

基于谷物蛋白质序列与 PPI 网络的功能预测研究

王钰¹, 刘静^{*1}, 管骁², 崔双龙¹, 汤杏华¹

(1. 上海海事大学 信息工程学院, 上海 201306; 2. 上海理工大学 健康科学与工程学院, 上海 200093)

摘要: 谷物中现存大量未经注释、功能未知的蛋白质, 且难以通过实验验证, 因此计算方法成为预测谷物蛋白质功能的主流方法之一。作者以玉米、小麦、粳稻、籼稻 4 种谷物蛋白质为研究对象, 利用数据库获取结构域相互作用信息。从蛋白质中较为稳定的结构域信息出发, 结合 AdaBoost 算法获得蛋白质相互作用信息并构建蛋白质相互作用网络, 将其与利用 blast 所获得的蛋白质序列相似性网络相结合, 利用协同分类和多层感知机两种算法实现对谷物蛋白质的功能预测。研究结果显示, 两种算法均能较为准确地预测蛋白质功能, 其中协同分类在召回率方面表现更优, 而多层感知机在准确率方面表现更优。本研究为谷物蛋白质的功能注释提供了新思路、新方法, 对谷物的加工与营养研究提供了依据。

关键词: 蛋白质相互作用网络; 结构域; 谷物; AdaBoost 算法; 协同分类; 多层感知机

中图分类号: TP391.4; S51 文章编号: 1673-1689(2023)04-0075-10 DOI: 10.3969/j.issn. 1673-1689.2023.04.009

Function Prediction Based on Grain Protein Sequence and PPI Network

WANG Yu¹, LIU Jing^{*1}, GUAN Xiao², CUI Shuanglong¹, TANG Xinghua¹

(1. School of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; 2. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: There are many unannotated proteins with unknown functions in cereals, which are difficult to be verified by experiments. However, computational methods have become one of the mainstream methods to evaluate the functions of the cereal proteins. In this study, maize, wheat, indica rice and japonica rice proteins were studied, and their structural domain interaction information was obtained from the related databases. The protein-protein interaction information was obtained and the protein-protein interaction network was constructed, starting from the relatively stable domain information of protein and combining the AdaBoost algorithm. Combining it with the biomolecular structure similarity network obtained by blast, the functions of the cereal proteins were predicted based on the cooperative classification and multi-layer perceptron algorithms. Results showed that both of cooperative classification algorithm and multi-layer perceptron algorithm could accurately predict the functions of the proteins. Moreover, collaborative classification algorithm showed a better recall rate, whereas multi-layer perceptron algorithm showed a better accuracy. This

收稿日期: 2021-05-31

基金项目: 上海市科委地方高校能力建设项目(20060502100); 上海市科技兴农项目(2021-02-08-00-12-F00780)。

* 通信作者: 刘静(1979—), 女, 博士, 副教授, 主要从事信息技术与食品安全交叉领域研究。E-mail: jingliu@shmtu.edu.cn

sudy revealed a new idea and method for the unannotated protein investigation, and provided a strong guarantee for cereal processing and nutrition research.

Keywords: protein-interaction network, structural domain, grain, AdaBoost algorithm, collaborative classification, multi-layer perceptron

谷物营养价值高,其蛋白质质量分数达8%~12%,是膳食中蛋白质的主要来源,研究谷物蛋白质对于食品工业的发展具有重要意义。

现有的谷物蛋白质数据中,存在着大量未经注释、功能未知的蛋白质,难以对其进行人工注释复核。因此,基于这些已测序的蛋白质数据,计算方法成为蛋白质功能预测的主流方法之一。

早期的研究主要基于蛋白质序列相似性进行功能预测,该方法认为序列水平上相似的蛋白质,其结构与功能基本相似,即利用蛋白质序列相似性实现了蛋白质家族划分^[1]。随着人们对蛋白质与蛋白质之间相互作用 (protein-protein interactions, PPI)关系研究的深入,发现当蛋白质之间发生了相互作用时,即可认为这两个蛋白质之间可能存在相同或相似的 GO 生物功能^[2]。Schwikowski 等人利用邻近的蛋白质已知功能预测未知蛋白质的功能^[3]。Vazquez 等人采用图分割理论与 PPI 网络拓扑特征,从图的角度出发预测蛋白质功能^[4]。Deng 等人通过构建马尔科夫随机场模型,使用逻辑回归法预测蛋白质的功能^[5]。Mei 等人基于图聚类算法分析蛋白质相互作用网络识别蛋白质功能模块^[6]。

结构域信息作为蛋白质中较为稳定的结构, Han 等人提出若两个蛋白质的结构域之间发生相互作用,则这两个蛋白质之间必然存在相互作用^[7-9]。在此基础上 Hayashid 等人使用条件随机场算法完成了从蛋白质结构域相互作用信息到蛋白质相互作用的预测^[10]。Singhal 等人利用已知的蛋白质相互作用信息,对参与相互作用的结构域进行评分,从而实现蛋白质相互作用的预测^[11]。

由于针对谷物蛋白质相互作用的研究较少,且难以获取足够的、可靠的相关数据,因此选择蛋白质中更加稳定且能够体现蛋白质功能的结构域信息作为实验数据。作者基于谷物蛋白质中结构域信息,使用集成学习方法得到谷物蛋白质相互作用关系,从而构建谷物蛋白质相互作用网络。作者将蛋白质相互作用网络与利用 blast 工具^[12]所获得的谷物蛋白质序列相似网络相结合,提出了基于协同分

