基于多特征融合预测蛋白质相互作用界面

陈心浩,胡 俭
(中南民族大学 生物医学工程学院,武汉 430074)

摘 要 为高效准确地预测蛋白质相互作用界面,提取了传统特征,并采用多种方法改进进化信息特征,利用特征选择构建了一个 14 维的预测模型. 通过 5 折交叉验证和独立测试,预测结果表明: 该预测模型不仅显著降低特征维度,而且选择的特征组合具有较好的预测能力和较强的泛化能力.

关键词 蛋白质-蛋白质界面; 分类; 进化; 特征选择
中图分类号 Q811.4 **文献标识码** A **文章编号** 1672-4321(2017)03-0080-04

Study on Protein-Protein Interfacial Classification Based on Multi -feature Fusion

Chen Xinhao, Hu Jian
(College of Biomedical Engineering, South-Central University for Nationalities, Wuhan 430074, China)

Abstract To build a model of efficient and accurate classification of protein-protein interfaces, this study constructs two characteristics of traditional features and evolutionary information, a 14-dimensional feature model is constructed by feature selection. By cross-validation of the main data set and independent test set testing, results show that selects the features combination has better predictive ability and strong extension ability. Compared with the best models at the present stage, this study significantly reduce the dimensionality of the model case classification has improved.

Keywords protein-protein interface; classification; evolutionary; feature selection

区分蛋白质晶体中的生物学相互作用界面(Biological interfaces)和无生物学意义的晶体学界面(Crystal interfaces),是结构生物信息学中的一个重要研究方向.

现有计算方法预测蛋白质相互作用界面的特征主要分成两大类: 第一类是以界面面积、疏水性和温度因子等几何特性和氨基酸理化特性为代表的传统特征[1]; 第二类则是以 EPPIC 方法为代表的进化特征[2]. 为获得良好的分类效果,目前的主要策略是将上述特征进行联合. 然而,这类融合方法也存在弊端,如现阶段分类效果最好的 Luo 方法[3],该方法具有较高的特征维度(46 维),且进化信息计算复杂,不利于快速构建本地分类模型. 因此,本文期望采用较为简便的方式计算进化特征,融合传统特征并使用特征选择技术,构建一个低维高效的蛋白质互作界面分类模型.

1 材料和方法

1.1 数据集,蛋白质界面残基、表面残基的定义

在构建和测试模型过程中使用了三个数据集,Duarte 数据集[2]作为主数据集用于构建模型和优化参数,Bernauer[4]和 Ponstingl[5]两个经典数据集作为独立测试集.

核心残基(Core)位于互作界面中心,主要由疏水性氨基酸构成. 核心残基周围环绕着一圈残基,此类型残基称之为环绕残基(Rim). 界面残基、表面残基、核心残基与环绕残基定义采用 Proface 方式定义[6].

1.2 传统特征

核心残基(Core)与环绕残基(Rim): 分别计算核心残基与环绕残基在界面残基中的比例,构成 Core 和 Rim 这两个特征. 核心残基数目(NoC): 每个蛋白质复合体的核心残基数构成本特征. 温度因子(BF):

收稿日期 2017-03-30
作者简介 陈心浩(1968-),男,副教授,研究方向: 医学图像处理与传输,E-mail: xinhaochen@mail.scuec.edu.cn
基金项目 国家自然科学基金资助项目(61002046); 中央高校基本科研业务专项基金项目(CZP17025)

将 PDB 中每个残基温度因子做 Z-score 归一化,将归一化后的界面残基温度因子平均值作为此蛋白质复合体温度因子. 局部包装密度(LD) ,热点残基数目(Nhs) ,氨基酸分布 (RP) 定义方式来自 Proface[6] ,界面疏水性 Hy) 采用 Jones 定义[1] .

