DOI:10.11918/j.issn.0367-6234.201803117

水稻组织特异性蛋白质相互作用网络构建方法

邢林林[1], 郭茂祖[1,2,3],刘晓燕[1], 李 傲[1]

(1.哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001; 2.北京建筑大学电气与信息工程学院, 北京 100044;
3. 建筑大数据智能处理方法研究北京市重点实验室,北京 100044)

摘 要: 组织特异的基因表达和蛋白质相互作用是研究基因调控、蛋白质功能、细胞过程的重要部分. 相较于其他模式生物在蛋白质相互作用研究方面的进展,高等模式植物水稻中组织特异性蛋白质相互作用的研究十分缺乏. 因此,提出了一种用于水稻组织特异性蛋白质相互作用网络构建的计算方法. 该方法主要包含三部分:第一,在统一标准下融合多数据识别组织特异的基因;第二,提出了新的同源映射方法,并集成6种模式生物相互作用数据构建和评估目标物种蛋白质相互作用网络;第三,构建不同组织的蛋白质相互作用子网,并筛选高可靠的蛋白质相互作用. 为了验证方法的有效性,构建并分析了水稻首个组织特异的蛋白质相互作用网络(PTSN4R:Predicted Tissue-Specific Network for Rice). PTSN4R 包含了水稻23个组织的组织特异基因及对应的组织特异蛋白质相互作用子网,为分析组织特异的基因表达和蛋白质相互作用提供了便利条件. PTSN4R 有助于理解水稻的生长调控机制,为水稻增产提供线索. 同时,提出的方法能够方便的应用到其他物种,促进组织特异的蛋白质相互作用网络的研究.

关键词: 组织特异基因;同源映射;蛋白质-蛋白质相互作用;组织特异的蛋白质相互作用
中图分类号: TP391 **文献标志码**: A **文章编号**: 0367-6234(2018)11-0001-09

A tissue-specific protein interaction network construction method for rice

XING Linlin[1], GUO Maozu[1,2,3], LIU Xiaoyan[1], LI Ao[1]

(1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001;
2. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044;
3. Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing 100044)

Abstract: The expression pattern of genes and protein interactions in specific tissues are important frameworks for studying gene regulation, protein function, and cellular processes. Compared with the research progress of other model organisms in the interactome, the tissue-specific protein interaction research and development in higher plants is very slow, especially in rice. With this motivation, we have proposed a computing framework to predict tissue-specific protein-protein networks for rice. This framework consists of three parts: (a) identification of tissue-specific genes by integrating multiple dataset under a unified criterion; (b) prediction and evaluation of the protein interaction network based on the resource of six model organisms by using the proposed novel Interolog mapping method; (c) tissue-specific subnet construction in each tissue and high reliable interactions filter based on co-expression correlation. To evaluate the effectiveness of our framework, PTSN4R (Predicted Tissue-Specific Networks for Rice) is constructed and analyzed. PTSN4R is the first integrated database for tissue-specific protein interactions of rice, which contains tissue-specific genes and the interaction networks of 23 rice tissues. And, it provides a tissue-specific perspective to conveniently analyze the gene expression and protein interaction. These resources can help researchers understand the intrinsic regulatory mechanisms of rice growth and development and provide clues for rice yield increase. In addition, the proposed framework can extend to other species easily to improve the research of tissue-specific protein interactions.

Keywords: tissue-specific genes; Interolog mapping; protein-protein interaction; tissue-specific network

水稻作为农作物研究的重要模式生物[1],相关的基因组信息[2-3]、转录组数据积累得越来越多,相关的研究在逐年增长. 充足的数据使得人们能够对水稻基因的表达模式进行详细的分析. 在过去的十多年里,人们使用统计学方法对水稻基因的表达模式进行大量的研究[4-7],包括组织转录特征[8]、水稻在外部胁迫下的转录响应[9]、田间环境表达模式[10]等. 研究人员也整合相关资源构建了便于检索的集成数据库[11-14],包括集成功能注释与基因组信息的

收稿日期: 2018-03-22
基金项目: 国家自然科学基金(61571163, 61532014, 61671189, 91735306);国家重点研发计划课题(2016YFC0901902)
作者简介: 邢林林(1987—),男,博士研究生;
郭茂祖(1966—),男,教授,博士生导师
通信作者: 郭茂祖,maozuguo@hit.edu.cn

OryzaBase[11] 和 RiceDB[12],综合的基因共表达网络 ATTED - II[13], 功能相关网络 RiceNetV2[14]. Chandran 和 Jung 对水稻相关的系统生物学的资源信息进行汇总,讨论了水稻组学数据的发展过程以及最近的进展[15].

与模式植物拟南芥相比,高等植物水稻的生长环境复杂,培育时间长,生物实验代价高昂. 使得水稻的研究主要集中在基因层面整体转录模式的分析,缺少蛋白质相互作用方面的研究. 目前用于预测新的蛋白质相互作用的计算方法包括基于蛋白质空间结构的方法[16],集成功能基因组信息的机器学习方法[17-18] 以及基于进化保守性的同源映射方法[19-21]. 已知空间结构信息较少和计算效率低下的问题使得基于空间结构的方法并不适用于大规模植物蛋白质相互作用网络预测. 与此同时,功能注释信息的缺乏和已知的相互作用较少使得基于 SVM 等的机器学习方法也不适用于水稻蛋白质相互作用的预测.

基于进化保守性的同源预测方法不需要功能注释信息和金标准数据,是目前研究的热点. 在 2011 年和 2012 年,Chen 的团队和 Matt 的团队基于同源映射方法分别预测了水稻的蛋白质相互作用网络,并使用收集的数据以及预测的亚细胞定位信息对计算的蛋白质相互作用进行了分析说明[22-23]. 这两项研究包含 5 573 个水稻基因,约占水稻基因总数 14%. 两项包含的基因重叠程度较高,但是预测得到的相互作用差异较大,重叠程度较低,Chen 的工作 PRIN 中只有 30%的相互作用在 Matt 的预测结果中出现. 而且,Chen 和 Matt 的工作是从整体上对水稻蛋白质相互作用的预测,并不能够反应蛋白质相互作用的组织特异性. 此外,这两项工作完成时间较早,对于后续出现大量的相互作用资源利用不够充分. 因此,水稻在基因和蛋白质的组织特异性表达与相互作用方面的研究十分滞后.