类以及基于多层感知机的两种谷物蛋白质功能预测算法。

1 数据来源

1.1 数据与方法

选取4种谷物作为研究谷物蛋白质功能预测的数据信息,分别为玉米 (Maize)、小麦 (*Triticum aestivum*) 以及水稻的两个亚种: 籼稻 (*Indica*) 和粳稻 (*Japonica*)。分别从蛋白质结构域相互作用数据库 UniDomInt^[13]和 UniProt 数据库^[14]中获取这4种谷物的蛋白质结构域相互作用信息与其相关数据信息。UniProt 数据库检索条件设置如表1所示。其中籼稻、粳稻的数据为2019年12月版,小麦、玉米的数据为2021年4月版。

表1 UniProt 检索条件

Table 1 UniProt search conditions

检索条件	关键词
annotation	type: "positional domain"
keyword	Molecular function[KW-9992]
database	type: pfam
reviewd	yes or no
organism	<i>Indica</i> ; <i>Japonica</i> ; <i>Maize</i> ; <i>Triticum aestivum</i>

1.2 特征处理

在预测蛋白质相互作用之前,需要对其蛋白质结构域数据进行处理。将任意两个蛋白质之间的结构域组合定义为:

$$D\text{-Pair}(P_i, P_j) = D(P_i) \times D(P_j) \quad (1)$$

$D(P_i)$ 与 $D(P_j)$ 分别代表蛋白质 P_i 、 P_j 中结构域的数目, $D\text{-Pair}(P_i, P_j)$ 为蛋白质 P_i 、 P_j 结构域的笛卡尔积。如图1所示,假设蛋白质 P_i 、 P_j 中分别拥有2个、3个结构域信息,进行笛卡尔积后,获得6组可能存在相互作用的结构域对,其中任意一组相互作用都有可能使得蛋白质 P_i 与 P_j 之间产生相互作用。

继而根据 pfam 编号查询结构域间的关系,获得组合结构域特征向量及其对应的类别标签。由于谷物蛋白质中相互作用的结构域较少,导致数据集不平衡,对其进行采样,最终获得蛋白质结构域组

合数据集 $T=\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 。

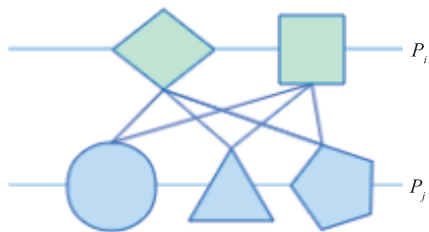


图 1 2 个谷物蛋白质之间所包含的结构域相互作用示意图

Fig. 1 Schematic diagram of the interaction between the structural domains contained in two grain proteins

1.3 功能预测模型

一个蛋白质可能拥有一个或多个 GO 功能注释,故蛋白质功能预测为多标记问题。基于多标记问题,作者首先提出基于集成算法的结构域预测相互作用模型,继而分别提出基于协同分类与基于多层感知机的蛋白质功能预测算法。

谷物 PPI 网络矩阵使用邻接矩阵 $M_{n \times n}$ 表示。将 PPI 网络表示为无向图 $G(V, E)$, 其中顶点集为 $V=\{v_1, v_2, \dots, v_n\}$, v_i 表示蛋白质,边集为 $E=\{e_{ij} | e_{ij}=(v_i, v_j), v_i, v_j \in V\}$, e_{ij} 表示蛋白质 v_i 与蛋白质 v_j 之间存在相互作用。 $M_{n \times n}$ 中每个元素取值定义如下:

$$m_{ij} = \begin{cases} 1, & e_{ij} \in E \\ 0, & e_{ij} \notin E \end{cases} \quad (2)$$

谷物序列相似性使用 $A_{n \times n}$ 表示,使用 blast 工具计算谷物蛋白质氨基酸序列之间的相似性,对蛋白质 V_i 使用 blast 工具进行序列比对后,获得序列相似度向量如下:

$$S(V_i)=[S_{i,1}, S_{i,2}, \dots, S_{i,j}, \dots, S_{i,n}] \quad (3)$$

式中: S_{ij} 表示蛋白质 v_i 与蛋白质 v_j 的序列相似度,当 $i=j$ 时, $S_{ij}=0$ 。为降低谷物蛋白质序列相似性网络中的冗余数据,构建更为精确的谷物蛋白质相似性网络,设置阈值参数 k ,当 $S_{ij}>k$ 时保留,否则将 S_{ij} 置为 0。 $A_{n \times n}$ 中每个元素 $a_{ij}(i=1, 2, \dots, n; j=1, 2, \dots, n)$ 取值定义如下:

$$a_{ij} = \begin{cases} S_{ij}, & S_{ij} \neq 0 \\ 0, & S_{ij} = 0 \end{cases} \quad (4)$$

使用 Cytoscape 工具^[15]对所获得的蛋白质相互作用关系与蛋白质序列相似性进行可视化分析。以水稻谷物的部分蛋白质为例,获得水稻蛋白质相互作用网络示意图见图 2。

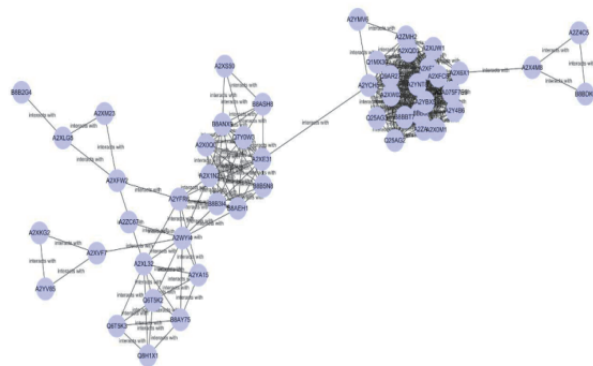


图 2 水稻蛋白质相互作用网络

Fig. 2 Indica rice protein interaction network

蛋白质功能注释矩阵用 $Y_{n \times m}$ 表示,其中 m 表示蛋白质 GO 功能注释标记的个数,其中 $y_{ij}(i=1, 2, \dots, n; j=1, 2, \dots, m)$ 取值定义如下:

$$y_{ij} = \begin{cases} 1, & \text{蛋白质 } v_i \text{ 有被 GO 功能术语 } j \text{ 注释} \\ 0, & \text{蛋白质 } v_i \text{ 未被 GO 功能术语 } j \text{ 注释} \end{cases} \quad (5)$$

