通过蛋白质互作网络预测已知部分功能的蛋白质的精细功能

李彦辉[①] 郭 政[①②*] 马文财[①] 杨 达[①] 王 栋[①] 张 敏[①] 朱 晶[①]
钟国才[①] 李永进[①] 姚 晨[①] 王 靖[①]

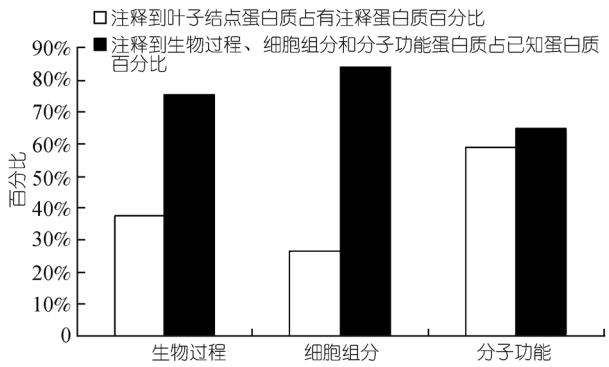
(① 电子科技大学生命科学与技术学院, 成都 610054; ② 哈尔滨医科大学生物信息学系, 哈尔滨 150086.
* 联系人, E-mail: guoz@ems.hrbmu.edu.cn)

摘要 基于高通量数据, 研究人员已经设计了许多算法用于寻找功能完全未知蛋白质的功能. 然而, 这些算法的效率受到一些根本因素的制约, 包括: (ⅰ) 功能完全未知的蛋白质参与一个精细功能的先验概率低; (ⅱ) 高通量互作数据中有大量的假阳性互作; (ⅲ) 蛋白质互作数据对功能类的覆盖不完全; (ⅳ) 训练算法的大量阴性样本数据是异质的; (ⅴ) 训练算法的蛋白质的精细功能知识不足. 因此, 本研究提出一种新的方法对已知部分功能的蛋白质进行功能预测, 即利用功能特异的蛋白质互作子网或者基因表达模式信息来寻找蛋白质更为精细的功能. 该方法能够通过恰当地定义预测范围和过滤假阳性数据减少上述提到的问题, 因此可以高效地发现蛋白质的新功能. 对于几千个已知部分功能的酵母与人类蛋白质, 该方法能够以超过 90%的精确率找到它们更为精细的功能. 预测的精细功能对于指导随后的湿实验和提供必要的功能知识来学习其他蛋白质的功能都具有重要的意义.

关键词 蛋白质互作 **gene ontology** 基因功能 算法 预测

基于大规模的基因表达和蛋白质互作数据, 研究人员已经设计了很多算法来寻找蛋白质的功能. 目前大多数算法的目标是预测功能完全未知的蛋白质, 我们将这类算法称为从头预测算法 [1~5]. 尽管期望寻找到所有蛋白质的精细功能, 然而从头预测算法的效率受到很多根本因素的限制, 其中之一是缺少蛋白质细致的功能知识来训练算法 [6,7]. 如图 1 所示, 按照SGD数据库(*Saccharomyces* genome database)[8], 即使对于研究比较深入的酵母, 也只有约三分之一的蛋白质注释到了GO (gene ontology)[9]中描述最细致的叶子结点. 实际上, 目前许多所谓已知功能的蛋白质只注释到了描述很不具体的功能类, 如"蛋白质生物合成"(图 1), 我们称其为已知部分功能的蛋白质. 显然, 寻找这些蛋白质的精细功能对于了解这些蛋白质和提供必要的数据来学习其他蛋白质的功能都具有重要意义.

为了寻找已知部分功能的蛋白质更精细的功能, 我们曾提出基于基因表达谱的深层预测方法 [10]: 将基因(蛋白质)从其已注释到的功能类向下预测一层. 本研究进一步发展了这样的深层预测策略: 利用蛋



结点编号	结点名字	*N*	*K*
GO: 0006412	蛋白生物合成	253	238
GO: 0000723	端粒维持	145	144
GO: 0007047	细胞壁构成及生物发生	104	96