综上所述,当前水稻组织特异性蛋白质相互作用的研究中存在如下几点问题:1)数据利用不充分,组织特异基因的识别没有统一的标准;2)目前的蛋白质相互作用网络的预测方法单一,数据滞后,网络覆盖程度不够;3)缺少组织特异性蛋白质相互作用状况的研究. 因此,水稻组织特异性的蛋白质相互作用的预测依旧是一个挑战,有待进一步的研究.

为了解决上述问题,本文以水稻为研究对象提出一种组织特异性蛋白质相互作用网络构建方法,并构建了水稻的组织特异性网络(PTSN4R). 其目的是为预测组织特异的蛋白质相互作用提供一个合理的框架,为研究人员从组织特异的角度来研究基因的表达和蛋白质相互作用提供便利. 首先,通过分析目前组织特异基因的识别方法确定了一致的识别标准,进而识别出水稻各组织中组织特异基因,并使用预测的水稻必需基因进行交叉验证. 然后,提出基于支持得分的同源映射方法并构建了水稻全基因组的蛋白质相互作用网络. 最后,根据识别的组织特异基因和预测的蛋白质相互网络构建了水稻各组织中组织特异的蛋白质相互作用网络,并根据基因表达相关性筛选出了高可靠的蛋白质相互作用.

1 组织特异性蛋白质相互作用网络构建方法

1.1 组织特异基因的识别

基因的组织特异性表达是基因重要特性之一. 完成组织特异表达基因的识别需要根据样本表达特征将数据按照合适的层次划分到不同组织中. 首先按照原始数据集中标注的组织划分,然后根据样本间的表达相关性,对相同以及相近组织的表达谱样本进行确认,根据相似度合并成为该组织的表达数据集. 不同品种的往往具有不同的表达特性,因此,将水稻数据集 GSE19024 中 MH,ZS 两个品种区分处理.

根据 Nadezda 与 Marc 对众多组织特异性表达指标的比较研究,使用由 Yanai 等提出 Tau[24] 作为表达特异性的指标.

Tau 的计算公式为

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1};\ \hat{x}_i = \frac{x_i}{\max\limits_{1\leqslant i\leqslant n}(x_i)}. \qquad (1)$$

式中:x_i 是基因 x 在第 i 个组织中的表达值,$\overline{x_i}$ 是使用最大分量值正则化的表达值,n 是组织的个数. 根据公式可看出,Tau 值越接近于 1 则表示这个基因是组织特异表达的.

同时,使用基于同源性方法预测了水稻的必需基因,用来与识别的差异表达基因进行对比验证. a)与 DEG 数据库的最佳比对必须满足 E-value<1e-10,Identity>40%;b)与 D-Non-EG 数据库的最佳比对结果必须满足 E-value<1e-10, Identity<40%.

1.2 直系同源基因的预测方法

根据目前蛋白质相互作用网络的完善程度选择 6 个参考物种(大肠杆菌,酵母菌,蠕虫,果蝇,人类,拟南芥)作为源物种. 使用局部序列比对搜索工具(BLAST)来识别目标物种基因在参考物种中的直系同源基因. 首先,将水稻以及 6 个参考物种的蛋白质序列制作成 BLASTP 程序可用的蛋白质序列局部

比对数据库. 然后，使用 BLASTP 程序在 6 个物种的蛋白质序列数据库中进行 BLAST 比对搜索水稻基因的同源基因. 计算水稻所有蛋白质的直系同源蛋白质序列以及每对同源基因对应的 E 值（E-value）和一致性（Identity）和比特得分（Bit Score）. 满足 E-value>1e-10 的序列认为与目标序列存在直系同源关系.

1.3　直系同源映射和蛋白质相互作用评分

现有直系同源映射方法均从源物种出发，搜索已有相互作用(A',B')在目标物种中的直系同源对作为预测的蛋白质相互作用. 这一策略不能确定预测的相互作用关系受到已有相互作用关系的支持程度. 为了解决这一问题，提出了一种新的基于直系同源映射原理的方法来预测蛋白质相互作用.

基于支持得分的同源映射方法步骤，见图 1. 其中，图 1(a)为主要的步骤说明，图 1(b)为直系同源基因识别以及直系同源映射的示意图. 首先根据直系同源基因预测方法分别识别出 A，B 的直系同源的所有基因，存入与 B'. 然后检查所有可能 A'_i 与 B'_j 在源物种中是否存在实验验证的相互作用关系. 如果 A'_i 与 B'_j 存在相互作用，而且同时满足 J_E<1e-40，J_I<0.7. 那么就认为 $<A'_i,B'_j>$ 就是支持 $<A,B>$ 存在相互作用的 PPI. 根据式(5)计算 $<A,B>$ 的支持得分. 如果在源物种中发现多对满足直系同源映射的相互作用关系，见图 1(b)，那么认为这一相互作用更为可靠. 基于同源映射方法来预测相互作用时往往会集成多个物种的相互作用数据.