1.3.1 AdaBoost 算法 根据两个蛋白质结构域信息,判断这两个蛋白质之间是否具有相互作用为二分类问题。作者使用由 Freund 等提出的 AdaBoost (Adaptive Boosting) 算法,利用不同的权重将弱分类器组装成一个强分类器^[16]。具体算法流程如下:

步骤一:输入蛋白质结构域组合数据集 $T=\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 。其中, $x_i \in X \subseteq R_n, y_i \in Y$, 迭代次数为 M 。

步骤二:初始化权值: $D_1=(w_{1,1}, w_{1,2}, \dots, w_{1,i})$, 权重设为均值 $\frac{1}{N}$, 即 $w_{1,i}=\frac{1}{N}, i=1, 2, \dots, N$ 。

步骤三:对 $m=1, 2, \dots, M$ 进行迭代处理。将更新后的权值分布 D_m 作用于训练数据中,获得弱分类器 $G_m(x)$, 并计算分类误差率:

$$e_m = \sum_{i=1}^N w_{m,i} I(G_m(x_i) \neq y_i) \quad (6)$$

计算弱分类器 $G_m(x)$ 在强分类器中所占的权重比例系数:

$$\alpha_m = \frac{1}{2} \lg \frac{1-e_m}{e_m} \quad (7)$$

更新权值分布矩阵,其中 z_m 为归一化因子。

$$w_{m+1,i} = \frac{w_{m,i} \exp(-\alpha_m y_i G_m(x_i))}{z_m} \quad (8)$$

$$z_m = \sum_{i=1}^N w_{m,i} \exp(-\alpha_m y_i G_m(x_i)) \quad (9)$$

步骤四:得到强分类器:

$$F(x) = \text{sign}(\sum_{i=1}^N a_i G_m(x)) \quad (10)$$

最终得到谷物蛋白质中结构域之间的相互作用关系,继而推导出谷物蛋白质之间的相互作用关系,并据此构建蛋白质相互作用网络,见图 3。

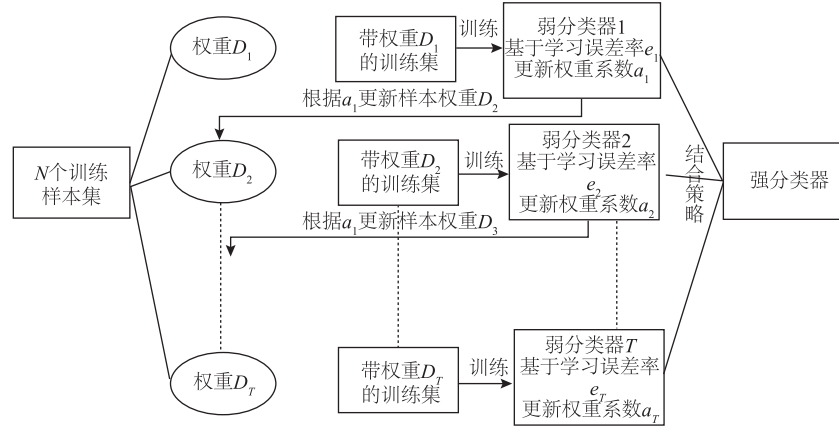


图 3 AdaBoost 算法流程图

Fig. 3 AdaBoost algorithm flow chart

1.3.2 协同分类算法模型 为充分利用谷物蛋白质相互作用与序列相似性网络中的信息,使用协同分类算法对这两个网络进行分析。该算法分为两个阶段:自引导阶段和迭代分类阶段。

在自引导阶段,使用权值投票分别作用于 PPI 网络 $M_{n \times n}$ 与序列相似性网络 $A_{n \times n}$, 计算出未标注蛋白质的初始概率分布 $\vec{a}_x$ 。对于一个未标注的蛋白质 V_x , 与其有相邻关系的节点用权值向量表示为:

$$\begin{cases} N_x^w = [w_{x1}, w_{x2}, \dots, w_{xi}, \dots, w_{xN_x}] \\ N_x^s = [S_{x1}, S_{x2}, \dots, S_{xi}, \dots, S_{xN_x}] \end{cases} \quad (11)$$

式中: N_x^w 表示蛋白质相互作用网络中与蛋白质 V_x 相邻的节点所构成的边的权重; N_x^s 表示蛋白质序列相似性网络中与蛋白质 V_x 相邻的节点所构成的边的权重。从而得到蛋白质 V_x 具有第 j 个 GO 功能注释 F_j 的概率为:

$$P_x^j = \lambda \frac{1}{Z_x^w} \sum_{i=1}^{N_x} w_{xi} f_{ij} + (1-\lambda) \frac{1}{Z_x^s} \sum_{i=1}^1 s_{xi} f_{ij} \quad (12)$$