图 1 已知部分功能蛋白质在 GO 中的分布

基于 2006 年 5 月 29 日下载自 GO 的酵母注释数据, 注释到生物过程、细胞组分和分子功能的酵母蛋白质占酵母总蛋白质的 77%, 84%和 66%, 然而, 只有 39%, 29%和 58%的蛋白质注释到了 GO 中生物过程、细胞组分及分子功能 3 个体系描述最细致的叶子结点. 包括已知部分功能蛋白质数目最多的前 3 个生物过程功能类也列在了表中. *N* 表示在一个功能类中部分已知蛋白质的数目, *K* 是被 BioGRID 数据集中蛋白质互作数据覆盖的已知部分功能的蛋白质数目. GO 的注释数据经常更新. 例如, 在 2006 年 5 月 29 日注释到 GO: 0006412("蛋白质生物合成")的 238 个已知部分功能蛋白质中的 197 个近期已经注释到了 GO: 0043037("翻译")

2007-05-21 收稿, 2007-09-25 接受
国家自然科学基金(批准号: 3037088 和 30670539)资助项目

白质互作数据，将蛋白质从其已注释到的功能类向下预测一层或多层，发现其更精细的功能. 显然，由于已知部分功能的蛋白质参与一个子功能类的先验几率增大，预测的可靠性可能会提高. 进一步，使用注释到同一个功能类中的蛋白质，可以过滤掉部分假阳性互作，因此该方法对于含有大量假阳性的高通量互作数据是稳健的. 大量假阳性互作是限制功能预测可靠性的另一个重要因素 [11~14].

用酵母的蛋白质互作数据和基因表达谱数据，我们展示用深层预测方法寻找已知部分功能的蛋白质的精细功能的效果，并且比较用两类数据预测紧密联系的子功能类的能力. 结果显示，对于许多注释到如“蛋白质生物合成”等粗泛功能类中的蛋白质，深层预测方法可以精确地将它们预测到“翻译起始”等精细功能类. 相反，对于同一组蛋白质，不利用它们已知功能的从头预测算法通常不能将它们以一个可以接受的精确率预测到同样精细的功能，以指导下一步的湿实验. 另外，从头预测算法的训练集合通常是高度不平衡的，大量松散定义的阴性数据导致了类之间的重合 [1,2]. 训练数据不平衡通常是导致很多预测算法预测效率低下的主要原因，特别是基于基因表达谱的预测 [6,12,15]. 因此，本研究也讨论不平衡的表达谱数据的处理方法 [12,16]对预测效果的影响.

基于蛋白质互作数据和深层预测方法，以高于90%的精确率，为几千个已知部分功能的酵母与人类蛋白质预测了精细的功能.

1 材料和方法

(ⅰ) 蛋白质互作网络和基因表达谱数据. 蛋白质互作数据包括物理互作和遗传互作，这两种数据都适合进行功能预测 [17]. 酵母和人类的蛋白质互作数据分别来源于基于文献的BioGRID数据库 [18]和HPRD数据库 [19]. 按如下的方法预处理数据: (1) 删除自身互作; (2) 只保留一个通过不同检测方法观察到的相同互作(酵母双杂交和免疫共沉淀); (3) 仅保留在GO中有注释的蛋白质. 处理后的BioGRID数据包括 5927 个蛋白质和 50434 条互作, HPRD数据包括 8338 个蛋白质和 31800 条互作. SGD的注释数据下载于 2006 年 5 月 29 日和 2006 年 9 月 29 日, EBI Uni-Prot[20]的人类注释数据下载于 2007 年 5 月 15 日.

酵母表达谱数据(下称为Stress173)包括 173 张芯片, 6152 个基因 [21]. 对数据进行标准化，使每张芯片的均值为 0，标准差为 1. 删除检测值缺失率超过10%的基因，其他缺失值采用K近邻(K = 15)方法估计 [22,23].

(ⅱ) 修改大数法用于二分类算法. 首先选定一个 GO 结点(功能类)作为深层预测的目标结点，定义它的任何一个祖先结点为预测空间(图 2). 按照 GO 的注释体系，将注释到预测空间而没有注释到它的任何一个子结点的蛋白质定义为已知部分功能蛋白质，即预测对象. 然后通过连接注释在预测空间中互作的蛋白质构建一个功能特异的互作子网，孤立的蛋白质被排除在外. 在互作子网中，注释到目标结点的蛋白质被当作阳性样本，而除预测对象外的其他蛋白质被当作阴性样本.