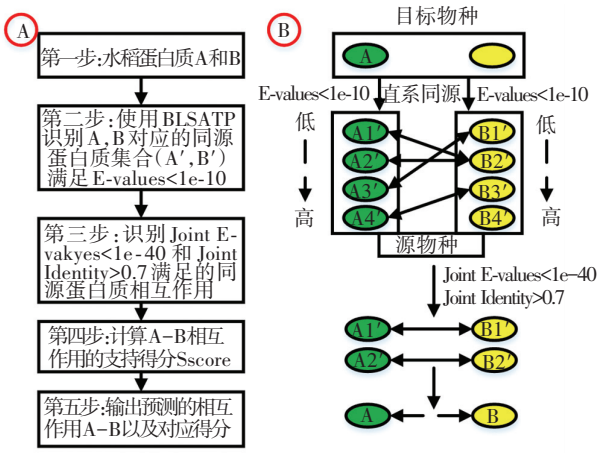


图 1　蛋白质相互作用预测及评分示意

Fig.1　Diagram of the prediction and evaluation of PPI

通常使用联合序列相似性作为直系同源映射可靠程度评价标准. 联合序列比对一致性 J_I，与联合 E 值 J_E 是常用的评价指标，定义如下：

$$J_I = \sqrt{I_A * I_B}. \tag{2}$$

$$J_E = \sqrt{E_A * E_B}. \tag{3}$$

式中：I_A，I_B 分别为蛋白质 A，B 的比对一致性值；E_A，E_B 分别为蛋白质 A，B 的 E 值. J_I，J_E 作为阈值筛选相互作用只能反应最佳匹配的同源映射对于预测的蛋白质相互作用的影响. 不能正确的评价多个匹配对于预测出的相互作用的影响. 为评价预测的蛋白质相互作用受到多个匹配的影响程度，提出了支持得分 Sscore 来衡量预测的相互作用被源物种中的相互作用的支持程度. 比特得分反映两序列的匹配程度，本文使用正规化的比特得分来构建支持得分 Sscore. 计算 Sscore 的基本步骤如下：

1)将 A，B 的直系同源基因的比特得分分别使用本组得分最大值正规化.

$$\hat{a}_i = \frac{a_i}{\max\limits_{1\leqslant i\leqslant n}(a_i)};\ \hat{b}_j = \frac{b_j}{\max\limits_{1\leqslant i\leqslant m}(b_j)}. \tag{4}$$

式中：a_i 为 A 的第 i 个同源蛋白质 A'_i 的比特得分. b_j 为 B 的第 j 个同源蛋白质 B'_j 的比特得分. n 为源物种中与目标物种蛋白质 A 存在直系同源关系的蛋白质的数目. m 为源物种中与目标物种蛋白质 B 存在直系同源关系的蛋白质的数目.

2)那么支持得分 S_{score} 为两个直系同源组中所有实验验证的相互作用的几何平均值的和.

$$S_{\text{score}} = \sum_{<A'_i-B'_j>\in Is} \sqrt{\hat{a}_i * \hat{b}_j}. \tag{5}$$

式中：Is 为物种中已经验证的所有的蛋白质相互作用的集合.

1.4　组织特异的蛋白质相互作用提取

从全基因组层面计算得到了蛋白质相互作用网络与具体的组织或者发育过程无关. 如果某一组织中相互作用对的一个蛋白质不存在，那么这一相互作用就不会发生. 针对每个特定组织，抽取特异表达的基因来构建组织特异的子网，并依据共表达相关性构建高可靠的组织特异的蛋白质相互作用网络. 具体构建方法如下：

算法 1：组织特异性蛋白质相互作用提取算法

输入：全部基因集 G_{all}，组织特异表达基因集 G_{tis}，完整蛋白质相互作用网 P_{all}

输出：特定组织的子网 P_{tis}，高可靠的相互作用子网 P_{hc}

$k \leftarrow \text{size}(G_{\text{tis}})$；$n \leftarrow \text{size}(G_{\text{all}})$；$P_{\text{tis}} \leftarrow \varnothing$

FOR 每个基因　$g_i \in G_{\text{tis}}$

FOR 每个基因 $g_j \in G_{\text{all}}$

　IF $(g_i \neq g_j)$

　　$t \leftarrow \text{PCC}(g_i, g_j)$

　ELSE

　　$t \leftarrow -1$

END

IF ($< g_i, g_j > \in P_{all}$)

$P_{tis} \leftarrow P_{tis} \cup < g_i, g_j >$

END

IF ($t \geqslant 0.7$&& $< g_i, g_j > \in P_{all}$)

$P_{hc} \leftarrow P_{hc} \cup < g_i, g_j >$

END

END

END

通过分析算法伪代码可知算法 1 包含两重循环,因此算法 1 的时间复杂度为 $O(n^2)$.

使用 NCBI Gene ID 作为统一的标识,使得所有节点可以直接链接到 NCBI 数据库,方便用户获取基因的注释信息并转向相关资源进行进一步分析.

1.5　组织特异的蛋白质相互网络构建整体方法

提出的组织特异性蛋白质相互作用网络构建方法见图 2. 构建过程包括 3 个主要步骤:1) 识别与验证组织特异表达的基因;2)基于同源映射方法预测全基因组蛋白质相互作用;3)构建组织特异基因的蛋白质相互作用子网,并筛选可靠性较高的蛋白质相互作用. 图 2 中蓝色框线标识的部分分别为上述主要步骤的结果,分别为预测的组织特异的基因,组织特异基因的共表达网络,预测的全基因组蛋白质相互作用网络,预测的组织特异的蛋白质相互作用子网.

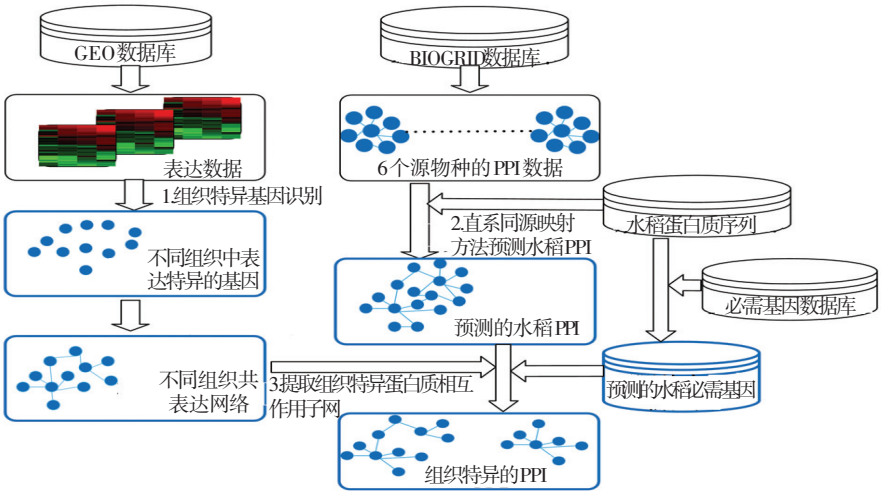


图 2　水稻组织特异性蛋白质相互作用网络计算方法

Fig.2　Tissue-specific protein-interaction network construction method for rice

2　实验结果及分析

根据上述方法构建了水稻的组织特异的蛋白质相互作用子网(PTSN4R). 图 2 中蓝色框标识的部分提供下载.

