《数据建模与计算案例集》

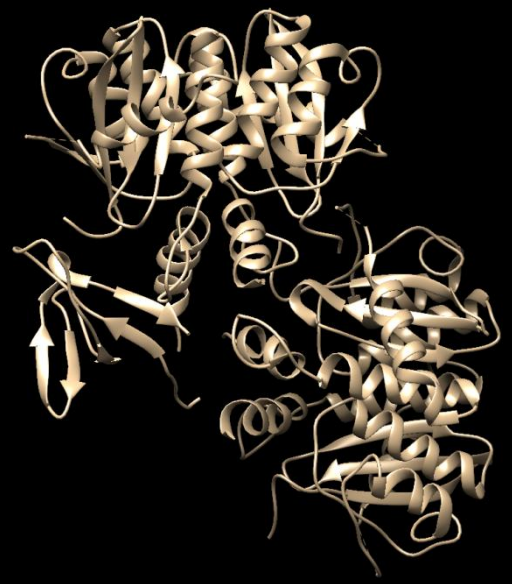
生物序列数据-蛋白质功能鉴定

交互融合特征表示与选择性集成的DNA结合蛋白质预测[1]

福建技术师范学院　游文杰

[1] “Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble,” *Knowledge-based Systems*, 2019, 163: 598-610.

案例特点



数据类型：生物序列

>1C6VX |
1QQSKNSKFKNFRVYYREGRDQLWKGPGELLWKGEGAVLLKVGTDIKVVPRRKAKIIKDYGGGKEVDSSSHMEDTGEAREVA
>4E0IA |
2MKAIDKMTDNPPQEGLSGRKIIYDEDGKPSRSCNTLLDFQYVTGKISNGLKNLSSNGKLAGTGALTGEASELMPGSRTYRKVDPPD
VEQLGRSSWTLLHSVAASYPAQPTDQQKGEMKQFLNIFSHIYPCNWSAKDFEKYIRENAPQVESREELGRWMCEAHNKVNKKLRKP
KFDCNFWEKRWKDGWDE
蛋白质一级结构(Protein primary structure)是肽或蛋白质中氨基酸的线性序列。按照惯例，蛋白质的一级结构被报道从氨基末端(N)端到羧基末端(C)端。蛋白质一级结构可以被直接蛋白质序列测序，或从DNA序列推断。

计算类型：密集计算

blast本地数据库的构建
从NCBI中的ftp库下载库，如nr.gz为非冗余的数据库(2020年227G大小),
或者数据库uniref50(2020年17G大小)

学科知识:(高代)矩阵

特征表示模型：高等代数中的矩阵知识，矩阵计算、分块矩阵等；
机器学习集成：基分类器如SVM或FLD，集成方法及选择性集成等。

案例目标

建模能力

实际问题 => 提出假设 => 数学建模 => 数据结构 =>
算法设计 => 编程计算 => 结果分析(可视化等)

计算能力

编程能力是计算能力的核心，最关键是算法设计能力以及由算法编写代码的能力。

摘要

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。

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。

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。

1. 背景介绍

问题

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。随着测序技术的飞速发展，序列数据呈爆炸性增长，序列与它们已知的结构和功能间的鸿沟越来越大。

意义

急需一个有效且可靠的基于生物序列的计算方法，也就是从氨基酸序列出发直接预测和建模蛋白质结构及其生物学功能，这也是后基因组时代蛋白质组研究领域中极具挑战性的研究课题之一。