1.3　进化特征

本文采用默认参数,使用 PSI-BLAST 程序对目标蛋白质在 NR 数据库中搜索其同源序列并构建位置特异性矩阵. 根据上述矩阵,采用 Capra 方法[7] ,对每一个残基位置分别计算了 SE(Shannon entropy of residues) , SERP (Shannon entropy of residue properties) ,VNE(von Neumann entropy) ,RE(Relative Entropy) 和 JSD(Jensen-Shannon divergence score) 5 种保守性分值,并且对于计算出来的 5 种保守性分值采用 3 窗口平均,构成另外 5 个保守性分值. 计算公式如下:

$$SE_i = -\sum_{\alpha \in AA} p(\alpha) \lg [p(\alpha)] , \quad (1)$$

$$SERP_i = -\sum_{\alpha \in Term)} p(\beta) \lg [p(\beta)] , \quad (2)$$

$$VNE_i = -Tr(\rho \lg(\rho)) ,$$

$$\rho = \mathrm{diag}([p1, p2, \cdots, p20) \cdot BLUSUM62] , \quad (3)$$

$$RE_i = -\sum_{\alpha \in AA} p(\alpha) \lg [p(\alpha)/q(\alpha)] , \quad (4)$$

$$JSD_i = \lambda \sum_{\alpha \in AA} p(\alpha) \lg [p(\alpha)/r(\alpha)] + (1-\lambda) \sum_{\alpha \in AA} q(\alpha) \lg [q(\alpha)/r(\alpha)] , \quad (5)$$

$$WindowScore_i = 0.5 Entropy_i + 0.5 \frac{\sum_{j \in Window} Entropy_j)}{|Window|} , \quad (6)$$

公式(1) 中,$p(\alpha)$ 是 20 种常见氨基酸在位置 i 出现的概率,公式(2) 中的 $p(\beta)$ 则是根据 Mirny 研究[8] 对氨基酸根据化学属性分成 6 组,计算出的每一组在整体出现的概率,具体分组可见表 1. VNE 计算方法[9] 如公式(3) 所示,特点是将原始的概率得分使用 BLUSUM62 矩阵重新计算. RE 的计算方式与 SE 接近,不同点是使用背景概率 $q(\alpha)$ 重新定义,其概率分布见表 1. JSD 是将 RE 做了背景频率改进[10] ,可以将保守性分数归一化 0 ~ 1 之间,在本文中 $\lambda = 0.5$. 公式(6) 即 3 窗口的算法,序列上第 i 个残基与其邻近的两个残基加权平均. 将上述获得的 5 个保守性分值和 5 个窗口保守性分值分别作 Z-score 变换,以消除不同蛋白质复合体间差异.

蛋白质残基保守性分值可以衡量残基在进化过程中变异程度,生物学界面残基,特别是生物学界面上的核心残基在进化过程中相对保守. 本文采用两种方式计算核心残基保守性分值[2] ,第一种是核心残基-界面残基保守性分值 (CI) ,计算核心残基保守性分值平均值与界面残基保守性分值平均值的差值,即将界面残基保守性分值作为基准. 第二种是核心残基-表面残基保守性分值(CS) ,计算核心残基保守性分值平均值与表面残基保守性分值平均值的差值. 最终构成 20 维进化信息特征.

表 1　氨基酸属性

Tab. 1　Amino acid properties

氨基酸	氨基酸分组(Term)	氨基酸背景频率 $q(\alpha)$	氨基酸	氨基酸分组(Term)	氨基酸背景频率 $q(\alpha)$
Asp	带负电	0.052	Ser	极性	0.059
Ser	极性	0.059	Thr	极性	0.055
Thr	极性	0.055	Gly	特殊类型	0.083
Gly	特殊类型	0.083	Pro	特殊类型	0.043
Pro	特殊类型	0.043	Ala	脂肪族	0.078
Ala	脂肪族	0.078	Cys	脂肪族	0.024
Cys	脂肪族	0.024	Ile	脂肪族	0.062
Ile	脂肪族	0.062	Leu	脂肪族	0.092
Leu	脂肪族	0.092	Met	脂肪族	0.024
Met	脂肪族	0.024	Val	脂肪族	0.072

1.4　特征选择、分类器与分类评价

增 L 去 R 选择算法是一种改进了的前向特征选择方法[11] . 算法初始特征选择从空集开始,每轮先加入 L 维特征,然后从中除去 R 个特征,将每一轮 AUC 最高的特征组合挑选出来作为下一轮初始特征组合.