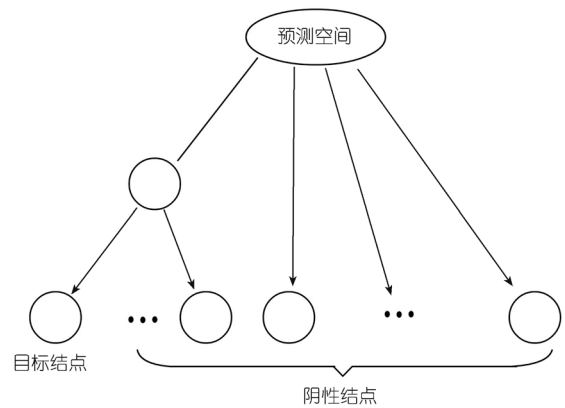


图 2 预测空间

预测空间中蛋白质分为 3 类: 预测对象，即已知部分功能蛋白质，它是指注释到预测空间而没有注释到它的任何一个子结点的蛋白质; 阳性样本，注释到目标结点的蛋白质; 阴性样本，预测空间中除预测对象和阳性样本外的蛋白质

按照“大数规则”[3]，一个蛋白质通常被赋予与其直接相互作用的邻居蛋白质中出现频率最高的几个功能(如 3 个)[3,24]. 尽管一个蛋白质可以执行多个功能，但一般无法知道究竟应该为一个蛋白质赋予几个功能. 再考虑到湿实验的成本，预测高度可信的功能可能更为符合实际需要. 所以，我们选择只为蛋白质赋予一个可信度最高的子功能，尽管这种保守的处理方法可能会遗漏蛋白质的其他功能. 具体地，我们修改“大数规则”以适合预测结果是阳性或者阴性的二类化问题: 在功能特异的蛋白质互作网络中，一个蛋白质被赋予最频繁的直接相互作用邻居蛋白质的标签(阳性或者阴性). 另外，如果一个蛋白质与同等数目的阳性和阴性样本(蛋白质)互作，则被判为阴

性. 显然, 因为目标结点中的阳性样本要和预测空间中所有其他子结点中的阴性样本进行竞争, 因此修改的大数法对于预测一个阳性结果来说是保守的.

(iii) 预测效果的评价. 我们采用留一法来评价分类器的预测效果. 每一个训练样本都要被轮流留出来作为测试样本. 计算真阳性(TP)、真阴性(TN)、假阳性(FP)和假阴性(FN), 再计算精确率、覆盖率和 F 指标.

$$\text{精确率}=\frac{\text{TP}}{\text{TP}+\text{FP}},\quad \text{覆盖率}=\frac{\text{TP}}{\text{TP}+\text{FN}},$$

$$F=\frac{2\times\text{精确率}\times\text{覆盖率}}{\text{精确率}+\text{覆盖率}}.$$

精确率是预测为阳性的样本中真阳性样本所占的比例, 而覆盖率是阳性样本中被预测为真阳性的比例. 对于一个给定的目标结点, F指标用来评价算法的整体效果 [6,7], 分值越高表示预测效果越好.

2 结果和分析

2.1 根据蛋白质互作数据从“蛋白质生物合成”向下预测

作为一个案例, 首先采用留一法评价了从 GO:0006412(“蛋白质生物合成”)向其子结点的预测(见材料和方法). 如表 1 所示, 基于 2006 年 5 月份的 GO 注释数据, 从“蛋白质生物合成”到它的一部分子结点的预测效果很好(F > 0.6). 例如, 尽管 GO:0006497 (蛋白质氨基酸的脂代谢)子功能类非常细致, 并且其由 35 个阳性样本和 144 个阴性样本组成的训练数据高度不平衡, 该结点的预测效果仍然很好, 精确率和覆盖率分别为 97%和 80%. 然而, 并不是所有子功能类都容易被识别. 蛋白质互作数据和注释数据对功能类的覆盖率显然是影响预测效果的一个重要因素. 例如, “蛋白质生物合成”的子结点 GO:0019988 (氨酰 tRNA 修饰), 是不适合进行预测分析的, 因为蛋白质互作网络只包括 2 个具有此功能的蛋白质. 同样, 子结点 GO:0006057 (甘露糖蛋白质的生物合成)也不适合进行预测, 因为 12 个注释到其中的蛋白质之间只有一对互作, 而它们大多数与注释到 GO:0009101(糖蛋白质的生物合成)的蛋白质互作. 按照我们的建议, GO 委员会近期已经将“甘露糖蛋白质生物合成”调整为“糖蛋白质生物合成”的子结点.