2. 案例内容

数据和特征决定
机器学习的上限

模型和算法能够
逼近前者的上限

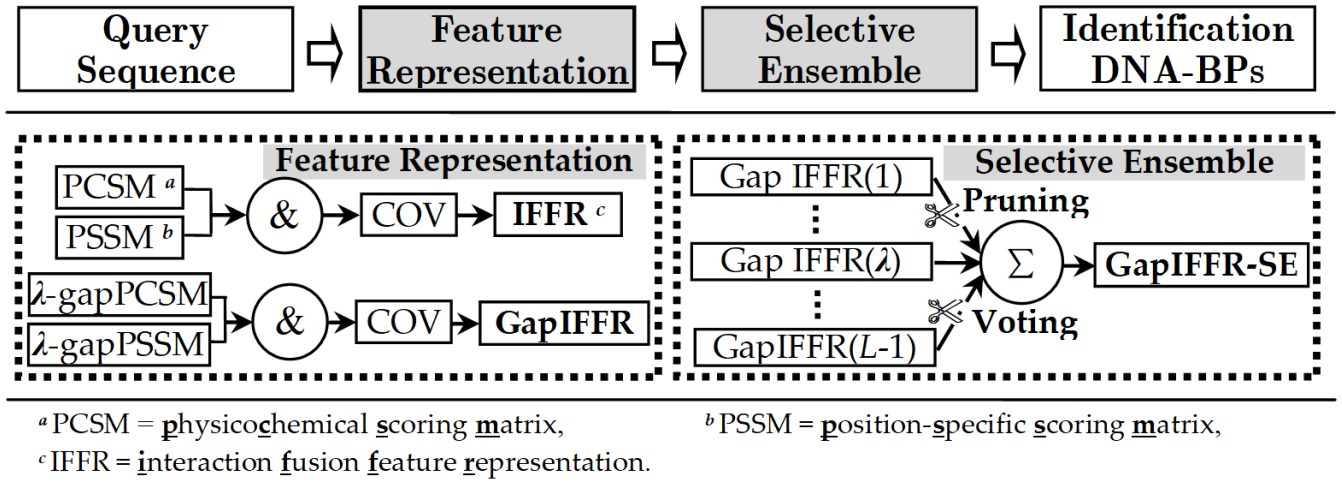


图1. DNA结合蛋白预测模型的框架图

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

从细胞中的生化反应来看，活体细胞中充满了各种的交互作用，如蛋白质之间的交互作用、氨基酸残基之间的着交互作用等。本案例将从具有交互效应的多信息融合的特征表示，考查不同属性(理化属性等)和信息(进化信息等)之间的交互效应。为此，提出科学问题的假设。

假设1： 不同的物化属性和进化信息之间存在交互效应

多源融合作为一种有效的信息处理技术，从信息论的角度来看，它能够(至少在理想情况下)提高对一个实体理解的特异性和全面性。为此，提出以下假设。

假设2： 在交互融合特征表示框架下，三信息融合优于二信息融合

2.2 模型

针对蛋白质序列的数据特点，本案例给出交互融合的特征表示模型，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)。

给出相关生物问题的数学描述：
定义1、定义2、定义3和定理1。

1. 定义1
(得分矩阵
Scoring Matrix)
2. 定义2
(λ-gap 得分矩阵
λ-gap Score Matrix)
3. 定义3
(得分协差阵
Score Covariance Matrix)
4. 定理1
(矩阵向量化
Matrix Vectorization)

2.2 模型

给出具有交互效应的多信息融合特征表示的数学模型。

所运用的知识点：矩阵、分块矩阵及其运算

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 λ-gap 得分矩阵(λ-gapSM)，即

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

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。

2.2 模型

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

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

2.3 算法

算法1(交互融合特征表示)

考虑物化属性和进化信息的交互效应，同时还考虑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**:

$$\boldsymbol{v} = (c_{1,1}, c_{2,1}, \cdots, c_{20+M,1}, c_{1,2}, c_{2,2}, \cdots, c_{20+M,2}, \cdots, c_{1,20+M}, c_{2,20+M}, \cdots, c_{20+M,20+M})$$

2.3 算法

算法2(选择性集成)

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，达到提升整体分类器的性能。详细算法如右：

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

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

Output: Y // class label of the test dataset S_{tst}.

(1) **Initialization process:**

—Set $T=\Phi$, L=minimum sequence length of S_{trn}, S_{val} and S_{tst}, calculate $D_{trn}(\lambda)$, $D_{val}(\lambda)$ and $D_{tst}(\lambda)$ by calling **GapIFFR** with $\lambda=\{1,2,\cdots,L\text{-}1\}$.

(2) **Training base classifiers process:**

—**For** i=1 to L-1 **do**

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

—**EndFor**

(3) **Selection (Pruning) process:**

—**For** j=1 to L-1 **do**

——Calculate M_j for each base classifier $C_j \in T$ on the validation dataset $D_{val}(j)$ using the evaluation criteria M.

—**EndFor**

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

(4) **Ensemble (Voting) process:**

—Predict the class label of the test dataset S_{tst},

$$Y = sign\{\sum_{t=1}^{k} C_{\lambda_t}(\mathbf{x})\}$$

where $C_{\lambda t}$ is the λ_t-th base predictor on the dataset $\mathbf{x} \in D_{tst}(\lambda_t)$, $\{\lambda_1, \lambda_2, \cdots, \lambda_k\} \subset \{1,2,\cdots,L\text{-}1\}$.