其中 Z_x^w 和 Z_x^s 是归一化函数。

在式(12)中, P_x^j 的值越大,则说明蛋白质 V_x 越可能具有第 j 个 GO 功能注释。因此蛋白质 V_x 的初始 GO 功能注释概率用特征向量表示为:

$$a_x = [P_x^1, P_x^2, \dots, P_x^m] \quad (13)$$

由于一个蛋白质可能拥有一个或多个蛋白质 GO 功能注释,因此,可用多个类别标签对蛋白质进行标记。在迭代分类阶段,对两种网络中的邻居节点进行投票,不断更新 GO 分子功能概率 $\vec{a}_x$, 共迭代

S 次。对于一个未标记的蛋白质 V_x , 对其进行功能标记可以用下式表示:

$$b_x^1 = \arg \max_{j \in [1, m]} P_x^j \quad (14)$$

其中 b_x^1 表示使得 P_x^j 最大的 j 值, b_x^2 表示使得 P_x^j 第二大的 j 值。那么 b_x^p 表示使得 P_x^j 第 p 大的 j 值。

为了确定两种网络中未知功能蛋白质的标签,将获取的功能标记组合起来。即对于所有蛋白质的功能合集,最终的分类结果是:

$$b_x = [b_x^1, b_x^2, \dots, b_x^m] \quad (15)$$

当分类算法完成所有的迭代过程后,可以得到一个 $s \times m$ 的二维矩阵。其表现形式如式(16)所示:

$$M_x = [b_{x1}, b_{x2}, \dots, b_{xs}]^T \quad (16)$$

在矩阵 M_x 的第一列,其按照 GO 功能注释预测数值进行排序,取数值最大为第一类预测结果,记为 P_x^1 。同样的,在矩阵第二列,将其按照 GO 功能注释预测数值进行排序,将去除 P_x^j 后数值最高的作为 GO 功能注释的第二类预测结果。最终,所获得的这个 GO 功能注释预测结果向量 P_x , 即为所求蛋白质功能预测向量,为:

$$p_x = [P_x^1, P_x^2, \dots, P_x^m] \quad (17)$$

1.3.3 多层感知机算法模型 为深度挖掘蛋白质 PPI 网络和序列相似性网络与其蛋白质功能注释之间的映射关系,作者选用具有单隐藏层的多层感知机 (Multilayer perceptron, MLP) 作为分类模型,其结构示意图见图 4。

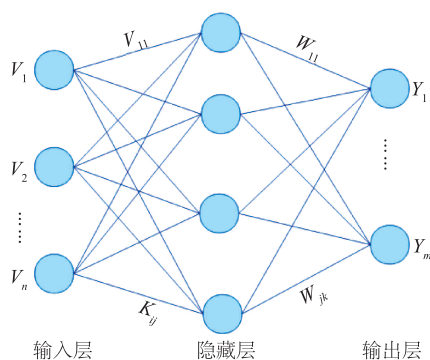


图4 多层感知机算法结构

Fig. 4 Algorithm structure of multi-layer perceptron

多层感知机每层神经元与下层神经元呈现全连接结构,同层神经元之间不存在连接情况。为解决网络线性预测问题,在输出层使用 sigmoid 作为激活函数,其定义如下:

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \quad (18)$$

算法具体流程如下:输入谷物蛋白质相关信息数据集 $D=\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 其中, x_i 为谷物蛋白质相互作用网络 $M_{n \times n}$ 或序列相似性网络 $A_{n \times n}$, y_i 为谷物蛋白质功能注释矩阵 $Y_{n \times m}$ 。构造一个具有 n 个输入层神经元、 m 个输出层神经元的单隐层前馈神经网络, 隐藏层节点数量依照经验公式设置为 $\sqrt{n \times m}$ 。随机选取初值连接系数 w , 偏置量 b , 训练过程中不断更新 w 和 b 的值, 则隐藏层输出 H 为:

$$H = \theta(Xw_h + b_h) \quad (19)$$

其中, θ 为激活函数, X 为输入样本数据, w_h 为隐藏层的连接系数, b_h 为隐藏层的偏置量。

输出层最终输出 O 为:

$$O = Hw_o + b_o \quad (20)$$

式中: w_o 为输出层的连接系数, b_o 为输出层的偏置量。

1.4 评价标准

1.4.1 AdaBoost 算法评价标准 基于混淆矩阵选用准确率(A)、精确率(P)、召回率(R)与 F1 值(F1-measure)对 AdaBoost 算法进行评价以衡量该算法预测效果。其定义如下:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (21)$$

$$P = \frac{TP}{TP+FP} \quad (22)$$

$$R = \frac{TP}{TP+FN} \quad (23)$$

$$F_1\text{-measure} = \frac{2PR}{P+R} \quad (24)$$

式中: TP 表示成功预测正例的个数; TN 表示成功预测反例的个数; FP 表示未成功预测正例的个数; FN 表示未成功预测反例的个数。

1.4.2 功能预测评价标准 蛋白质功能预测问题为多标记分类问题。因此选用 3 种主流的多标记分类评价指标汉明损失 (Hamming Loss)、Macro-F1 (maF1) 与 Micro-F1 (miF1) 以及 1.4.1 提及的精确率 (Precision)、召回率 (Recall) 对该模型进行评价。Hamming Loss 定义如下:

$$HL(h) = \frac{1}{N} \sum_{i=1}^N \frac{|h(x_i) \oplus y_i|}{y} \quad (25)$$

其中, $\oplus$ 表示蛋白质中预测标记集合与真实标记集合的对称差分; N 为谷物中所有 GO 功能注释的数量。

Macro-F1 与 Micro-F1 定义如下:

$$\left. \begin{aligned} \text{maF1}(h) &= \frac{1}{N} \frac{\sum_{i=1}^N y_i h_i(x_i)}{\sum_{i=1}^N y_i + \sum_{i=1}^N h(x_i)} \\ \text{miF1}(h) &= \frac{2 * \sum_{i=1}^N \langle h(x_i, y_i) \rangle}{\sum_{i=1}^N |h(x_i)| + \sum_{i=1}^N y_i} \end{aligned} \right\} \quad (26)$$

