFFR 的参数 λ 对 MCC 指标的影响(柱状图)，其中以二信息融合特征表示算法 IFFR 为基准(黑虚线)，并和选择性集成 GapIFFR-SE(黑实线)进行比较 (Jackknife test)

Note: IFFR means that the model integrates the evolutionary- and sequence-based dual information; λ-gapIFFR means that the model integrates the evolutionary-, physicochemical- and sequence-based triple information;

以算法IFFR的MCC指标值做为基准线(黑虚线)进行比较，从图4容易看出，在每个数据集上，柱状图的部分柱子穿过黑虚线出现在黑虚线上方，这就是说算法GapIFFR总存在个别的参数λ，它们的MCC指标值能够超过基准线(黑虚线)。同时，参数λ对算法GapIFFR的性能指标MCC的影响非常大，表现为柱状图柱子间高度差异较大，这表明不同参数λ会产生具有差异性的基分类器。

选择性集成GapIFFR-SE的MCC指标值(黑实线)更是大幅度的超越特征表示算法IFFR对应的基准线。从图4可以看出，黑实线在全部柱状图的上方，也就是选择性集成GapIFFR-SE所对应的MCC指标值均超越了算法GapIFFR的最大值，说明选择性集成算法GapIFFR-SE能够显著提升集成分类器的性能。因此，对参数λ进行扰动的选择性集成方法，一方面能够保证基分类器具有较高的准确性，另一方面能够保证基分类器之间存在差异性，进而达到提升整体学习器的泛化性能的目的。

3.5 与现有方法的进一步比较

对特征表示进一步做比较，以下采用原始文献中训练集与测试集的固定分割，也即使用 HoldOut 方法评估特征表示模型的性能，即在给定的训练集上生成特征，在所生成特征空间上训练分类器；然后，在给定的测试集的相应特征空间上验证分类模型；最后，分类器的识别率被用于间接评估不同的特征表示的性能。表 5 给出所提模型框架下的不同特征表示方法(CovPCSM, CovPSSM, CFFR 和

IFFR)和 3 个经典特征表示算法(PsePSSM, PseAAC 和 AAC)在测试集上的实验结果。从表 5 容易看出，在三个综合指标 ACC、MCC 和 AUC 上，CFFR 取得最大值，分别为 87.50%，0.7562 和 0.9297。然而，本案例更加感兴趣的是正例(DNA 结合蛋白)，从混淆矩阵和指标 SE(Sensitivity)容易看出，算法 PsePSSM 和 IFFR 的 SE 指标值均超过 80%，它们分别识别出 77 个正例和 74 个正例。同时，本案例也注意到分类器(LinearSVC)所使用的支持向量(nSV)数目，较少的 nSV 表明分类模型有更强的泛化能力。因此，综合比较容易看出，基于 IFFR 的特征表示有更好的表现，综合指标 MCC 达到 0.7204，正例识别 74 个，分类模型使用了最少的支持向量。这在一定程度上也说明基于 IFFR 特征表示的有效性。相比较，特征表示 PsePSSM 的敏感性指标达到最大值(SE=83.70%)，但是特异性指标较低，且综合指标 MCC 也不理想。这种现象也符合前面图 2 的结果。

Table 5
Performance comparisons with different methods on Testing dataset containing 92 DNAbinder proteins and 100 non DNAbinder proteins.

Method	Evaluation indices: Confusion Matrix	SE (%)	SP (%)	ACC (%)	MCC	AUC	nSV
AAC (d=20)	63 | 29 / 21 | 79	68.48	79.00	73.96	0.4781	0.7765	(87,90)
PseAAC (d=420)	70 | 22 / 14 | 86	76.09	86.00	81.25	0.6252	0.8843	(99,102)
PsePSSM (Ref. λ=0~24)	77 | 15 / 18 | 82	83.70	82.00	82.81	0.6564	0.8935	(89,98)
CovPCSM	57 | 35 / 2 | 98	61.96	98.00	80.73	0.6492	0.9104	(127,127)
CovPSSM	71 | 21 / 7 | 93	77.17	93.00	85.42	0.7138	0.9218	(74,78)
CFFR	73 | 19 / 5 | 95	79.35	95.00	87.50	0.7562	0.9297	(72,79)
IFFR	74 | 18 / 9 | 91	80.43	91.00	85.94	0.7204	0.9230	(69,77)
	a | *b* / *c* | *d*	*a*=True positive; *b*=False negative (Type II error); *c*=False positive (Type I error); *d*=True negative.					