Return Y

3. 实验

3.1 实验数据[1]与评价指标

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%

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

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

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

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

[1] 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

3. 实验

3.2 二信息交互融合特征表示的评估

比较并评估基于物化属性与进化信息的二重信息交互融合特征表示IFFR的性能。

Table 3
Prediction performance comparison 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**

3. 实验

3.2 二信息交互融合特征表示的评估

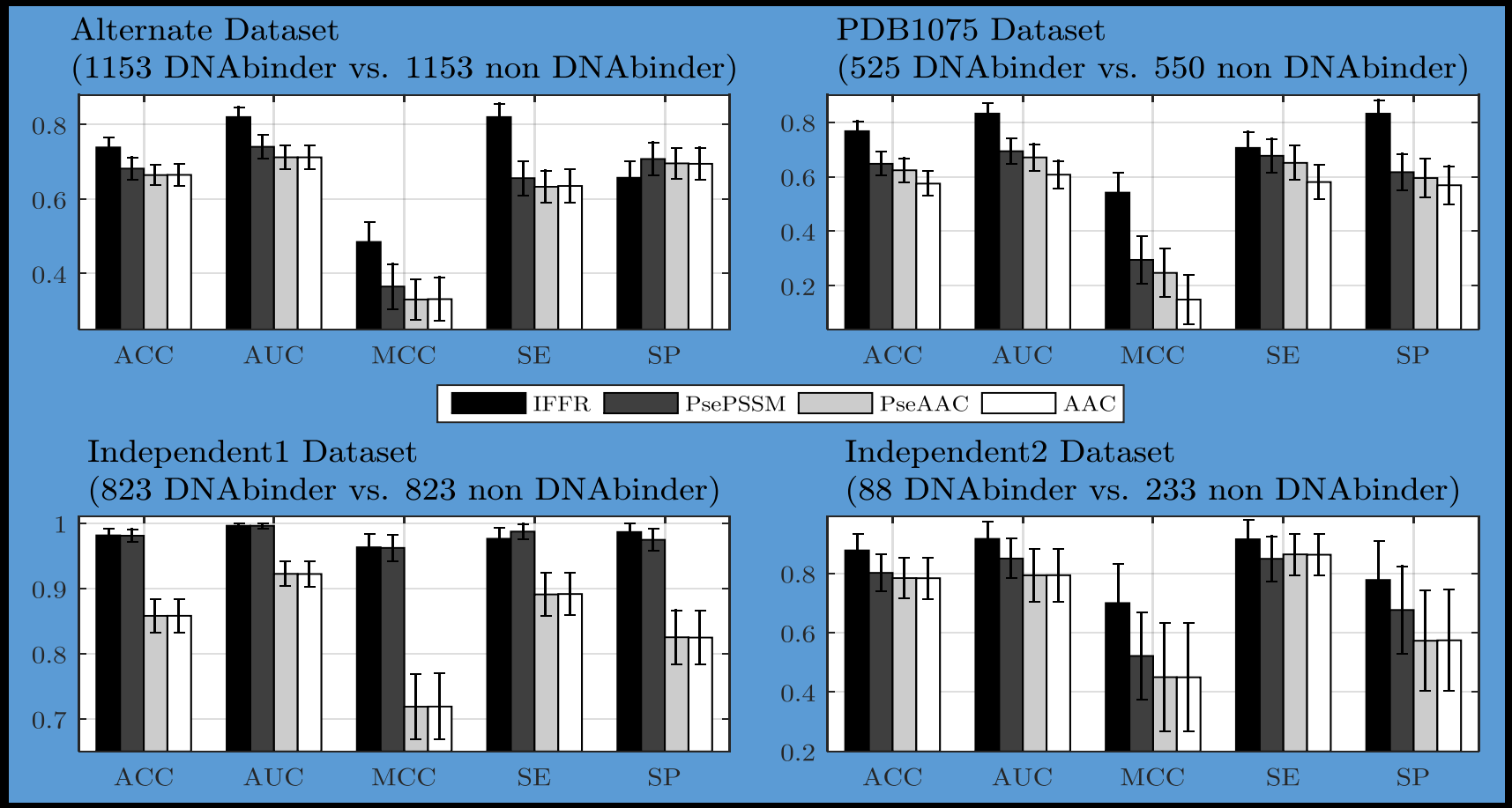


图2 不同特征方法的性能指标(Accuracy, AUC, MCC, Sensitivity and Specificity)比较(30次的10-fold CV).