其中 $\langle \cdot \rangle$ 表示模型中预测标记与真实标记之间的数量积。

实验流程如图 5 所示。

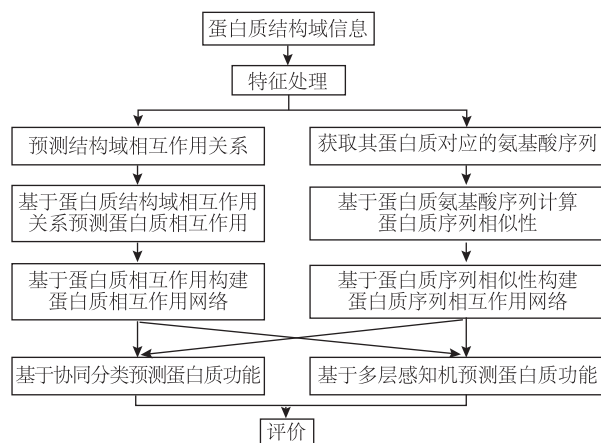


图5 论文实验流程

Fig. 5 Experimental process of the paper

2 结果与分析

2.1 AdaBoost 算法性能评价

使用上述评价标准对 AdaBoost 算法预测谷物蛋白质结构域相互作用的效果进行评价, 水稻和粳稻的结构域相互作用预测结果如表 2 所示。

由表 2 可以发现, 对于这 4 种谷物, 预测准确率均达到了 85% 及以上, 其他指标也较为优秀, 说明其预测效果较好。

2.2 功能预测算法性能评价

采用十折验证法对数据集进行训练, 利用协同

分类算法与多层感知机算法对蛋白质功能进行预测。在构建蛋白质相似性网络时, 设定阈值为 k , 以 0.1 的大小在 [0.1, 0.5] 区间递增, 最终预测结果见表 3。

表 2 AdaBoost 预测结构域相互作用关系结果

Table 2 AdaBoost prediction results of the interaction relationship between the domains

数据	准确率	精确率	召回率	F_1 值
籼稻 (<i>Indica</i>)	0.8579	0.7627	0.6145	0.6807
粳稻 (<i>Japonica</i>)	0.9782	0.9606	0.9711	0.9658
玉米 (<i>Maize</i>)	0.8625	0.6875	0.6471	0.6667
小麦 (<i>Triticum aestivum</i>)	0.9533	0.9516	0.8565	0.9016

表 3 谷物蛋白质功能预测结果

Table 3 Prediction results of grain protein function

数据	方法	参数	汉明损失 ↓	MaF1 值 ↑	MiF1 值 ↑	精确率 ↑	召回率 ↑
籼稻 (<i>Indica</i>)	协同分类	$k=0.1$	0.077 2	0.642 9	0.922 8	0.228 1	0.742 0
	MLP		0.016 3	0.910 8	0.983 6	0.598 7	0.665 8
	协同分类	$k=0.2$	0.077 7	0.641 4	0.922 3	0.226 1	0.742 0
	MLP		0.016 0	0.911 5	0.983 9	0.605 2	0.665 8
	协同分类	$k=0.3$	0.077 0	0.642 5	0.923 0	0.227 5	0.743 9
	MLP		0.017 0	0.905 2	0.982 9	0.579 7	0.668 2
	协同分类	$k=0.4$	0.076 3	0.630 1	0.923 7	0.212 7	0.671 3
	MLP		0.018 8	0.904 1	0.981 1	0.550 2	0.548 9
	协同分类	$k=0.5$	0.077 6	0.622 5	0.922 4	0.205 9	0.623 8
	MLP		0.019 4	0.897 3	0.980 5	0.537 0	0.501 1
粳稻 (<i>Japonica</i>)	协同分类	$k=0.1$	0.026 0	0.707 0	0.974 0	0.367 4	0.613 9
	MLP		0.012 0	0.914 5	0.987 9	0.538 1	0.479 8
	协同分类	$k=0.2$	0.025 9	0.707 4	0.974 1	0.368 1	0.613 9
	MLP		0.012 0	0.914 5	0.987 9	0.538 1	0.479 8
	协同分类	$k=0.3$	0.026 1	0.707 5	0.973 9	0.361 8	0.630 0
	MLP		0.012 5	0.914 4	0.987 5	0.522 0	0.475 7
	协同分类	$k=0.4$	0.026 0	0.700 6	0.974 0	0.362 7	0.601 8
	MLP		0.012 7	0.914 3	0.987 2	0.509 2	0.420 2
	协同分类	$k=0.5$	0.026 6	0.691 2	0.973 4	0.346 7	0.561 5
	MLP		0.013 4	0.910 8	0.986 5	0.474 7	0.364 7
玉米 (<i>Maize</i>)	协同分类	$k=0.1$	0.005 9	0.829 4	0.994 1	0.678 4	0.723 4
	MLP		0.006 9	0.914 7	0.993 1	0.533 8	0.248 6
	协同分类	$k=0.2$	0.005 9	0.829 4	0.994 1	0.678 4	0.723 4
	MLP		0.006 8	0.914 7	0.993 1	0.534 3	0.245 1
	协同分类	$k=0.3$	0.005 2	0.859 7	0.994 8	0.728 2	0.777 0
	MLP		0.006 6	0.917 8	0.993 3	0.583 7	0.225 9
	协同分类	$k=0.4$	0.004 5	0.868 2	0.995 5	0.737 3	0.789 3
	MLP		0.006 5	0.920 8	0.993 4	0.645 5	0.178 6
	协同分类	$k=0.5$	0.003 9	0.879 2	0.996 1	0.756 2	0.806 1
	MLP		0.006 5	0.919 1	0.993 4	0.666 7	0.136 6