对预测方法进一步做比较，以下在基准数据集 PDB1075 上，对所提预测方法选择性集成 GapIFFR-SE 和其它预测方法进行比较，其中用于比较的 8 个卓越方法包括有：iDNA-Prot|dis, PseDNA-Pro, iDNA-Prot, DNA-Prot, DNAbinder, iDNAPri-PseAAC, Kmer1+AAC 和 Local-DPP。基于 Jackknife 校验的比较结果如表 6 所示，容易看出，在众多的比较方法中，本案例的择性集成算法 GapIFFR-SE 具有最好的预测性能，也即识别率达到最大值 79.91%，MCC 指标取得最大值 0.61，SE 指标也取得最大值 87.43。因此，相比较于现有

的其它方法，所提方法具有更加卓越的性能，这也间接表明本案例所提交互融合特征表示能够生成携带有强判别信息的特征，同时选择性集成还能进一步提升整体学习器的泛化能力，最终能够保证对 DNA 结合蛋白的准确预测。

Table 6
Results of the proposed method and state-of-the-art predictors on the dataset PDB1075 (Jackknife test).

Methods	Evaluation indices			
	ACC (%)	MCC	SE (%)	SP (%)
iDNA-Prot|dis (Liu, 2014)	77.30	0.54	79.40	75.27
PseDNA-Pro (Liu, 2015a)	76.55	0.53	79.61	73.63
iDNA-Prot (Lin, 2011)	75.40	0.50	83.81	64.73
DNA-Prot (Kumar, 2009)	72.55	0.44	82.67	59.76
DNAbinder (dimension=400) (Kumar, 2007)	73.58	0.47	66.47	**80.36**
DNAbinder (dimension=21) (Kumar, 2007)	73.95	0.48	68.57	79.09
iDNAPro-PseAAC (Liu, 2015b)	76.56	0.53	75.62	77.45
Kmer1+AAC (Dong, 2015)	75.23	0.50	76.76	73.76
Local-DPP(n=3, lambda=1) (Wei, 2017)	79.10	0.59	84.80	73.60
Local-DPP(n=2, lambda=2) (Wei, 2017)	79.20	0.59	84.00	74.50
The proposed method	**79.91**	**0.61**	**87.43**	72.73

4 案例小结

从蛋白质序列(一级结构)出发，利用机器学习方法对蛋白质的结构和功能进行预测，是目前生物信息学研究的热点问题，也是一种重要研究手段。如何从序列数据中充分且有效地表达特征信息，是目前关注的焦点之一。针对蛋白质序列，通常是考虑氨基酸组成、多肽(相邻残基)组成、伪氨基酸(非相邻残基)组成，以及不同物化属性和进化信息等，生成显式的特征，并将这些特征向量进行(串联)组合，这类组合式融合特征表示 CFFR 能够取得较理想的效果。

本案例提出多信息交互式融合的特征表示算法，其实质是考虑蛋白质序列的不同物化属性和进化信息之间存在交互效应，以及考虑蛋白质序列中不同氨基酸残基间的相互作用。实验表明，从交互作用的视角，对不种物化属性、进化信息和非相邻残基间的作用信息，进行特征级融合，可以显著提高 DNA 结合蛋白的预测性能，这表明本案例的特征表示能够充分挖掘隐藏在蛋白质序列背后的潜在信息，所生成特征向量能够更好的识别和理解 DNA 结合蛋白。进一步，对特征表示算法参数进行扰动，生成不同的输入特征空间，选择性集成算法通过选择(或修剪)得到具有差异性的基分类器，提升整体分类器的泛化能力。本案例所提方法可以应用于蛋白质的结构与功能预测的其它相关领域，对辅助分析蛋白质序列