2.1　数据来源

2.1.1　表达数据

从 GEO 数据库收集了 3 个平台的 5 组公开的水稻表达谱数据,详见表 1.

表 1　使用的水稻表达数据集

Tab.1　Used expression data of rice

数据编号	平台	组织数/个	样本数/个	探针数/个
GSE6893	GPL2025	7	45	57 381
GSE13161	GPL1829	40	220	60 727
GSE19024	GPL2025	31	190	57 381
GSE21396	GPL6864	13	143	45 220
GSE14304	GPL2025	25	98*	57 381

* 不包含 16 个 Aglient 平台的样本.

2.1.2　相互作用数据

开放获取的蛋白质相互作用数据库 BIOGRID 作为模式生物相互作用数据的来源. 从中收集整理了 6 个参考物种(大肠杆菌 Escherichia coli,酵母菌 Saccharomyces cerevisiae, 蠕虫 Caenorhabditis elegans, 果蝇 Drosophila melanogaster, 人类 Homo sapiens,拟南芥 Arabidopsis thaliana)的数据作为同源性映射方法中源物种的相互作用,见表 2.

表 2　6 个参考物种相互作用网络统计

Tab.2　The number of interactions in six reference species

物种名称	相互作用数/个	基因数/个
大肠杆菌	164 717	3 965
酵母菌	231 326	6 535
蠕虫	8 037	3 949
果蝇	40 211	8 369
人类	278 538	21 247
拟南芥	35 641	9 529

2.1.3　蛋白质序列数据

水稻(粳稻，Oryza sativa subsp. japonica)以及 6 个参考物种的蛋白质序列数据从 NCBI 参考基因组数据库获得. 必需基因的蛋白质序列数据从必需基因数据库 DEG[25] 收集.

2.2　水稻组织特异基因的识别结果

在五组数据上分别进行组织特异表达基因的识别. 由于同一数据集中相似度较高的样本会降低特异表达识别的准确度，在数据预处理阶段将相关系数大于 90%的样本合并为一组. 各个数据集上基因的 Tau 值的分布见图 3. 从图 3 可看出 Tau 值在靠近 0.8 的位置出现单峰. 根据 Tau 值的分布以及参考文献[26]，选取 Tau>0.8 的基因作为组织特异的基因. GSE13161 的生物样本为不同的细胞类型，先按照各个细胞类型进行组织特异基因的识别，然后将各类型细胞的识别结果归并到对应的组织中.

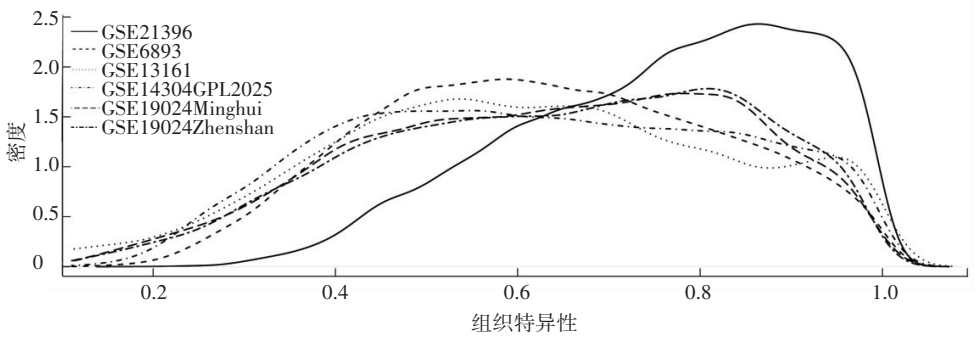


图 3　各个数据集上 Tau 值的分布

Fig.3　The distribution of Tau on each dataset

此时得到水稻 23 个不同的组织的特异表达基因. 各个数据集中检测到的特异表达基因数目并不一致，这些差异同时也表明当前的数据数量、质量和计算方法不能够提供组织特异表达的完全确定的证据. 人们对于特异表达的研究还需要进一步的细化.

数据集 GSE14304，GSE19024，GSE21396 对应的文献中，作者提供了部分组织特异的基因. 图 4 展示了识别的组织特异的基因与对应参考文献识别结果的重叠程度. 从图 4 可看出使用 Tau 指标识别的组织特异基因基本包含参考文献给出的基因，并且 Tau 方法识别了更多的组织特异的基因[26].

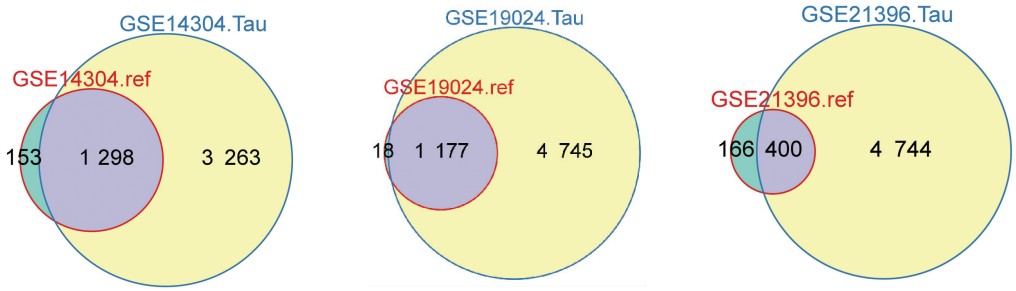


图 4　基于 Tau 识别的组织特异基因与参考文献比对结果

Fig.4　Comparison of tissue-specific genes in ref papers and by Tau method

根据必需基因的定义[25]，必需基因不会出现组织特异性的表达. 根据同源性预测方法(见 1.3)得到了水稻的 2 880 个必需基因. 根据定义，预测的必需基因与组织特异基因之间的交集必定较小. 将这些必需基因与识别的组织特异的基因进行比较，发现识别的组织特异的基因与预测的必需基因只有极少数重叠.

