从加权网络中预测蛋白质复合物*

汤希玮，李勇帆，胡秋玲
（湖南第一师范学院 信息科学与工程系，长沙 410205）

摘　要：实验产生的蛋白质相互作用数据不可避免地伴随着假阳性和假阴性，因而，基于蛋白质相互作用数据预测蛋白质复合物的计算方法天然具有较大的误差。为了弥补这种数据先天性不足，基因表达谱被结合进来，构造了新的加权蛋白质网络。为了验证网络的生物学意义，马尔可夫聚类算法被用于从加权与非加权网络中预测蛋白质复合物，预测到的复合物与基准复合物进行匹配分析。分析结果表明，加权网络比非加权网络具有更高的生物学意义。

关键词：蛋白质相互作用；基因表达谱；蛋白质复合物；匹配统计
中图分类号：TP301.6；Q31；Q71　　**文献标志码**：A　　**文章编号**：1001-3695(2012)12-4500-03
doi：10.3969/j.issn.1001-3695.2012.12.025

Prediction of protein complexes based on weighted network

TANG Xi-wei，LI Yong-fan，HU Qiu-ling
（*Dept. of Information Science & Engineering*，*Hunan First Normal University*，*Changsha* 410205，*China*）

Abstract：The computational methods predicting protein complexes in the protein-protein interaction network have a great error because of the high false positive rate and false negative rate of protein interaction data. To compensate for this，this paper constructed a new weighted protein interaction network via the integration of the protein-protein interaction and gene expression data. To evaluate the biological significance of the network，it used MCL algorithm to detect protein complexes in the weighted network and unweighted one. Matching analysis was performed between the derived complexes and benchmark complexes. The results show that the weighted network outperforms the unweighted network on biology.

Key words：protein-protein interaction；gene expression profiles；protein complexes；matching statistics

0　引言

高通量实验技术和计算预测方法产生了很多生物体的大规模生物网络。最近几年，为了获得细胞组织与功能的线索，越来越多的研究者致力于生物网络的分析。在全局网络的分析中，聚类（clustering）也许是最常见的方法，它们频繁地被应用于预测功能模块、蛋白质复合物和蛋白质功能[1~5]。大量的面向生物网络的聚类算法也先后出现[1,6~18]。这些算法将整个生物网络视为一张图，图中的节点代表蛋白质，节点之间的边代表蛋白质之间的相互作用；功能相似的蛋白质以蛋白质复合物的形式存在，对应图中紧密相连的子图。算法的目的是找出这些具有生物学意义的子图（即蛋白质复合物）。人们发现，通过分析算法预测到的复合物，尽管这些算法的性能有高下，但是它们都预测到了一些具有生物学意义的蛋白质复合物。

实际上，在预测蛋白质复合物的研究中，人们不得不接受一个尴尬的事实：数据源的可靠性有待提高。例如，在酿酒酵母的蛋白质相互作用（protein-protein interaction，PPI）数据中，存在数据的不完整性以及产生这些数据的生物实验带来的假阳性和假阴性。这是因为，有些瞬时的相互作用根本无法通过高通量实验技术捕捉到，而且由于实验环境的影响和实验条件的限制，不可避免地会产生虚假的相互作用。本文通过深入研究，发现同一物种的不同生物数据之间存在良好的互补性，能有力弥补单一生物数据的不足。例如，酿酒酵母的基因表达数据能有效弥补 PPI 数据不足，从而大大提高计算方法的精度。这一发现同时也带来一个如何整合不同数据源的问题。

本文利用皮尔逊相关系数（Pearson correlation coefficient，PCC）整合酿酒酵母的基因表达数据和 PPI 数据，成功地将基因表达谱携带的时间信息附加到蛋白质相互作用网络中，构建了新的加权网络。为了验证新网络的可靠性，采用著名的马尔可夫聚类 MCL 算法[11]，分别从酿酒酵母的 PPI 网络以及新的加权 PPI 网络中预测复合物，然后以基准蛋白质复合物为参照，对预测到的不同来源（加权 PPI 网络和非加权 PPI 网络）的蛋白质复合物分别进行匹配分析。