信息及其前沿问题的理解，有着信息学与生物学意义。

在本案例中为使比较结果具有可信性，仅使用相关文献所列举的 6 组物化指数进行分析，尽管所使用的 6 组物化指数，在一定程度上能够反应氨基酸的性质，例如其中亲水性和疏水性在蛋白质高级结构形成过程中起着极其重要的作用，当某个亲水性残基变成疏水性残基就可能使该蛋白质功能丧失。但是仅使用这 6 组物化指数还不够充分，考虑使用其它物化指数，从 AAindex 数据库中选取更加有效的物化指数进行分析，以便更大程度提高识别效果，是后继可以深入探讨的工作。

附录 1

首先，下载一个去冗余的蛋白质序列数据集 nr (ftp://ftp.ncbi.nlm.nih.gov/blast/db/)；

其次，运行 PSI-BLAST，输入待测蛋白质序列，设置参数 E 值为 0.001(evalue=0.001)和三次循环(num_iterations=3)，比对该去除冗余蛋白质序列的数据集 nr；

最后，得到待测蛋白质序列上每一个蛋白质残基关于 20 种氨基酸的得分，即 PSSM 矩阵。

附录 2

Table2.
The values of the six physicochemical properties for each amino acid.

Amino Acid		Physicochemical Index					
		Hydrophobicity	Hydrophilicity	Mass	pKa1(-COOH)	pKa2(-NH2)	pI(25°C)
		$Q^{(1)}$	$Q^{(2)}$	$Q^{(3)}$	$Q^{(4)}$	$Q^{(5)}$	$Q^{(6)}$
A	Ala	0.62	-0.5	15	2.35	9.87	6.11
C	Cys	0.29	-1.0	47	1.71	10.78	5.02
D	Asp	-0.9	3.0	59	1.88	9.60	2.98
E	Glu	-0.74	3.0	73	2.19	9.67	3.08
F	Phe	1.19	-2.5	91	2.58	9.24	5.91
G	Gly	0.48	0.0	1	2.34	9.60	6.06
H	His	-0.40	-0.5	82	1.78	8.97	7.64
I	Ile	1.38	-1.8	57	2.32	9.76	6.04
K	Lys	-1.50	3.0	73	2.20	8.90	9.47
L	Leu	1.06	-1.8	57	2.36	9.60	6.04
M	Met	0.64	-1.3	75	2.28	9.21	5.74
N	Asn	-0.78	0.2	58	2.18	9.09	10.76

P	Pro	0.12	0.0	42	1.99	10.60	6.30
Q	Gln	-0.85	0.2	72	2.17	9.13	5.65
R	Arg	-2.53	3.0	101	2.18	9.09	10.76
S	Ser	-0.18	0.3	31	2.21	9.15	5.68
T	Thr	-0.05	-0.4	45	2.15	9.12	5.60
V	Val	1.08	-1.5	43	2.29	9.74	6.02
W	Trp	0.81	-3.4	130	2.38	9.39	5.88
Y	Tyr	0.26	-2.3	107	2.20	9.11	5.63

参考文献

[Ptashne, 2005] Ptashne M. Regulation of transcription: From Lambda to eukaryotes [J]. Trends Biochem Sci, 2005, 30: 275-279.

[Jones, 1987] Jones, K. A., Kadonaga, J. T., Rosenfeld, P. J., Kelly, T. J., & Tjian, R. (1987). A cellular DNA-binding protein that activates eukaryotic transcription and DNA replication. Cell, 48(1), 79-89.