2.3　水稻全基因组蛋白质相互作用

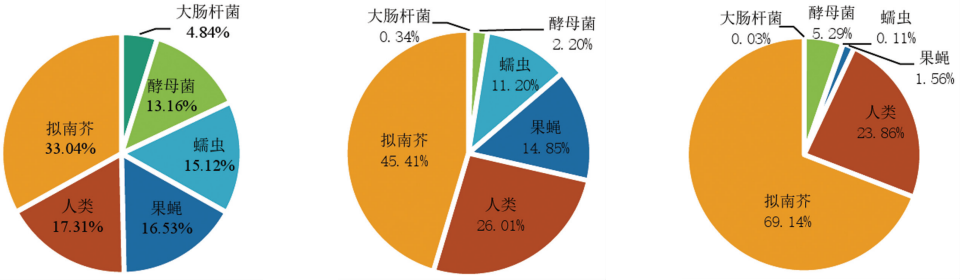
2.3.1　预测的水稻蛋白质相互作用网络

水稻与 6 个参考物种的进化距离远近不一，从各个参考物种中得到的直系同源基因的数量也有较大差异. 图 5(a)展示了水稻与各个参考物种存在直系同源关系的基因的数目. 图 5(b)是 BLASTP 程序发现的各个基因的直系同源蛋白质组的总数. 共计得到 6 560 基因的 116 393 条相互作用关系. 图 5(c)展示了这些蛋白质相互作用的来源分布.

水稻与 6 个物种的进化距离由近到远分别是：拟南芥，人类，果蝇，蠕虫，酵母菌，大肠杆菌. 从图 5 可看出，水稻与拟南芥进化距离最近，具有直系同源关系的基因数目最多，在拟南芥中命中的直系同源基因占所有 6 个物种的 33%，BLASTP 检测到的直系同源物则占到 46%. 那么根据进化保守性原理，水稻从这 6 个物种中继承来的蛋白质相互作用关系的数目应该由多到少，即来自与拟南芥的最多，来自

于大肠杆菌的最少. 从图 5(c)中可看出预测的蛋白质相互作用网络中来自拟南芥的相互作用最多. 因此,水稻的蛋白质相互作用网以及基因调控网络将会更类似于拟南芥.

图 6 展示了预测的蛋白质相互作用的支持得分 Sscore 的分布,柱形图表示各个分数段中蛋白质相互作用的数目,曲线图表示累积百分比.



(a)命中基因来源　　(b)BLASTP结果来源　　(c)预测的蛋白质相互作用来源

图 5　同源映射方法预测的水稻蛋白质相互作用的分布

Fig.5　The distribution of predicted PPIs by Interolog method

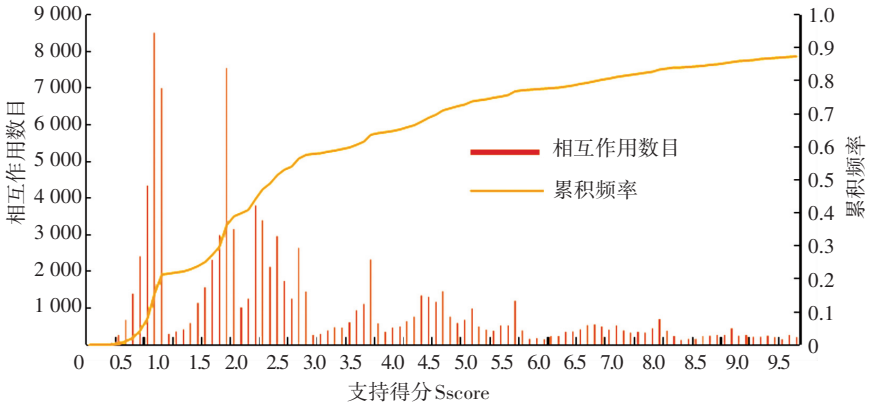


图 6　支持得分 Sscore 的分布情况

Fig.6　The distribution of support score

支持得分小于 1 表示源物种的直系同源物中没有找到最佳匹配的蛋白质对来支持目标物种中的蛋白质相互作用. 根据支持得分含义以及分布情况,选择 Sscore>1 的相互作用作为高可靠性的相互作用. 经过筛选后,预测的蛋白质相互作用总共包含 6 088 个基因,98 696 条相互作用关系.

对于上述筛选的蛋白质相互作用,从网络的拓扑性质,GO 注释,共表达相关性 3 个方面进行可靠性分析与验证.

2.3.2　预测的网络的拓扑性质

图 7 是预测的蛋白质相互作用网络的具有代表性的网络拓扑结构的参数.

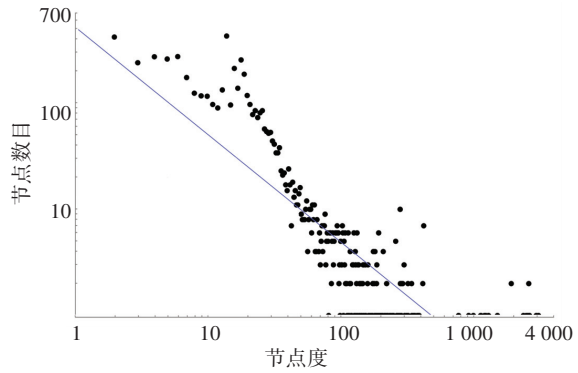
从网络中节点的度分布情况(图 7(a))可看出预测的网络呈现出无标度性质,蛋白质相互作用网络的无标度特性使得生物对于环境变化或者刺激具有更强大的承受能力. 从图 7(b)网络的平均聚类系数的分布情况可看出本文预测的网络的平均聚类系数较高. 可见预测的网络具有与真实网络类似的模块化特性. 图 7(c)是平均最短路径长度的分布情况,预测的网络的最短路径长度多数落在 2~5 之间. 由图 7(b)、(c)可知预测的网络是小世界网络. 蛋白质相互作用网络的小世界特性有利于生物信号在网络中的迅速传播,提高了生物响应外界刺激的速度与环境适应能力. 从上述分析可看出预测的网络的拓扑性质与真实的蛋白质相互作用网络的拓扑性质类似.