1　生物数据

1.1　蛋白质相互作用网络

在各种生物中，相对而言酿酒酵母的蛋白质相互作用网络是最完整的，而 DIP（database of interacting proteins）数据库[19]中酿酒酵母的 PPI 数据都是通过生物实验而不是计算方法获得的，因此，酿酒酵母的 PPI 数据较为可靠。本文采用该数据集进行研究，它包含了 5 093 个节点（蛋白质）和 24 743 条边（相互作用）。

收稿日期：2012-05-15；**修回日期**：2012-06-28　**基金项目**：湖南省教育厅科研资助项目（11C0281）；湖南省科技厅科技计划资助项目（2011GK3138，2010GK3049）；湖南省高校科技创新团队支持计划资助项目（湘教通[2010] 212 号）

作者简介：汤希玮（1973-），男，湖南安乡人，博士，主要研究方向为生物信息学（nudt_xiwei@126.com）；李勇帆（1959-），男，湖南邵阳人，教授，硕士，主要研究方向为信息技术学和生物信息学；胡秋玲（1976-），女，湖南常德人，主要研究方向为生物信息学.

1.2　基因表达谱

文献[20]通过严格的生物实验提取了 36 个不同时刻的酿酒酵母的基因表达谱,总共包括 6 777 个不同的基因。该数据集构成了一个 6 777 × 36 的矩阵,在本研究中被用于给酿酒酵母的 PPI 网络加权。

1.3　基准蛋白质复合物

CYC2008[21]包含了 408 个基准蛋白质复合物,它们通过给被预测的复合物提供匹配标准来评估 PPI 网络加权方法的成功与否。

2　评估方法与已知复合物匹配

为了评估加权 PPI 网络的可靠性,首先需要分析现存算法(如马尔可夫聚类 MCL)从该网络中预测出来的复合物的质量。文献[1]提出的与已知复合物匹配的方法被广泛采用。该方法采用重叠得分(overlap score,OS)描述预测到的复合物与基准复合物的匹配程度。*OS* 的计算公式为

$$OS = \frac{i^2}{a \times b} \tag{1}$$

其中:a 是指预测到的复合物中蛋白质数,b 是指已知复合物中蛋白质数,i 是二者的交集。*OS* 为 0 表示二者完全不匹配;*OS* 为 1 表示预测到的复合物与基准复合物完全重合。本文在 *OS* 的不同阈值区间计算衡量复合物质量的三个指标:灵敏度(sensitivity)、特异性(specificity)和二者的调和评价值(*f*-measure)。在此之前需要先描述真阳性、假阳性和假阴性。

真阳性(TP)是指 *OS* 大于某一阈值时预测到的复合物数;假阳性(FP)是指预测到的复合物的总数减去真阳性(TP)的差;假阴性(FN)是指不能匹配已知复合物预测到的复合物数。进而定义灵敏度(S_n)为[TP/(TP + FN)][1],特异性(S_p)为[TP/(TP + FP)][1];二者的调和平均值(*f*-measure)为

$$f\text{-measure} = \frac{2 \times S_n \times S_p}{S_n + S_p} \tag{2}$$

本文开发了一个 Perl 程序自动完成复合物的匹配分析和三个评价指标的计算。

3　算法描述

3.1　构建加权 PPI 网络

在生物信息学中,两个基因的相似性可以通过它们对应的表达谱的相似性刻画,而表达谱的相似性又可以通过计算两列表达谱的皮尔逊相关系数得出。因此,本文计算酿酒酵母 PPI 网络中每对相互作用的蛋白质的编码基因的皮尔逊相关系数,并将之作为该相互作用的权值。具体算法步骤如下:

a)从 PPI 数据集(网络)中,取第一对相互作用的蛋白质 *A* 和 *B*,然后在基因表达数据集中查询是否有这两个蛋白质编码的基因 *X* 和 *Y*,如果没有,就认为这两个基因不相关,设定它们之间的权值为 -1;否则进入 b)。

b)计算 *X* 和 *Y* 的皮尔逊相关系数 PCC。将这两个基因的表达谱(或表达序列)记为 $X = (x_1, x_2, \cdots, x_i, \cdots, x_{36})$,$Y = (y_1, y_2, \cdots, y_i, \cdots, y_{36})$,则 PCC 为