续表 3

数据	方法	参数	汉明损失 ↓	MaF1 值 ↑	MiF1 值 ↑	精确率 ↑	召回率 ↑
小麦 (<i>Triticum aestivum</i>)	协同分类	<i>k</i> =0.1	0.006 7	0.917 0	0.993 3	0.823 8	0.877 6
	MLP		0.013 7	0.919 5	0.986 2	0.837 2	0.336 4
	协同分类	<i>k</i> =0.2	0.006 7	0.917 0	0.993 3	0.823 8	0.877 6
	MLP		0.013 7	0.919 5	0.986 5	0.837 2	0.336 4
	协同分类	<i>k</i> =0.3	0.004 5	0.936 4	0.995 5	0.862 2	0.908 9
	MLP		0.013 2	0.921 5	0.986 7	0.890 2	0.341 1
	协同分类	<i>k</i> =0.4	0.003 5	0.945 1	0.996 5	0.883 1	0.922 9
	MLP		0.013 6	0.919 3	0.986 3	0.848 8	0.341 1
	协同分类	<i>k</i> =0.5	0.004 2	0.929 5	0.995 8	0.860 7	0.877 0
	MLP		0.014 2	0.917 6	0.985 7	0.819 2	0.317 7

注:字体加粗为两种算法中表现较优的数据结果。

表 3 可以发现,对于不同的谷物信息,可能拥有的最优阈值参数并不相同。原因是减小 *k* 值会使得数据中信息含量增多,同时数据噪声也会增多,增大 *k* 值反之。

对于籼稻、粳稻来说,当 *k*=0.2 时整体效果较优。对于玉米来说,当 *k*=0.5 时整体效果较优。对于小麦来说,当 *k*=0.4 时整体效果较优。

同时,多层感知机的最优阈值参数总是等于或略小于协同分类算法时的阈值参数,可能是由于多

层感知机更希望拥有足够多的数据进行分析,抵抗数据噪声能力较强,而协同分类算法更希望拥有更加准确稳定的信息,对数据噪声抵抗能力不如多层感知机算法。

为分析实验中部分谷物准确率偏低的原因,从 4 种谷物的实验结果中分别选出 3 条蛋白质数据,并将其预测的蛋白质功能与真实功能相比较。由表 4 可得,协同分类与 MLP 两种算法对于谷物蛋白质均能够较好地预测出蛋白质的功能。

表 4 使用协同分类算法的谷物蛋白质预测功能对比

Table 4 Comparison of cereal protein prediction functions using collaborative classification algorithm

谷物	蛋白质	实际功能	协同分类预测结果	MLP 预测结果
籼稻 (<i>Indica</i>)	A2YQ56	GO:0004176 GO:0004252 GO:0005524 GO:0003677× GO:0004674× GO:0008289× GO:0042803×	GO:0004176 GO:0004252 GO:0005524 GO:0000155× GO:0003677× GO:0004674× GO:0008289× GO:0042803×	GO:0004176 GO:0004252 GO:0005524 GO:0003677× GO:0003952× GO:0004359×
	A2WM14	GO:0003677 GO:0003700	GO:0003677 GO:0003700 GO:0016491×	GO:0003677 GO:0003700 GO:0003735× GO:0019843×
	A2XQD3	GO:0004674 GO:0005524 GO:0030246	GO:0004674 GO:0005524 GO:0030246 GO:0000155× GO:0004707× GO:0016491×	GO:0004674 GO:0005524 GO:0030246

续表 4

谷物	蛋白质	实际功能	协同分类预测结果	MLP 预测结果
梗稻 (<i>Japonica</i>)	Q75V57	GO:0004674 GO:0005524	GO:0004674 GO:0005524 GO:0005516× GO:0009931×	GO:0004674 GO:0005524
	A1A698	GO:0000155 GO:0004673× GO:0019955×	GO:0000155 GO:0003677× GO:0005509× GO:0005516× GO:0005524× GO:0043424×	GO:0000155 GO:0043424×
	Q6YUL8	GO:0003777 GO:0005524 GO:0008017 GO:0016887 GO:0043565×	GO:0005524 GO:0008017 GO:0016887 GO:0003779× GO:0005509× GO:0005516× GO:0008569× GO:0009931× GO:0043621×	GO:0003777 GO:0005524 GO:0008017 GO:0016887 GO:0000287× GO:0042803× GO:0043531×
玉米 (<i>Maize</i>)	P52855	GO:0003735 GO:0044877 GO:0030295× GO:0043021×	GO:0003735 GO:0044877	GO:0003735 GO:0044877
	P35083	GO:0003785 GO:0070064	GO:0003785 GO:0070064	GO:0003785 GO:0070064
	Q43257	GO:0005506 GO:0016709 GO:0020037 GO:0036190×	GO:0005506 GO:0016709 GO:0020037 GO:0036192×	GO:0005506 GO:0016709 GO:0020037 GO:0010333× GO:0097007×
小麦 (<i>Triticum aestivum</i>)	Q41575	GO:0003677 GO:0046982	GO:0003677 GO:0046982	GO:0003677 GO:0046982
	P01083	GO:0004867 GO:0015066	GO:0004867 GO:0015066 GO:0019863×	GO:0004867 GO:0015066 GO:0019863×
	Q9ST57	GO:0004867	GO:0004867	GO:0004867