实验结果表明：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。验证了假设1的结论。

3. 实验

3.3 参数敏感性分析与模型比较

参数选择问题，考查算法的参数 λ 的敏感性，也即不同的跳空距离对结果的影响。

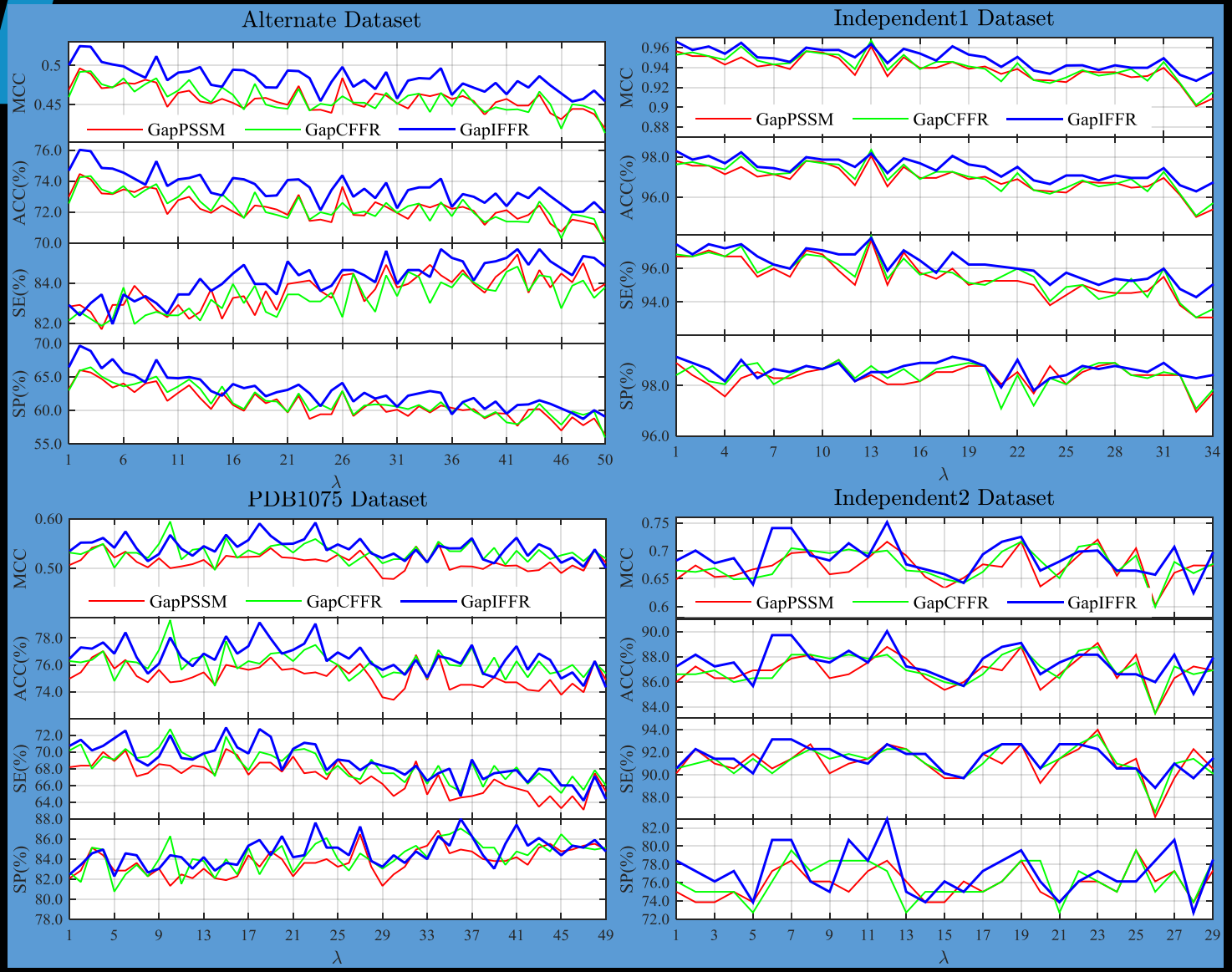


图3 多信息融合特征表示GapPSSM, GapCFFR和GapIFFR的性能(MCC, Accuracy, Sensitivity和Specificity)比较以及参数λ对结果的影响，(Jackknife validation test, base classifier: linear SVM)