分类器采用 R 语言下随机森林包,所涉及参数均采用默认值.

对单个特征和联合特征测试均在 Duarte 数据集上完成,采用 5 折交叉验证. 为排除随机影响,5 折交叉验证采用 50 次独立分组取平均的结果,两个独立测试采用 50 次重复平均结果. 分类效果评价采用敏感度(SN) 、特异度(SP) 、准确性、马修相关性系数(MCC) 、受试者工作曲线(ROC) 及 ROC 曲线下面积(AUC) 6 个指标. *MCC* 范围是 [-1, 1] ,当 *MCC* 大于 0 代表正确的分类效果,越接近 1 代表分类效果越好. 一般来说,当 *MCC* 大于 0.3 表示有一定分类效果,大约 0.5 时分类效果较好. *AUC* 也有类似的评价标准,当 *AUC* 处于 0.5 到 0.6 之间表示只有微弱的分类效果,当 *AUC* 大于 0.6 表示此特征有一定的区分样本能力,当 *AUC* 大约 0.8 表示分类效果很理想.

2　结果与讨论

2.1　特征分类效果

根据表 2,在传统特征中,*Hy*、*Core*、*Rim*、*RP* 和

Nhs 5 个特征的单独使用分类 *AUC* 均到达 0.7 以上，除 *Nhs* 每个特征的 *MCC* 都超过 0.4，显示出这些特征在生物学界面和晶体学界面上有较大的分布差异性. *BF* 和 *NoC* 的 *AUC* 处于 0.6 到 0.7 之间，*MCC* 大于 0.3，有一定分类效果. *LD* 分类效果较差，*AUC* 不到 0.6.

表2 特征独立使用分类效果

Tab.2 Independent feature classification results

特征	ACC	MCC	AUC	特征	ACC	MCC	AUC
BF	0.647	0.302	0.693	CI-VNE	0.574	0.239	0.553
LD	0.602	0.21	0.595	CI-VNE-3WIN	0.559	0.225	0.58
Hy	0.751	0.511	0.787	CI-SERP	0.649	0.316	0.677
Nhs	0.673	0.351	0.703	CI-SERP-3WIN	0.609	0.279	0.636
RP	0.683	0.445	0.732	CS-SE	0.672	0.36	0.708
Core	0.733	0.478	0.772	CS-SE-3WIN	0.61	0.239	0.646
Rim	0.732	0.478	0.772	CS-RE	0.651	0.313	0.656
NoC	0.632	0.302	0.65	CS-RE-3WIN	0.642	0.315	0.641
CI-SE	0.647	0.3	0.698	CS-JSD	0.591	0.265	0.664
CI-SE-3WIN	0.677	0.362	0.732	CS-JSD-3WIN	0.652	0.342	0.674
CI-RE	0.674	0.369	0.679	CS-VNE	0.514	0.022	0.448
CI-RE-3WIN	0.579	0.297	0.651	CS-VNE-3WIN	0.53	0.134	0.491
CI-JSD	0.662	0.326	0.676	CS-SERP	0.713	0.484	0.758
CI-JSD-3WIN	0.579	0.297	0.651	CS-SERP-3WIN	0.692	0.43	0.747

在进化信息特征中，并非所有的保守性分值算法都适合本问题，如 CS-VNE 的 *AUC* 小于 0.5，产生相反的分类效果. 若以 *AUC* 为评价准则，整体上来说，相同算法计算出的 *CS* 要略优于 *CI*，这与 Duarte 得出的结论相同. 在原始保守性分值与 3 维窗口计算出的保守性分数比较中，不同种算法产生了不同的效果，如 CI-SE-3WIN 相比 CI-SE 分类效果提升明显，而 CS-SE-3WIN 相比 CS-SE 分类效果却变差. 在 20 个进化特征中，CS-SERP *AUC* 达到 0.758，*MCC* 达到 0.484，是 28 个特征中分类效果最好的特征之一.