注:预测失败的功能使用×标出。

由表 4 中可以看出,当蛋白质拥有一个或多个 GO 功能注释时,两种算法均能做出较为准确的预测,例如编号为 P35083、Q41575、Q9ST57 的蛋白质,其预测功能与实际功能完全一致。还有一部分蛋白质其功能注释全部都得到正确预测,例如编号为 A2WM14、A2XQD3、Q75V57、P01083 的蛋白质。同时表 4 可以看出,梗稻与籼稻单个蛋白质大多都同时拥有多于 3 个的 GO 功能注释,且两种算法均能正确预测出其中的蛋白质功能注释信息,可能由于初始信息不全,未能完全包含可预测出其它 GO 功

能注释所需的谷物蛋白质信息。同时评价指标 pre 的公式设置,也在一定程度上降低了蛋白质功能预测的准确率。例如,对于编号为 P52855 的蛋白质来说,两种方法均未预测出 GO:0030295、GO:0043021 这两个 GO 功能注释,但对于其他功能注释的预测并未发生错误。同时由于相同的蛋白质结构域可能服务于不同的蛋白质功能,因此通过结构域进行相互作用的预测分析时可能预测出多个蛋白质相互作用,从而使得预测出未发生在本蛋白质而可能发生在其他蛋白质中的功能,降低准确率。例如编号

为 Q43257 的蛋白质, 分别预测出了未存在的 GO:0036192、GO:0010333、GO:0097007, 其均与多个 GO 功能注释存在相连关系。除此以外, 观察到编号为 P01083 的蛋白质, 在正确预测出其蛋白质功能注释的基础上, 均预测出了蛋白质 GO 注释库中未显示拥有的 GO:0019863, 其含义为与 IgE 同种型的免疫球蛋白选择性和非共价相互作用, 这为以后实验研究提供了新的方向, 即通过生物方法尝试验证此蛋白质中是否真正含有此 GO 功能注释。同时也提供了新的研究思路, 利用计算方法对蛋白质功能注释的搜索与发现寻求新途径。

为更直观地比较两种算法的实验结果, 将其预测结果用柱状图表示。其中, 图 6~9 依次为籼稻、粳

稻、玉米、小麦两种算法预测结果。

比较图 6—图 9 籼稻、粳稻、玉米和小麦的结果数据图, 多层感知机算法拥有较好的 MaF1 值与准确率, 而协同分类算法拥有较高的召回率, 同时注意到玉米的实验结果受阈值 k 影响较为明显。

总而言之, 对于不同的谷物数据不仅拥有不同的阈值参数, 且适用的算法也不相同。籼稻与粳稻同为水稻的亚种, 因此数据结果整体表现差别不大, 且在多层感知机算法上呈现出较好效果。而对于玉米和小麦, 整体来说使用协同分类算法效果较好。对于谷物其他品种而言, 作者另选燕麦、荞麦做蛋白质功能预测实验, 均能获得较好的预测效果, 多层感知机算法拥有更优的准确率, 协同分类算法拥

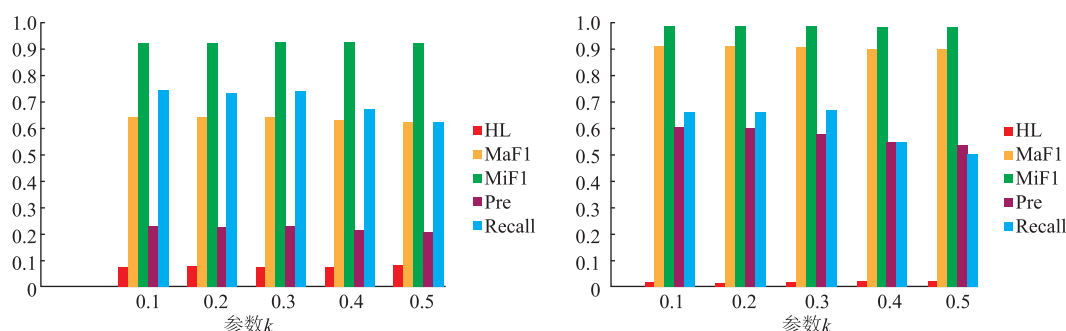


图 6 籼稻协同分类(左)与 MLP(右)结果数据图

Fig. 6 Results of collaborative classification of *Indica* rice (left) and MLP (right)

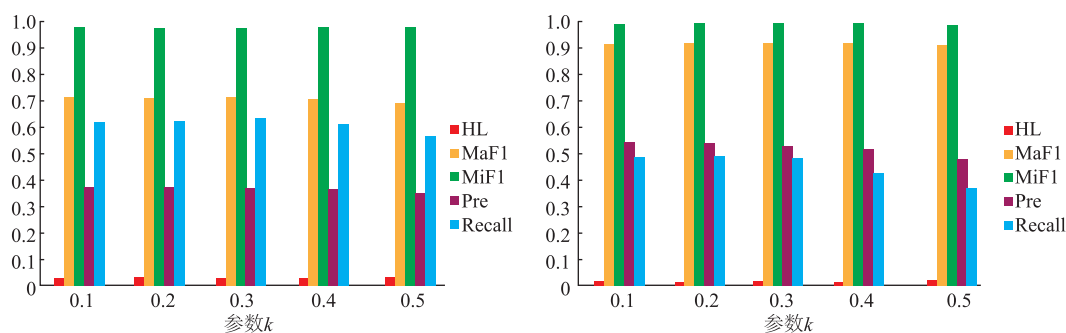


图 7 粳稻协同分类(左)与 MLP(右)结果数据图

Fig. 7 Results of collaborative classification of *japonica* rice (left) and MLP (right)