根据“蛋白质生物合成”, 能够以很高的精确率(表 1)预测许多已知部分功能的蛋白质到精细的功能类. 例如, 根据 2006 年 5 月份的 GO 注释数据, 可以以高达 95%的精确率(表明每一个预测结果都有 95%的概率是真的)将 238 个已知部分功能的蛋白质中的 91 个预测到 GO:0043037(翻译). 实际上, 91 个被深层预测的蛋白质中的 77 个的最近更新注释(2006 年 9 月份)和预测结果完全一致.

对于 238 个直接注释到“蛋白质生物合成”的蛋白质(2006 年 5 月份), 197 个已经在近期注释到了“翻译”功能类(2006 年 9 月份). 在更新的数据上, 从“蛋白质生物合成”到“翻译”的预测效果得到了改善, 精确率和覆盖率分别达到了 97%和 95%. 在“蛋白质生物合成”没有更新注释的 41 个蛋白质中有 18 个可以被预测到“翻译”(表 2). 因为直系同源蛋白质一般具有相同的功能 [25], 通过在其他物种中寻找直系同源

表 1 “GO:0006412: 蛋白生物合成”一些子结点的预测效果

GO 编号	名称	精确率	覆盖率	F 指标
GO:0009101	糖蛋白生物合成	0.88	0.92	0.90
GO:0042158	脂蛋白生物合成	0.97	0.80	0.88
GO:0043037	翻译	0.95	0.98	0.96
GO:0006486	蛋白氨基酸糖基化	0.82	0.91	0.86
GO:0006497	蛋白氨基酸脂质化	0.97	0.80	0.88
GO:0006487	蛋白氨基酸 N-连接糖基化	0.81	0.71	0.76
GO:0006506	GPI 锚生物合成	0.85	0.69	0.76
GO:0018342	蛋白异戊烯化	1	0.80	0.89
GO:0018377	蛋白豆蔻酰化	1	0.80	0.89
GO:0016255	GPI 锚向蛋白的黏附	1	0.60	0.75
GO:0018319	蛋白氨基酸豆蔻酰化	1	0.80	0.89
GO:0018346	蛋白氨基酸异戊烯化	1	0.80	0.89
GO:0006499	蛋白 N 末端豆蔻酰化	1	0.80	0.89
GO:0006413	翻译起始	0.67	0.86	0.75
GO:0006414	翻译延伸	0.82	0.56	0.67

表 2 从“蛋白生物合成”到“翻译”功能类的预测 a)

目标功能类	预测	阴性邻居	阳性邻居
GO:0043037	IST1	无	SUI3, GCN3, GCD7, GCD6, GCD1, GCD2
GO:0043037	TPD3	无	RPS24B, RPG1, SSB2, RPS1B, RPS6A, RPS17A
GO:0043037	THS1	无	RPL18A
GO:0043037	CBP6	无	GCN3
GO:0043037	MTO1	无	PPQ1
GO:0043037	DOM34	无	RPS30A, HBS1
GO:0043037	PET309	无	MSS51
GO:0043037	FES1	无	TIF1, SSA1
GO:0043037	PPH21	无	PPE1
GO:0043037	SRO9	无	TIF4631
GO:0043037	CDC55	无	RPS1B, TEF4, RPS17A, YEF3
GO:0043037	PET122	无	MRP17, MRP1, PET123
GO:0043037	PPH22	无	PPE1, MKT1
GO:0043037	PET111	无	MSS51
GO:0043037	ILS1	无	SUI1, DED81, FRS1
GO:0043037	YJR136C	无	RSM23
GO:0043037	HCR1	HOC1	SUI1, RPS18A, RPG1, TIF34, NIP1, RLI1, TIF5, RPS8A, PRT1, TIF35
GO:0043037	RTS1	无	GUS1

a) 基于蛋白互作数据的预测: 精确率为 0.97, 覆盖率为 0.95; 基于基因表达谱数据的预测: 精确率为 0.93, 覆盖率为 0.9. TPD3, RPL19B, FES1, CBP6, SRO9, SLM5, PPH21, CBS2, YDR341C, PET122, CDC55, MTO1, PET54, RRF1, MTG2, THS1, PET309, RPM2, MSS1, MTG1, AEP2, DOM34, IST1, PET494, MSD1, MSF1

蛋白质, 可以为一些预测结果提供证据. 例如, TPD3被预测到了“翻译”. 在NCBI的HomoloGene[26]中, 它的人类直系同源基因*PPP2R1A*编码的蛋白质磷酸酶2A PR65/A亚基注释到了“翻译调控”. 此外, 有报道认为, TPD3 参与翻译的起始 [27]、延长 [28]和终止 [29].