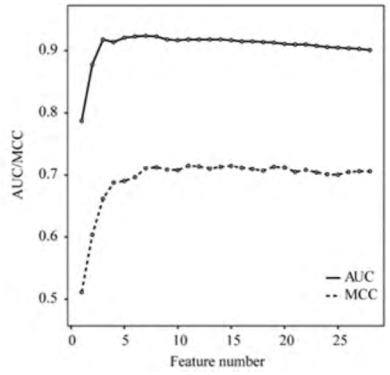

图1 特征选择

Fig.1 Feature selection

2.2 特征选择

以 *AUC* 为选择标准，本文采用增 2 去 1 选择算法对 28 个特征做特征选择. 对于每一轮选择出的特征组合，计算 *AUC* 和 *MCC*，绘制的曲线如图 1 所示，随着特征数目的增加，*AUC* 先快速上升，在第 8 轮特征选择后达到顶点，而后 *AUC* 缓慢下降；*MCC* 上升速度相比于 *AUC* 较慢，而且在达到顶点后并没有明显的下降趋势. 综合 *AUC* 和 *MCC* 分值，最终选择第 14 个特征组合，分别是 Hy、Core、CS-SERP-3WIN、CI-SE-3WIN、RP、Nhs、CI-SE、BF、CI-RE、CI-SERP、CI-JSD-3WIN、CI-RE-3WIN、CI-SERP-3WIN、LD. 选择出的 14 个特征 *AUC* 为 0.918，*MCC* 为 0.713，而全部 28 个特征 *AUC* 为 0.901，*MCC* 为 0.706，可见本文在消减了一半特征维度情况下，AUC 还是获得了较大程度提升，说明本文采用的特征选择确实可以在保证预测准确性条件下选择出更有意义的特征组合.

在特征选择中没有被选择出来的特征，其中 *Rim* 是因为与 *Core* 成对偶关系，所包含的信息是完全一致的；*NoC* 是因为在本文中多个特征涉及到核心残基，信息上存在冗余因而没有被选择出来. 信息冗余同样存在于 20 个进化信息特征上，因此只有 8 个进化信息特征被选择出来. 虽然 *CS* 单个特征效果略好，但是在选择出的 8 个进化信息特征中只有一个 *CS*，而独立使用 *LD* 分类效果较差却可以被选择出，说明并非联合较强特征一定会取得良好的分类效果，还需要考虑各个特征之间的组合效应.

2.3 交叉验证与独立测试效果

表 3 所示的是 Duarte 数据集 5 折交叉验证结果和两个独立测试集的分类效果，图 2 所示的是相应的 ROC 曲线. 可以看到，本文在 Duarte 数据集上取得了 *AUC* 为 0.918，*MCC* 为 0.713 这样良好的分类效果. 将本方法应用于两个独立测试集上，Bernauer 数据集 *AUC* 达到 0.955，*MCC* 达到 0.745 的 *MCC*，Ponstingl 数据集 *AUC* 为 0.962，*MCC* 为 0.842，均获得了良好的的分类效果，可见本方法有较强的泛化能力.

表 3　Duarte 数据集 5 折交叉验证和独立测试集预测效果

Tab. 3　Duarte 5-fold cross validation test and two independent prediction results

数据集	SN	SP	ACC	MCC	AUC
Duarte	0.893	0.814	0.853	0.713	0.918
Bernauer	0.927	0.863	0.881	0.745	0.955
Ponstingl	0.912	0.936	0.922	0.842	0.962