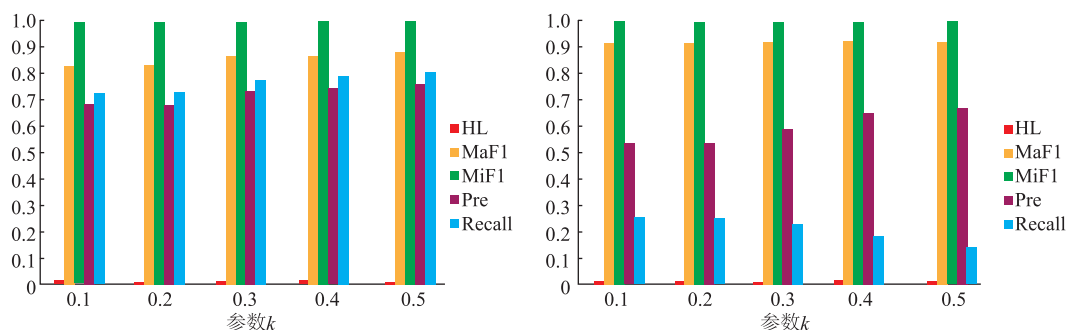


图 8 玉米协同分类(左)与 MLP(右)结果数据图

Fig. 8 Results of collaborative classification of *Maize* (left) and MLP (right)

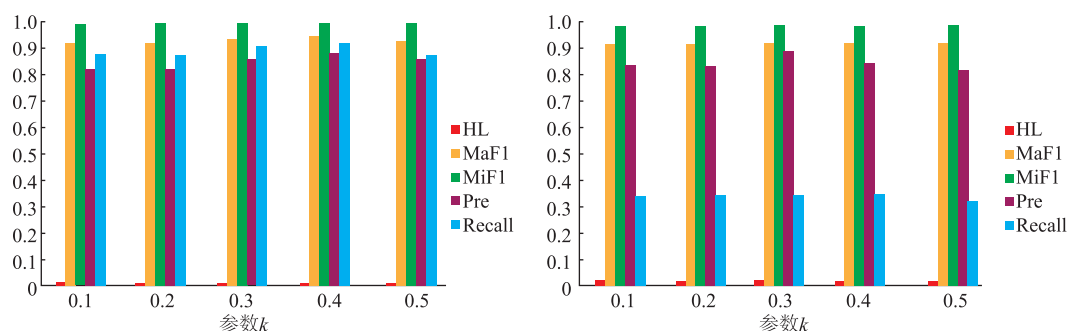


图9 小麦协同分类(左)与MLP(右)结果数据图

Fig. 9 Results of collaborative classification of *Triticum aestivum* (left) and MLP (right)

有更优的召回率,符合上述规律。

3 结语

作者使用籼稻、粳稻、玉米和小麦4种谷物蛋白质信息进行实验,结果表明,对于不同的谷物种类,不同的阈值参数或使用不同的预测方法对预测性能均有不同的影响。整体而言,两种算法均能实

参考文献:

- [1] 时逢宽,李炜疆. 序列相似性网络聚类与蛋白质家族划分[J]. 食品与生物技术学报, 2014, 33(1): 98-103.
- [2] STEPHN O. Proteomics: Guilt-by-association goes global[J]. *Nature*, 2000, 403(6770).
- [3] SCHWIKOWSKI B, UETZ P, FIELDS S. A network of protein-protein interactions in yeast[J]. *Nature Biotechnology*, 2000, 18: 1257-1261.
- [4] VAZQUEZ A, FLAMMINI A, MARITAN A, et al. Global protein function prediction from protein-protein interaction networks [J]. *Nature Biotechnology*, 2003, 21(6): 697-700.
- [5] DENG M, ZHANG K, MEHTA S, et al. Prediction of protein function using protein-protein interaction data [C]. Proceedings. IEEE Computer Society Bioinformatics Conference[A]. IEEE, 2002.
- [6] 梅娟,何胜,李炜疆. 基于图聚类的蛋白质相互作用网络功能模块探测[J]. 食品与生物技术学报, 2011, 30(1): 95-100.
- [7] HAN D, KIM H S, SEO J, et al. A domain combination based probabilistic framework for protein-protein interaction prediction [J]. *Genome Informatics*, 2003, 14: 250-259.
- [8] HAN D, KIM H S, JANG W H, et al. PreSPI: a domain combination based prediction system for protein-protein interaction[J]. *Nucleic Acids Research*, 2004, 32(21): 6312-6320.
- [9] HAN D, KIM H S, JANG W H, et al. PreSPI: design and implementation of protein-protein interaction prediction service system [J]. *Genome Informatics*, 2004, 15(2): 171-180.
- [10] HAYASHID A M, KAMAD A M, SONG J, et al. Conditional random field approach to prediction of protein-protein interactions using domain information[J]. *BMC Systems Biology*, 2011, 5(1): S8.
- [11] SINGHAL M, RESAT H. A domain-based approach to predict protein-protein interactions[J]. *Bmc Bioinformatics*, 2007, 8(1): 199.
- [12] ALTSCHUL S F. Basic local alignment search tool (BLAST)[J]. *Journal of Molecular Biology*, 1990, 215(3): 403-410.
- [13] BJORKHOLM P, SONNHAMMER E L L. Comparative analysis and unification of domain - domain interaction networks[J]. *Bioinformatics*, 2009, 25(22): 3020-3025.
- [14] ZVELEBIL M J. UniProt. Dictionary of Bioinformatics and Computational Biology[M]. American Cancer Society, 2014.
- [15] KOHL M, WIESE S, WARSCHIED B. Cytoscape: software for visualization and analysis of biological networks[M]. Humana Press, 2011: 291-303.
- [16] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]. European Conference on Computational Learning Theory[A]. Berlin: Springer, 1995: 23-37.

(责任编辑:李春丽)