然而, 对于同一组已知部分功能的蛋白质, 如果使用从头预测算法确定它们是否具有“蛋白质生物合成”的子结点所描述的功能, 则一般不使用这些蛋白质先前已知的功能知识(“蛋白质生物合成”), 而要从根结点开始预测. 在从头预测算法中, 已知部分功能的蛋白质是被预测的对象, 它们被当作功能完全未知的蛋白质进行分析. 为了评价这种从头预测的可靠性, 按通常的从头预测算法, 定义注释到目标结点的蛋白质为阳性样本, 注释到“生物过程”中的蛋白质(除了预测对象)为阴性样本. 结果显示, 按照 F 指标, 从头预测到“蛋白质生物合成”的子结点的预测效果较差. 不仅覆盖率很低, 精确率也都低于 50%. 预测效果最好的“翻译”结点的精确率和覆盖率也分别只有 48%和 10% ($F = 0.16$). 显然, 48%的真阳性发现率不足以指导进行下一步的湿实验. 尽管深层预测与从头预测在总体上是不可比较的, 因为它们的适用范围不同, 然而, 二者在寻找一组相同的已知部分功能的蛋白质的精细功能的效率上是可以比较的. 本研究结果说明, 假设不使用已知部分功能的蛋白质的先验功能信息, 从头预测这些蛋白质精细功能的精确率和覆盖率一般会很低.

2.2 预测酵母与人类蛋白质的精细功能

考虑到湿实验的成本, 即使以覆盖率为代价, 产生高度可信的阳性预测也是有意义的 [7]. 因此, 我们只报道精确率高于 90%的预测结果. 如果一个蛋白质被预测到几个有父子关系的子功能类, 则只保留最深层的预测, 称为一个独立预测. 使用BioGRID[18]中的蛋白质互作数据和来自SGD[8](2006 年 9 月)的GO注释数据, 我们在生物过程、细胞组分和分子功能 3 个体系中分别为 674, 1024 和 138 个蛋白质做了 766, 1077 和 143 个独立预测. 即使使用注释率较低的人类蛋白质互作数据, 按照 90%的精确率, 在生物过程、细胞组分和分子功能 3 个体系中也可以分别为 1342, 2070 和 617 个人类蛋白质预测 1674, 2214 和 638 个精细功能(结果见http://www.systembiology.cn/supplementary/finerfunc.htm).

2.3 基于基因表达谱的预测

为了比较, 我们采用K近邻分类器 [2]评价使用表达谱数据进行深层预测的效果. 将注释到一个目标结点的有表达检测值的基因作为阳性样本, 预测空间中除预测对象外的其他基因作为阴性样本. 采用基因表达值之间的欧氏距离衡量两个基因之间的相

似性. 一个测试样本的类别由K个最近的训练样本中的多数标签决定. K近邻算法的简单决策规则不仅比许多复杂的算法例如支持向量机更容易理解, 而且一般可以保持使用基因表达信息预测的效率[2].

用Stress173数据和2006年5月份的GO注释数据, 采用留一法对从“蛋白质生物合成”到其子功能类的预测进行了评价. 结果显示, 根据基因表达谱数据, 除了到“翻译”功能类的预测之外, 到“蛋白质生物合成”的其他子结点的预测效果相对较差(F都小于 0.6). 这表明“翻译”子功能类对实验条件有较特异的应答[21], 而其他子功能类的表达特点较难被识别.