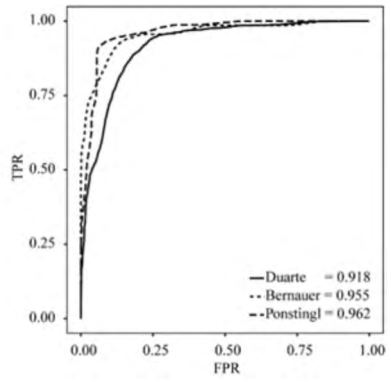

图 2　Duarte 数据集 5 折交叉验证和独立测试集 ROC 曲线

Fig2　The ROC curves of 5-fold cross validation test and two independent datasets

2.4　与现有方法比较

为更加全面地评价本方法，本文采用现阶段分类效果最好的两个分类器，即 Luo 方法和 EPPIC 方法对 Duarte 数据集做 5 折交叉验证，与本方法得到的结果进行比较. EPPIC 方法的预测效果直接取自文献报道; 对 Luo 使用的特征数据，采用与本文相同的 50 次 5 折交叉验证进行评价. 本方法与这两种方法比较见图 3，从对比结果上来看，除 *SN* 本方法与现有方法相仿之外，*SP*、*ACC* 和 *MCC* 本方法均有显著提升，采用符号秩和检验 *SP*、*ACC* 和 *MCC* 本方法差异达到 5.24E-10、3.01E-09、5.19E-09，可以得出本方法在 Duarte 数据集上分类效果优于上述两种方法的结论.

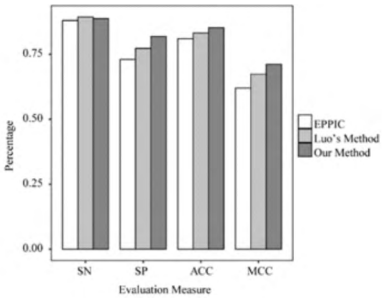

图 3　本文方法与 Luo 方法、EPPIC 比较

Fig. 3　Comparison of the performances of our method and Luo's Method and EPPIC

3　结语

本文提取了进化特征和传统特征，通过特征选择构建了一个高效的蛋白质相互作用界面分类模型. 交叉验证和独立测试的结果表明本方法可以达到较为理想的预测效果. 与现有方法相比，本方法大幅度降低了特征维度，却并没有降低分类效果. 然而也有不完善的地方，如备选特征数目较少，对特征的生物学意义挖掘不深等，这些问题将是作者下一步研究的重点.

参　考　文　献

[1] Jones S, JM Thornton. Analysis of protein-protein interaction sites using surface patches [J]. Journal of Molecular Biology, 1997, 272(1): 121-132.

[2] Duarte J M, Srebniak A, Scharer, M A, et al. Protein interface classification by evolutionary analysis [J]. BMC Bioinformatics, 2012, 13(1): 334-334.

[3] Luo J, Guo Y, Fu Y, et al. Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants [J]. Proteins, 2014, 82(11): 3090-3100.

[4] Bernauer J, Bahadur R P, Rodier, et al. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions [J]. Bioinformatics, 2008, 24(5): 652-658.

[5] Ponstingl H, Kabir T, Thornton J M. Automatic inference of protein quaternary structure from crystals [J]. Journal of Applied Crystallography, 2003, 36(5): 1116-1122.

[6] Saha R P, Bahadur R P, Pal A, et al. ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces [J]. BMC Struct Biol, 2006, 6: 11.

[7] Capra J A, Singh M. Predicting functionally important residues from sequence conservation [J]. Bioinformatics, 2007, 23(15): 1875-1882.

[8] Mirny L A, Shakhnovich E I. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function [J]. Journal of Molecular Biology, 1999, 291(1): 177-196.

[9] Caffrey D R, Somaroo S, Hughes J, et al. Are protein - protein interfaces more conserved in sequence than the rest of the protein surface [J]. Protein Science, 2004, 13(1): p. 190-202.

[10] Lin J, Divergence measures based on the Shannon entropy [J]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151.

[11] 姚　旭，王晓丹，张玉玺，等. 特征选择方法综述 [J]. 控制与决策，2012，27(2): 161-166.