传统生物实验技术

[Cajone, 1989]Cajone F, Salina M, Benelli-Zazzera A. 4-hydroxynonenal induces a DNA-binding protein similar to the heat-shock factor [J]. Biochem, 1989, 262; 977-979.

[Buck, 2004] Buck MJ, Lieb J D. CHlP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments [J]. Genomics, 2004,83:349-360.

[Chou, 2003] Chou C C, Lin T W, Chen C Y, et al. Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso 10b2 at a resolution of 1.85 angstroms [J]. Bacteriol, 2003, 185: 4066-4073.

基于蛋白质结构的预测

[Zhang, 2010] Zhang H, Yang Y, Zhou Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function [J]. Bioinformatics, 2010,26: 1857-1863.

[Tjong, 2007] Tjong H, Zhou H X. DISPLAR; an accurate method for prediction DNA-binding sites on protein surfaces [J]. Nucleic Acids Res, 2007,35: 1465-1477.

基于蛋白质序列的预测

[Robert, 2010]Robert E L,Hui L. Boosting the prediction and understanding of DNA binding domains from sequence [J]. Nucleic Acids Res, 2010, 38:3149-3185.

[Huang, 2011]Huang H L, Lin I C, Liou Y F, et al. Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties [J]. BMC Bioinformatics,2011, 12: S 47.

[Kumar, 2007]Kumar M, Gromiha M,Raghava G. Identification of DNA-binding proteins using support vector machines and evolutionary profiles [J]. BMC Bioinforma,2007, 8: 463.

[Shao, 2009] Shao X Y, Tian Y J, Wu L Y, et al. Prediction DNA- and RNA-binding proteins from sequences with kernel methods [J]. J Theor Biol, 2009, 258: 289-293.

[Lin, 2011]Lin W Z, Fang J A, Xiao X,et al. iDNA-prot: identification of DNA-binding proteins using random forest with grey model [J]. Plos One, 2011, 6: e24756.

[Cai, 2004]Cai Y D, Doig A J. Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition [J]. Bioinformatics, 2004, 20: 1292-1300.

[Szilágyi, 2006] A. Szilágyi , J. Skolnick , Efficient prediction of nucleic acid binding function from low-resolution protein structures, J. Mol. Biol. 358 (2006) 922–933 .

[Yu, 2006] X. Yu , J. Cao , Y. Cai , T. Shi , Y. Li , Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines, J. Theor. Biol. 240 (2006) 175–184 .

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

[Chou, 2001] Chou K C. Prediction Of Protein Cellular Attributes Using Pseudo-amino Acid Composition.[J]. Proteins, 2001, 43(3): 246–55.

[Chou, 2005] Chou K C. Using Amphiphilic Pseudo Amino Acid Composition To Predict Enzyme Subfamily Classes[J]. Bioinformatics, 2005, 21(1): 10–19.

[Chou, 2001] Chou K C, Prediction of protein cellular attributes using pseudo amino acid composition [J]. Proteins Struct Funct Genet, 2001,43: 246-255.

[Ho, 2007] S.-Y. Ho , F.-C. Yu , C.-Y. Chang , H.-L. Huang ,Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method, Biosystems 90 (2007) 234–241 .

[Xu, 2015] R. Xu , J. Zhou , B. Liu , Y. He , Q. Zou , X. Wang , K.-C. Chou , Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. 33 (2015) 1720–1730 .

[Zhang, 2016] Zhang, J., Gao, B., Chai, H., Ma, Z., & Yang, G. (2016). Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm. BMC bioinformatics, 17(1), 323.