经留一法交叉验证, 从“蛋白生物合成”到“翻译”的预测的精确率和覆盖率分别达到了 87%和 79%(F = 0.83), 其中训练集合由 121 个阳性样本和 111 个阴性样本组成. 包含在 Stress173 数据库中的已知部分功能的 221 个蛋白质中有 181 个从“蛋白质生物合成”预测到“翻译”. 在 2006 年 9 月份更新的 GO 数据中, 181 个被预测的蛋白质有 163 个更新的功能注释和预测一致, 其余 18 个没有更新注释. 然后, 使用更新的数据进一步评价了从“蛋白质生物合成”到其子结点的预测效果, 也是只有到“翻译”的预测效果较好: 精确率和覆盖率分别达到了 93%和 90% (F = 0.92), 该训练集合阳性样本有 308 个, 阴性样本有 111 个. 对于 41 个在“蛋白质生物合成”中没有更新注释的蛋白质, 37 个基因在表达谱上有检测值, 其中 26 个可以被预测到“翻译”(表 2). 在蛋白质互作网络中, 这被深层预测的 26 个蛋白质中的 13 个在“蛋白质生物合成”功能类中有互作的邻居, 其中 12 个也被预测到了“翻译”功能. 最后, 我们发现从头预测到“翻译”的效果很差, 精确率和覆盖率分别为 45%和 12% (F = 0.19). 该训练数据集高度不平衡, 阳性样本和阴性样本分别为 121 个和 3976 个. 按照 F 指标, 通常的不平衡处理方法并不能使预测效果提高(图 3).

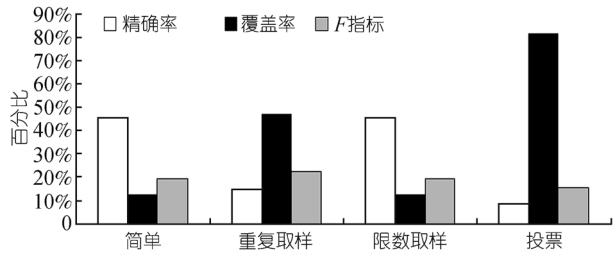


图 3 不平衡处理方法对预测效果的影响

“简单”示未经过不平衡处理的预测效果. 为了平衡训练样本中的阳性和阴性样本, “重复取样”方法是从小的类中随机抽取样本直到平衡, 而“限数取样”则是从大的类中随机抽取一个小的集合. “投票”是基于小分类器的多数投票原则, 每一个小分类器是通过从大类中抽取小的集合来构成平衡样本训练得到的. 对于从头预测算法, 这三种不平衡处理方法按照 F 指标不能够提高分类器预测效果

以上分析提示, 与使用蛋白质互作数据不同, 利用基因表达谱预测蛋白质(基因)的功能存在较大的局限性. 因为同一个预测空间中的子功能类倾向于表现出相近的表达模式, 除非其中一个子功能类(如本数据集中的“翻译”)对实验条件的应答有特异性, 用表达谱数据区分它们是困难的[15, 30~32].

3 讨论

深层预测算法能够通过使用许多功能描述粗浅的蛋白质已知功能信息, 可靠地预测它们更精细的功能. 而对于同一组蛋白质, 利用传统的从头预测算法寻找它们的精细功能, 其精确率通常不足以指导下一步的湿实验. 值得特别指出, 深层预测与从头预测这两种不同的预测策略是互补的: 从头预测算法能够为较多的功能未知及已知部分功能的蛋白质预测功能, 但是一般精确率较差; 而深层预测算法则适合预测已知部分功能的蛋白质, 利用其已知功能信息来更可靠地寻找它们更精细的功能.

一般地, 根据深层预测策略, 在功能特异的互作子网中, 可以使用所有的为从头预测发展的算法(或假设), 例如, 如果互作的蛋白质与功能相似的蛋白质集合互作, 则赋予它们相同的功能[33,34]; 或者通过最小化注释到不同的子功能类的蛋白质互作的数目来赋予蛋白质新的功能[35,36]. 本研究的深层预测算法中采用了通常使用的基于“连坐”(guilty by association)思想的大数规则[3]: 在功能子网中, 利用已知部分功能的蛋白质的邻居蛋白质的功能信息推测其精细功能. 该大数规则算法的基础假设是在一个功能类中协同工作的蛋白质有很高的机会互作. 按照功能类的可分性得分[37], 许多功能类在整个网络中表现为分割的子网证实了这一点. 本研究显示, 大量的功能子类可以被简单的大数规则很好地区分, 这也反映了在这些功能类中的蛋白质有较高的机会互作. 但是, 按照目前的蛋白质功能划分体系, 并非任何一个功能类中的蛋白质都倾向更紧密地互作. 例如, 本文提到, 在原来的GO功能体系中(2006年5月), 在子结点“甘露糖蛋白质的生物合成”中注释的蛋白质大多数与注释到“糖蛋白质的生物合成”的蛋白质互作. 尽管GO为计算分析蛋白质功能提供了一个很好的框