3. 实验

3.4 基于参数扰动的选择性集成的评估

基于不同跳空距离λ的选择性集成，即对参数进行扰动，生成不同的输入特征空间，以构建具有差异性的基分类器，而提升整体学习器的泛化能力。

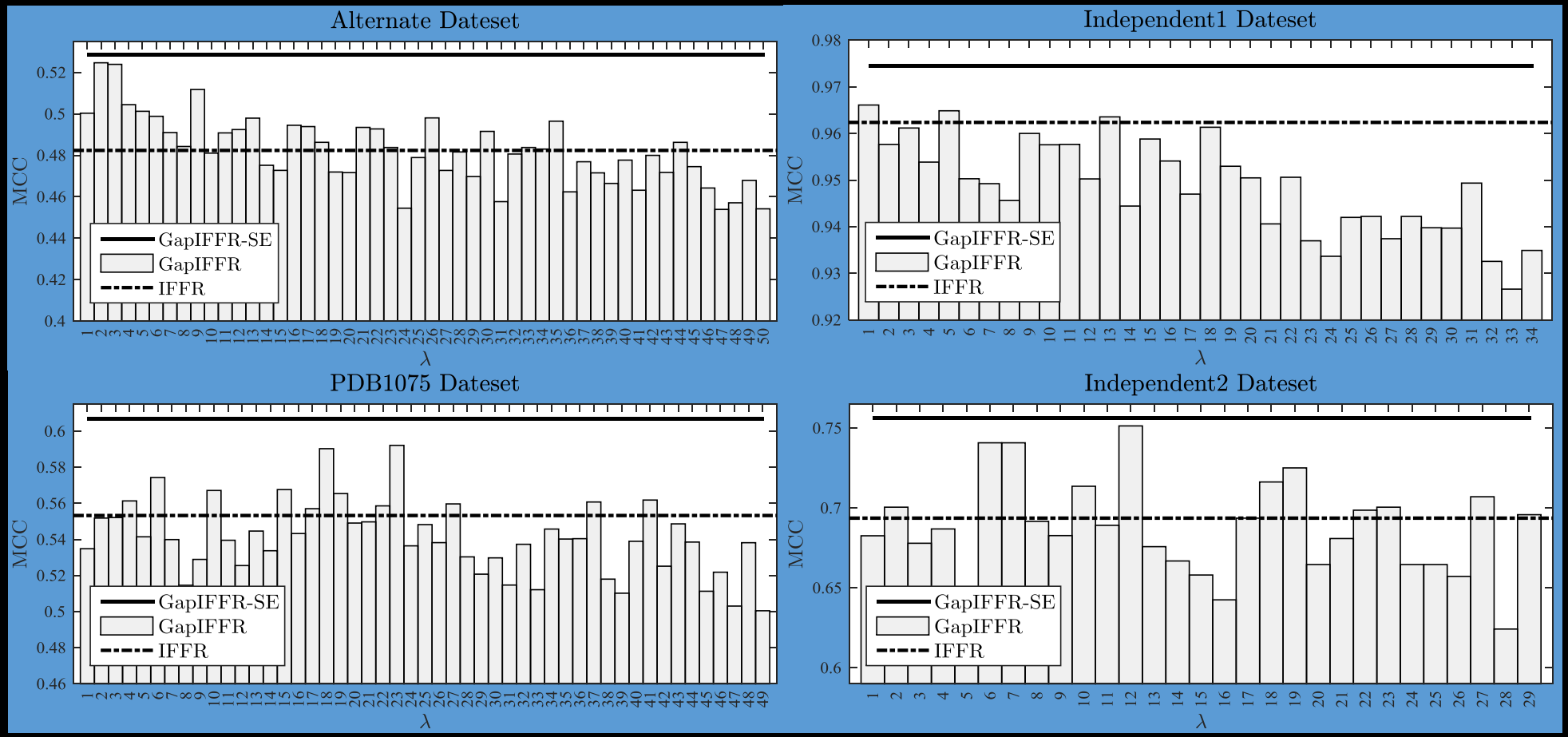


图4 三信息融合特征表示GapIFFR的参数λ对MCC指标的影响(柱状图)，其中以二信息融合特征表示算法IFFR为基准(黑虚线)，并和选择性集成GapIFFR-SE(黑实线)进行比较(Jackknife test)

3. 实验

3.5 与现有方法的进一步比较

在基准数据集PDB1075上，对选择性集成和其它预测方法进行比较。

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，本案例的特征表示模型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。

对算法参数进行扰动，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。

本案例的模型与算法可以应用于蛋白质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。

感谢

Many Thanks!