2.3.3　GO 共注释分析

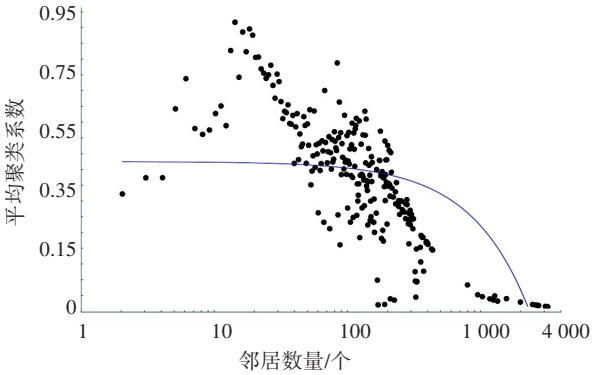
图 8 展示了具有相同注释的蛋白质相互作用的 GO 术语分布情况. GO 术语分为分子功能(MF, Molecular Function),生物过程(BP, Biological Process),细胞组成(CC, Cellular Component)3 个大类. 具有相同注释的蛋白质发生相互作用的可能性更高.

目前水稻的 GO 注释的丰度较低,在 3 个 BP, CC, MF3 个类别上的平均注释个数为 1.44, 1.06, 1.60. 对具有注释的蛋白质进行统计,发现 3 个 GO 类别中具有相同注释的相互作用的比例分别为:CC 中 26.3%, BP 中 18.3%, MF 中 23.4%. 从图 8(a)为的

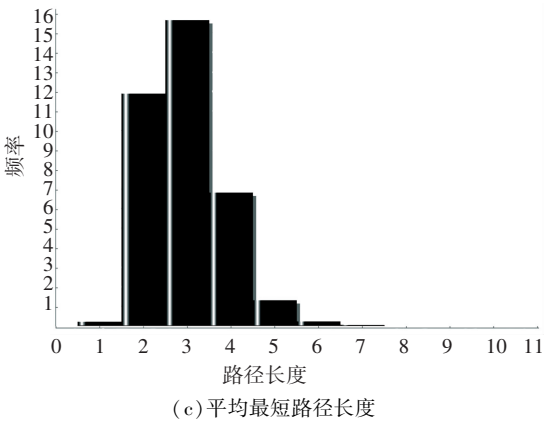
MF 类别看，超过 60%的相互作用被标记为“绑定”功能. 在 BP 类别(图 8(b))中发现中接近一半(48.3%)的蛋白质相互作用与翻译和转录相关. 而在 CC 分类(图 8(c))中，前三名分别被标记为细胞质(38.78%)与细胞核(37.21%)，膜结构(12.06%). 这是蛋白质相互作用的主要场所. 这说明预测蛋白质相互作用可靠性较高.



(a)节点度分布



(b)平均聚类系数



(c)平均最短路径长度

图 7　预测的蛋白质相互作用网络拓扑性质

Fig. 7　Topological properties of predicted protein interaction network

2.3.4　预测的蛋白质相互作用共表达情况

表达水平的相关性一定程度上能够反应蛋白质相互作用的可靠程度. 因此，可以使用预测的蛋白质相互作用在转录层面的相关性来验证预测的蛋白质相互作用网络的可靠程度. ATTED-II 是一个针对拟南芥、水稻等模式植物的共表达信息库，使用比相关系数更可靠的 Mutual-Rank(MR，相互排序)来衡量基因对之间的共表达关系. MR 值越小，说明基因之间的共表达关系越强. 图 9 是预测的蛋白质相互作用网络中所有蛋白质对对应的基因的共表达分布情况. 从图上可看出 30.04%的蛋白质对的共表达相关性排在前 15%，46.36%的蛋白质对的共表达相关性排在前 30%. 这说明预测的蛋白质相互作用有显著地共表达倾向，即预测的蛋白质相互作用网络可信程度较高.

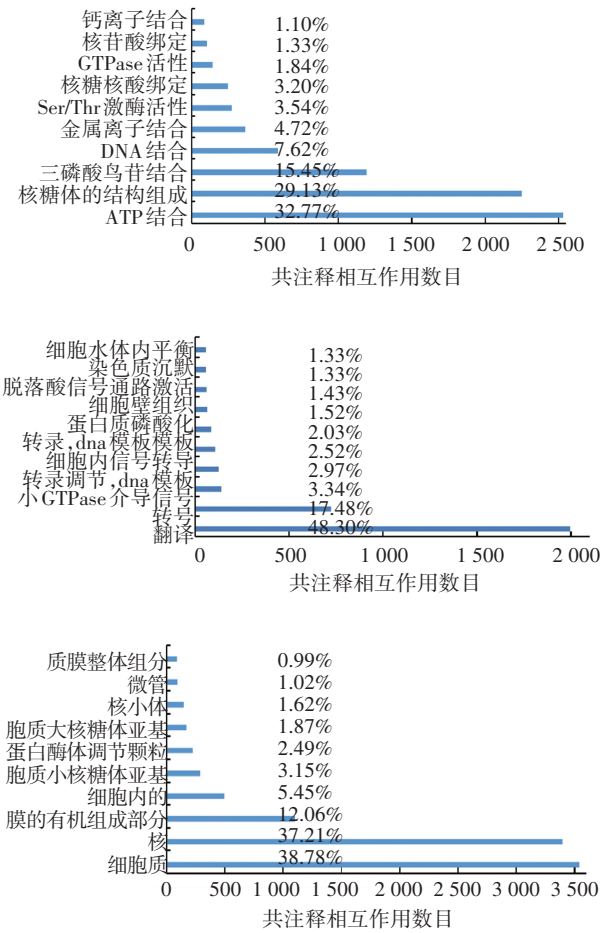


图 8　共注释的蛋白质相互作用的 GO 术语分布

Fig. 8　GO term distribution of co-annotation interactions

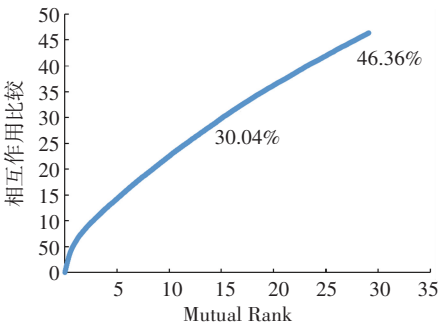
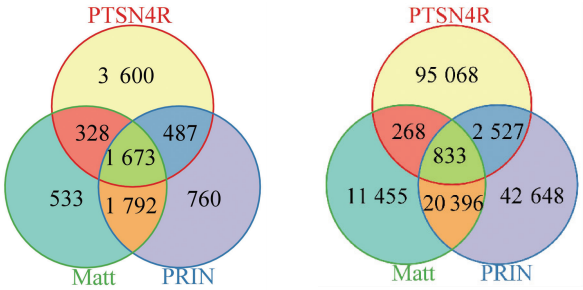