架，但是按照现有的蛋白质互作数据，一些功能类的边界仍然是模糊的[37]. 实际上，这些功能类的模糊定义造成的不确定性是许多预测方法效果不好的根本原因[6]. 因此，发展一个更为系统的蛋白质功能定义体系将会是以后工作的重点.

如从“蛋白质生物合成”的深层预测的结果所示，利用蛋白质互作数据，许多功能相近的子结点的蛋白质可以被很好的区分，这或许是因为成对的蛋白质互作数据能够捕捉到蛋白质功能的局部联系[6]. 而根据基因表达谱数据，只有对实验条件表现出不同表达模式的功能类才可以被预测，而没有被实验条件激活的功能类很难被识别[1,2,15]. 此外，在许多实验条件下，功能相关的基因倾向于表现出相似的表达模式[15,31]. 因此，即使使用深层预测算法，区分表达模式相似的子功能类通常也是困难的. Stress173数据的应用结果就说明了这一点. 但是，蛋白质互作数据也存在弱点，它不包含蛋白质互作发生的条件信息，一个蛋白质的邻居蛋白质可能参与几个信号通路[38]. 因此，为了提高预测的效果，通过贝叶斯网络等方法整合多种数据源是有价值的[4,39,40].

使用注释到同一个功能类中的蛋白质，可以过滤掉部分假阳性互作，因此基于它们的预测的可信度一般会有提高[11~14,41]. 然而，大量的假阳性互作仍然是影响预测效果的一个不利因素. 通过基因共表达来过滤互作数据是一个合理的方法[11]. 此外，因为目前蛋白质互作数据只覆盖了整个互作组的一小部分[6,7]，在预测空间中一些蛋白质直接的互作关系可能缺失. 对于被蛋白质互作数据不完全覆盖的功能类，可以进一步探讨通过间接的蛋白质互作、蛋白质互作子网的拓扑结构信息[30,42]和功能子类之间的语义相似性[43~45]等来寻找它们的精细功能.

参 考 文 献

1 Brown M P, Grundy W N, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA, 2000, 97(1): 262—267[DOI]

2 Kuramochi M, Karypis G. Gene classification using expression profiles: A feasibility Study. 2nd. IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, USA, 2001

3 Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol, 2000, 18(12): 1257—1261[DOI]

4 Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. Nucleic Acids Res, 2004, 32(21): 6414—6424[DOI]

5 孙景春，徐晋麟，李亦学，等. 大规模蛋白质相互作用数据的分析与应用. 科学通报, 2005, 50(19): 2055—2060

6 Jansen R, Gerstein M. Analyzing protein function on a genomic scale: The importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol, 2004, 7(5): 535—545[DOI]

7 Myers C L, Barrett D R, Hibbs M A, et al. Finding function: Evaluation methods for functional genomic data. BMC Genomics, 2006, 7: 187[DOI]

8 Dwight S S, Harris M A, Dolinski K, et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res, 2002, 30(1): 69—72[DOI]

9 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25(1): 25—29[DOI]

10 Tu K, Yu H, Guo Z, et al. Learnability-based further prediction of gene functions in Gene Ontology, Genomics, 2004, 84(6): 922—928[DOI]

11 Deng M, Sun F, Chen T. Assessment of the reliability of protein-protein interactions and protein function prediction. Pac Symp Biocomput, 2003: 140—151

12 Patil A, Nakamura H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC Bioinformatics, 2005, 6: 100[DOI]

13 Suthram S, Shlomi T, Ruppin E, et al. A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics, 2006, 7: 360[DOI]

14 Lin N, Wu B, Jansen R, et al. Information assessment on predicting protein-protein interactions. BMC Bioinformatics, 2004, 5: 154[DOI]

15 Mateos A, Dopazo J, Jansen R, et al. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res, 2002, 12(11): 1703—1715[DOI]