[Li, 2014]Li, L., Cui, X., Yu, S., Zhang, Y., Luo, Z., Yang, H., ... & Zheng, X. (2014). PSSP-RFE: accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations. PLoS One, 9(3), e92863.
集成学习
[Zhou, 2009] Zhou Z-H. Ensemble Learning[J]. Encyclopedia of Biometrics, Springer US, 2009: 270–273.
[Zhou, 2002]Zhou, Zhi-Hua, Jianxin Wu, and Wei Tang. "Ensembling neural networks: many could be better than all." Artificial intelligence 137.1-2 (2002): 239-263.
[Tsoumakas, 2008]Tsoumakas G, Partalas I, Vlahavas I (2008) A taxonomy and short review of ensemble selection. In: ECAI 2008, workshop on supervised and unsupervised ensemble methods and their applications
[Martínez, 2009]Martínez-Muñoz, G., Hernández-Lobato, D., & Suárez, A. (2009). An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2), 245-259.
[Rokach, 2010] Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33(1), 1-39.
[Zhang, 2009] Zhang C X, Zhang J S, Zhang G Y. Using boosting to prune double-bagging ensembles[J]. Computational Statistics & Data Analysis, 2009, 53(4): 1218-1231.
数据集
[Kumar, 2007]Kumar M, Gromiha M M, Raghava G P S. Identification of DNA-binding proteins using support vector machines and evolutionary profiles[J]. BMC bioinformatics, 2007, 8(1): 463.
[Stawiski, 2003]Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. J Mol Biol. 2003;326:1065–1079. doi: 10.1016/S0022-2836(03)00031-7.
[Wang and Brown, 2006] Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 2006;34:W243–W248. doi: 10.1093/nar/gkl298.
[Liu, 2015]Liu, B., Wang, S., & Wang, X. (2015). DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. Scientific reports, 5, 15479.
[Kumar, 2009]Kumar, K. K., Pugalenthi, G., & 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(6), 679-686.
[Liu, 2014] B. Liu , J. Xu , X. Lan , R. Xu , J. Zhou , X. Wang , K.-C. 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 (2014) e106691 .
[Lin, 2011] W.-Z. Lin , J.-A. Fang , X. Xiao , K.-C. Chou , iDNA-Prot: identification of DNA binding proteins using random forest with grey model, PLoS One 6 (2011) e24756 .
[Kumar, 2009] K.K. Kumar , G. Pugalenthi , P. Suganthan , DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest, J. Biomol. Struct. Dyn. 26 (2009) 679–686 .
[Liu, 2015a] B. Liu , J. Xu , S. Fan , R. Xu , J. Zhou , X. Wang , PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation, Mol. Inf. 34 (2015) 8–17 .
[Kumar, 2007] M. Kumar , M.M. Gromiha , G.P. Raghava , Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioin- form. 8 (2007) 463 .
[Liu, 2015b] B. Liu , S. Wang , X. Wang , DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation, Scient. Rep. 5 (2015) 15479 .
[Dong, 2015] Q. Dong , S. Wang , K. Wang , X. Liu , B. Liu , Identification of DNA-binding proteins by auto-cross covariance transformation, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE, 2015, pp. 470–475 .
[Wei, 2017] Wei L, Tang J, Zou Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information[J]. Information Sciences, 2017, 384: 135-144.
[Altschul, 1997]Altschul S. F., Madden T. L., Schaier A.A., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[C]. Nucleic Acids Res. 1997 Sep 1; 25(17) pp:3389-402

[Schäffer, 2001]Schäffer A A, Aravind L, Madden T L, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements[J]. Nucleic acids research, 2001, 29(14): 2994-3005.
[Moreira M, 1998] Moreira, M., & Mayoraz, E. (1998). Improved pairwise coupling classification with correcting classifiers. Machine Learning: ECML-98, 160-171.
[Kawashima, 2008]Kawashima S, Polarowski P,Pokarowska M, et al. AAindex: amino acid index database, progress report 2008 [DB]. Nucleic acids research, 2008,36: D202-D205.
[Shen, 2008] Shen, H. B., & Chou, K. C. (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical biochemistry, 373(2), 386-388.
(6 个特化属性)
Cortes C, Jackel L D, Chiang W P. Limits on learning machine accuracy imposed by data quality[C]//Advances in Neural Information Processing Systems. 1995: 239-246.