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

其中:$\bar{x}$ 表示基因 *X* 在 36 个时刻的表达值的平均值,$\bar{y}$ 表示基因 *Y* 在 36 个时刻的表达值的平均值。将计算出的 PCC 作为网络中 *A* 和 *B* 之间的边的权值。

c)遍历 PPI 网络中的每一条边,循环执行 a)b),直到给每条边添加了权值。由于 PCC 的值在 -1 ~ 1,为方便处理,将所有权值分别加 1,消除为负的权值。最后 PPI 网络中的每一条边都获得一个权值,从而构造了加权 PPI 网络。

3.2　预测蛋白质复合物

为了评价加权网络是否成功,需要从网络中预测具有相同功能的蛋白质构成的复合物。因为马尔可夫聚类 MCL 算法[11]能将网络中全部的蛋白质归类到不同的密度子图(对应不同蛋白质复合物)中,所以本文采用 MCL 从酿酒酵母的加权蛋白质网络中预测蛋白质复合物;为了比较,也将 MCL 运用于非加权网络预测蛋白质复合物。

4　实验结果比较和分析

将预测出来的两类复合物(一类来自加权网络,另一类来自非加权网络)分别进行匹配分析和 GO 功能富集分析。表 1 和 2 显示了匹配分析的结果。在这两个表中,P_c 表示从网络中预测出来的全部复合物数,MP_c 表示 P_c 中能与基准复合物匹配的复合物数,MK_c 表示基准复合物集合中能与预测到的复合物匹配的复合物数,S_n 表示灵敏度,S_p 表示特异性,*f*-measure 表示 S_n 和 S_p 的调和平均值。例如,表 1 第三行表示,MCL 总共从加权网络中预测到了 947 个复合物,在重叠得分 *OS* 不小于 0.2 的条件下,有 187 个预测到的复合物能与基准复合物集合中 207 个已知复合物匹配,灵敏度为 0.482,特异性为 0.197,*f*-measure 为 0.280。类似地,表 2 第三行表示,MCL 总共从非加权网络中预测到 918 个复合物,在同样的重叠得分条件下,有 163 个预测到的复合物与基准复合物集合中 180 个已知复合物匹配,灵敏度为 0.417,特异性为 0.178,*f*-measure 为 0.249。对照表 1 和 2 可以发现,在不同的重叠得分条件下,加权网络的各项评价指标都比非加权网络的要高。这表明与非加权网络相比,同样的算法在加权网络上预测到的复合物的质量较高,从而说明加权网络的生物学意义比非加权网络的要显著。

表 1　加权网络中预测出来的复合物的情况

条件	P_c	MP_c	MK_c	S_n	S_p	*f*-measure
$OS \geqslant 0.1$	947	306	313	0.763	0.323	0.454
$OS \geqslant 0.2$	947	187	207	0.482	0.197	0.280
$OS \geqslant 0.3$	947	131	146	0.333	0.138	0.196
$OS \geqslant 0.4$	947	106	116	0.266	0.112	0.158
$OS \geqslant 0.5$	947	86	90	0.213	0.091	0.127
$OS \geqslant 0.6$	947	61	63	0.150	0.064	0.090
$OS \geqslant 0.7$	947	33	35	0.081	0.035	0.049
$OS \geqslant 0.8$	947	26	26	0.064	0.027	0.038
$OS \geqslant 0.9$	947	18	18	0.044	0.019	0.027
$OS \geqslant 1$	947	17	17	0.042	0.018	0.025