16 Chen J J, Tsai C A, Young J F, et al. Classification ensembles for unbalanced class sizes in predictive toxicology. SAR QSAR Environ Res, 2005, 16(6): 517—529[DOI]

17 Reguly T, Breitkreutz A, Boucher L, et al. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. J Biol, 2006, 5(4): 11[DOI]

18 Stark C, Breitkreutz B J, Reguly T, et al. BioGRID: A general repository for interaction datasets. Nucleic Acids Res, 2006, 34(Database issue): D535—D539[DOI]

19 Mishra G R, Suresh M, Kumaran K, et al. Human protein reference database—2006 update. Nucleic Acids Res, 2006, 34(Database issue): D411—D444[DOI]

20 Wu C H, Apweiler R, Bairoch A, et al. The Universal Protein Re-

source (UniProt): An expanding universe of protein information. Nucleic Acids Res, 2006, 34: D187—D191[DOI]

21 Gasch A P, Spellman P T, Kao C M, et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell, 2000, 11(12): 4241—4257

22 Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics, 2001, 17(6): 520—525[DOI]

23 Wang D, Lv Y, Guo Z, et al. Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. Bioinformatics, 2006, 22(23): 2883—2889[DOI]

24 Jiang T, Keating A E. AVID: An integrative framework for discovering functional relationships among proteins. BMC Bioinformatics, 2005, 6(1): 136[DOI]

25 Taher L, Rinner O, Garg S, et al. AGenDA: Homology-based gene prediction. Bioinformatics, 2003, 19(12): 1575—1577[DOI]

26 Wheeler D L, Barrett T, Benson D A, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2006, 34(Database issue): D173—D180[DOI]

27 Di Como C J, Arndt K T. Nutrients, via the Tor proteins, stimulate the association of Tap42 with type 2A phosphatases. Genes Dev, 1996, 10(15): 1904—1916[DOI]

28 Browne G J, Proud C G. Regulation of peptide-chain elongation in mammalian cells. Eur J Biochem, 2002, 269(22): 5360—5368[DOI]

29 Andjelkovic N, Zolnierowicz S, van Hoof C, et al. The catalytic subunit of protein phosphatase 2A associates with the translation termination factor eRF1. EMBO J, 1996, 15(24): 7156—7167

30 Chua H N, Sung W K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics, 2006, 22(13): 1623—1630[DOI]

31 Guo Z, Zhang T, Li X, et al. Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics, 2005, 6: 58[DOI]

32 张敏, 朱晶, 郭政, 等. 利用亚细胞位置特异的基因功能模块与表达调控网络识别疾病特征基因. 科学通报, 2006, 51(13): 1545—1551

33 Samanta M P, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci USA, 2003, 100(22): 12579—12583[DOI]

34 Okada K, Kanaya S, Asai K. Accurate extraction of functional associations between proteins based on common interaction partners and common domains. Bioinformatics, 2005, 21(9): 2043—2048[DOI]

35 Karaoz U, Murali T M, Letovsky S, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA, 2004, 101(9): 2888—2893[DOI]

36 Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol, 2003, 21(6): 697—700[DOI]

37 Yook S H, Oltvai Z N, Barabasi A L. Functional and topological characterization of protein interaction networks. Proteomics, 2004, 4(4): 928—942[DOI]

38 Han J D, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 2004, 430(6995): 88—93[DOI]

39 Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 2003, 302(5644): 449—453[DOI]

40 Troyanskaya O G, Dolinski K, Owen A B, et al. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci USA, 2003, 100(14): 8348—8353[DOI]

41 Lu L J, Xia Y, Paccanaro A, et al. Assessing the limits of genomic data integration for predicting protein networks. Genome Res, 2005, 15(7): 945—953[DOI]

42 Massjouni N, Rivera C G, Murali T M. VIRGO: Computational prediction of gene functions. Nucleic Acids Res, 2006, 34(Web Server issue): W340—W344[DOI]

43 Yu H, Gao L, Tu K, et al. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene, 2005, 352: 75—81[DOI]

44 Zhu M, Gao L, Guo Z, et al. Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities. Gene, 2007, 391(1-2): 113—119[DOI]

45 高磊, 李霞, 郭政, 等. 结合蛋白质互作与基因表达谱信息大范围预测蛋白质的精细功能. 中国科学C辑: 生命科学, 2006, 36(5): 441—450