图 9　预测的网络中蛋白质对的共表达分布情况

Fig. 9　Co-expression rank distribution of protein pairs in predicted network

2.3.5　三项预测工作的比较

图 10 展示了预测的结果 PTSN4R 与之前类似工作 PRIN 以及 Matt 预测的结果的异同. 图 10(a)表示 PTSN4R 与 Matt 的工作以及 PRIN 中所涉及的基因的重叠程度. 从图 10(a)可看出,PRIN 以及 Matt 二者包含的基因重叠程度较高,而本文预测的网络所涉及的基因与二者有较大不同. 这是因为 PTSN4R 中有 3 600 个基因是先前的工作没有涉及的. 图 10(b)展示了三项研究预测的蛋白质相互作用网络的重叠程度. 从图 10(b)可看出,本文预测的网络与前两项工作中预测网络重合程度也有较大差异,新发现 95 068 条蛋白质相互作用关系,极大的丰富了水稻相互作用研究的数据,拓展了人们对水稻蛋白质相互作用的认识. 经过分析发现,新发现的相互作用主要来自拟南芥. 这与近年来拟南芥蛋白质相互作用研究的快速发展有直接关系. 随着蛋白质相互作用研究的不断发展,同源映射方法还有较大的发展空间.



(a)基因交叠程度　　(b)预测的相互作用交叠程度

图 10　三项工作涉及的基因与预测相互作用重叠程度

Fig.10　Overlap of three research work

2.4　组织特异的蛋白质相互作用

为更好地对组织特异的蛋白质相互作用进行研究,将 PRIN、Matt 等人预测的网络以及本文的预测结果合并构建了总的蛋白质相互作用网络. 然后,基于组织特异的基因和预测的蛋白质相互作用网络,构建了水稻 23 个组织的蛋白质相互作用子网. 最后,根据共表达相关性筛选出了各个组织中可靠性较高的蛋白质相互作用. 表 3 列出了各个子网以及筛选的蛋白质相互作用的详细信息.

研究发现,组织特异性的相互作用共计 101 430 条,占全部预测的相互作用的 58.6%. 在这些组织特异的子网中,花药、叶片、根以及胚乳组织的特异表达的基因由多个细胞类型合并而来,因此形成的子网中相互作用数目最多. 从表 3 可看出各个子网都具有相对较高的聚类系数和网络中心度. 即使是相互作用较少的细分组织(心皮,外稃,胚芽,胚根,小穗,柱头)依旧具有较高的网络中心性. 这说明各个组织特异的调控网络是构成整体小世界网络的一部分.

表 3　组织特异的蛋白质相互作用网络

Tab.3　Tissue-specific protein interaction networks

组织名称	相互作用数/个	平均聚类系数	网络直径	平均邻居数/个	中心性	可靠相互作用数/个
花药	26 105	0.439	10	9.92	0.505	42
愈伤组织	8 304	0.624	10	4.68	0.748	101
心皮	320	0	10	2.16	0.3	5
胚胎	12 941	0.103	10	7.52	0.194	54
胚乳	18 124	0.529	10	6.7	0.665	79
发芽种子	6 800	0.732	9	4.3	0.841	60
花序	10 091	0.577	10	5.23	0.689	80
叶片	15 706	0.074	11	7.77	0.09	372
叶鞘	2 712	0.09	8	3.77	0.24	5
外稃	1 456	0.042	15	4.52	0.133	58
卵巢	7 314	0.374	7	7.03	0.274	34
内稃	3 043	0.059	12	3.78	0.172	121
花粉	9 873	0.144	11	6.56	0.171	10
圆锥花絮	851	0.012	8	3.09	0.208	2
胚芽	134	0	3	2.53	0.849	0
幼根	1 619	0.098	10	3.8	0.413	4
根	39 023	0.213	9	11.6	0.399	1 081
种子	3 386	0.147	10	4.46	0.377	14
芽	12 271	0.119	12	6.78	0.142	280
小穗	879	0.012	9	3.39	0.275	13
雄蕊	5 922	0.078	10	5.46	0.121	165
茎	4 851	0.18	11	5.09	0.27	36
柱头	1 139	0.026	8	3.66	0.18	7

从表 3 中可看到在各个组织特异的子网中发现的表达相关性较高的相互作用较少. 进一步分析发现,主要原因有 3 点:1)蛋白质相互作用网络覆盖度不高;2)各个平台包含的基因集合不相同,多平台数据整合过程中不交叉的基因被忽略;3)整合的数据中存在较多检测失败的数据,导致整体的表达相关性较低.

3　总　结

本文针对模式植物水稻,提出了融合多组学数据构建组织特异性蛋白质相互作用网络的计算方法,然后构建了 23 个水稻重要组织或者器官对应的组织特异的蛋白质相互作用子网(PTSN4R),并从多个角度对预测结果进行分析,说明了方法的合理性,有效性.

组织特异的基因和蛋白质相互作用作为 PTSN4R 的核心,拓展了人们对水稻组织特异基因和蛋白质相互作用网络的认知,对基因调控机制和

蛋白质功能的研究具有重要意义. 我们期望 PTSN4R 能够促进组织特异的蛋白质相互作用的研究和对于水稻产量相关性状的调控机制的理解.

目前，农作物相关的转录组以及蛋白质组学数据正在大量产出，对于组织或者细胞特异表达的研究还在进一步的发展. 后续可能的研究思路包括：将提出的方法应用于其他的动植物物种中；研究更可靠的直系同源性度量方法，提高准确性；融合更丰富的数据提高相互作用网络的覆盖度及质量.

参考文献

[1] SHIMAMOTO K, KYOZUKA J. Rice as a model for comparative genomics of plants [J]. Annu Rev Plant Biol, 2002, 53(1): 399. DOI: 10.1146/annurev.arplant.53.092401.134447

[2] GOFF S A, RICKE D, LAN T H, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica) [J]. Science, 2002, 296 (5565): 92. DOI: 10.1126/science.1068275

[3] SAKAI H, LEE S S, TANAKA T, et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics [J]. Plant & cell physiology, 2013, 54(2): e6. DOI: 10.1093/pcp/pcs183

[4] JIAO Yuling, TAUSTA S L, GANDOTRA N, et al. A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies [J]. Nature genetics, 2009, 41(2): 258. DOI: 10.1038/ng.282

[5] FUJITA M, HORIUCHI Y, UEDA Y, et al. Rice expression atlas in reproductive development [J]. Plant & cell physiology, 2010, 51 (12): 2060. DOI: 10.1093/pcp/pcq165

[6] WANG Lei, XIE Weibo, CHEN Ying, et al. A dynamic gene expression atlas covering the entire life cycle of rice [J]. The Plant Journal: For Cell and Molecular Biology, 2010, 61(5): 752. DOI: 10.1111/j.1365-313X.2009.04100.x

[7] CHO K, SHIBATO J, KUBO A, et al. Genome-wide mapping of the ozone-responsive transcriptomes in rice panicle and seed tissues reveals novel insight into their regulatory events [J]. Biotechnology Letters, 2013, 35(4): 647. DOI: 10.1007/s10529-012-1118-x

[8] HAMADA K, HONGO K, SUWABE K, et al. OryzaExpress: an integrated database of gene expression networks and omics annotations in rice [J]. Plant & Cell Physiology, 2011, 52(2): 220. DOI: 10.1093/pcp/pcq195

[9] KUDO T, AKIYAMA K, KOJIMA M, et al. UniVIO: a multiple omics database with hormonome and transcriptome data from rice [J]. Plant & Cell Physiology, 2013, 54(2): e9. DOI: 10.1093/pcp/pct003

[10] SATO Y, ANTONIO B, NAMIKI N, et al. Field transcriptome revealed critical developmental and physiological transitions involved in the expression of growth potential in japonica rice [J]. BMC Plant Biol, 2011, 11(10). DOI: 10.1186/1471-2229-11-10

[11] YAMAZAKI Y, SAKANIWA S, TSUCHIYA R, et al. Oryzabase: an integrated information resource for rice science [J]. Breeding Science, 2010, 60(5): 544. DOI: 10.1270/jsbbs.60.544

[12] NARSAI R, DEVENISH J, CASTLEDEN I, et al. Rice DB: an Oryza Information Portal linking annotation, subcellular location, function, expression, regulation, and evolutionary information for rice and Arabidopsis [J]. The Plant Journal: For Cell and Molecular Biology, 2013, 76(6): 1057. DOI: 10.1111/tpj.12357

[13] OBAYASHI T, OKAMURA Y, ITO S, et al. ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants [J]. Plant & Cell Physiology, 2014, 55(1): e6. DOI: 10.1093/pcp/pct178

[14] LEE T, OH T, YANG S, et al. RiceNet v2: an improved network prioritization server for rice genes [J]. Nucleic Acids Research, 2015, 43(W1): W122-7. DOI: 10.1093/nar/gkv253

[15] CHANDRAN A K N, JUNG K H. Resources for systems biology in rice [J]. J Plant Biol, 2014, 57(2): 80. DOI: 10.1007/s12374-014-0903-6

[16] SMITH G R, STERNBERG M J. Prediction of protein-protein interactions by docking methods [J]. Curr Opin Struct Biol, 2002, 12 (1): 28. DOI: 10.1016/S0959-440X(02)00285-3

[17] WANG Yongcui, WANG Jiguang, YANG Zhixia, et al. Sequence-based protein-protein interaction prediction via support vector machine [J]. J Syst Sci Complex, 2010, 23(5): 1012. DOI: 10.1007/s11424-010-0214-z

[18] SHEN Juwen, ZHANG Jian, LUO Xiaomin, et al. Predicting protein-protein interactions based only on sequences information [J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104 (11): 4337. DOI: 10.1073/pnas.0607879104

[19] MATTHEWS L R, VAGLIO P, REBOUL J, et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs" [J]. Genome Research, 2001, 11(12): 2120. DOI: 10.1101/gr.205301

[20] YU Haiyuan, LUSCOMBE N M, LU H X, et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs [J]. Genome Research, 2004, 14(6): 1107. DOI: 10.1101/gr.1774904

[21] LO Yushu, CHEN Chunchen, HSU K C, et al. Rank-based interolog mapping for predicting proteinprotein interactions between genomes [C]// 2013 7th International Conference on Systems Biology (ISB). Huangshan, China: IEEE, 2013: 55. DOI: 10.1109/ISB.2013.6623794

[22] GU Haibin, ZHU Pengcheng, JIAO Yinming, et al. PRIN: a predicted rice interactome network [J]. Bmc Bioinformatics, 2011, 12 (161). DOI: 10.1186/1471-2105-12-161

[23] HO C L, WU Yinzhou, SHEN Hongbin, et al. A predicted protein interactome for rice [J]. Rice (N Y), 2012, 5(1): 15. DOI: 10.1186/1939-8433-5-15

[24] YANAI I, BENJAMIN H, SHMOISH M, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification [J]. Bioinformatics, 2005, 21(5): 650. DOI: 10.1093/bioinformatics/bti042

[25] ZHANG Ren, LIN Yan. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes [J]. Nucleic Acids Research, 2009, 37(Database issue): D455. DOI: 10.1093/nar/gkn858

[26] KRYUCHKOVA-MOSTACCI N, ROBINSON-RECHAVI M. A benchmark of gene expression tissue-specificity metrics [J]. Brief Bioinform, 2017, 18(2): 205. DOI: 10.1093/bib/bbw008

（编辑　苗秀芝）