表 2　非加权网络中预测出来的复合物的情况

条件	P_c	MP_c	MK_c	S_n	S_p	*f*-measure
$OS \geqslant 0.1$	918	280	298	0.718	0.305	0.428
$OS \geqslant 0.2$	918	163	180	0.417	0.178	0.249
$OS \geqslant 0.3$	918	118	126	0.295	0.129	0.179
$OS \geqslant 0.4$	918	94	102	0.235	0.102	0.143
$OS \geqslant 0.5$	918	76	80	0.188	0.083	0.115
$OS \geqslant 0.6$	918	46	47	0.113	0.050	0.069
$OS \geqslant 0.7$	918	26	26	0.064	0.028	0.039
$OS \geqslant 0.8$	918	19	19	0.047	0.021	0.029
$OS \geqslant 0.9$	918	13	13	0.032	0.014	0.020
$OS \geqslant 1$	918	13	13	0.032	0.014	0.020

5 结束语

本文将基因表达谱成功地结合到蛋白质相互作用网络中，构建了高可靠性的加权蛋白质相互作用网络。为了验证加权网络的生物学意义是否比非加权网络更显著，本文用著名的图聚类算法 MCL 从两种网络中预测蛋白质复合物，并对预测出的复合物执行严格的匹配分析。匹配分析的结果显示加权网络有更高的生物显著性。本文的研究验证了“同一物种的不同生物数据之间具有良好的互补性”这一结论。未来将使用数学模型沟通不同数据源，从而结合多数据源对生物信息学的其他研究领域如关键基因预测、致病基因预测和功能模块预测等展开研究。

参考文献：

[1] BADER G D, HOGUE C W. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatics, 2003, 4(2): 59.

[2] HARTWELL L H, HOPFIELD J J, LEIBLER S, *et al*. From molecular to modular cell biology [J]. Nature, 1999, 402(6761): C47-C52.

[3] PEREIR-LEAL J B, ENRIGHT A J, OUZOUNIS C A, *et al*. Detection of functional modules from protein interaction networks [J]. Proteins, 2004, 54(1): 49-57.

[4] RIVES A W, GALITSKI T. Modular organization of cellular networks [J]. Proceedings of National Academy of Sciences, 2003, 100(3): 1128-1133.

[5] SPIRIN V, MIRNY L A. Protein complexes and functional modules in molecular networks [J]. Proceedings of National Academy of Sciences, 2003, 100(21): 12123-12128.

[6] ALTAF-UL-AMIN M, SHINBO Y, MIHARA K, *et al*. Development and implementation of an algorithm for detection of protein complexes in large interaction networks [J]. BMC Bioinformatics, 2006, 7(1): 207.

[7] BLATT M, WISEMAN S, DOMANY E. Superparamagnetic clustering of data [J]. Physical Review Letters, 1996, 76(18): 3251-3254.

[8] BRUN C, CHEVENET F, MARTIN D, *et al*. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network [J]. Genome Biology, 2004, 5(1): 6.

[9] CHEN Jing-chun, YUAN Bo. Detecting functional modules in the yeast protein-protein interaction network [J]. Bioinformatics, 2006, 22(18): 2283-2290.

[10] COLAK R, HORMOZDIAR I F, MOSER F, *et al*. Dense graphlet statistics of protein interaction and random networks [C]//Proc of Pacific Symposium on Biocomputing. 2009: 178-189.

[11] ENRIGHT A J, Van DONGEN S, OUIOUNIS C A, *et al*. An efficient algorithm for large-scale detection of protein families [J]. Nucleic Acids Research, 2002, 30(7): 1575-1584.

[12] GEORGII E, DIETMANN S, UNO T, *et al*. Enumeration of condition-dependent dense modules in protein interaction networks [J]. Bioinformatics, 2009, 25(7): 933-940.

[13] KING A D, PRZULJ N, JURISICA I, *et al*. Protein complex prediction via cost-based clustering [J]. Bioinformatics, 2004, 20(17): 3013-3020.

[14] LOEWENSTEIN Y, PORTUGALY E, FROMER M, *et al*. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space [J]. Bioinformatics, 2008, 24(13): 41-49.

[15] NAVLAKHA S, SCHATI M C, KINGSFORD C, *et al*. Revealing biological modules via graph summarization [J]. Journal Computational Biology, 2009, 16(2): 253-264.

[16] PALLA G, DERENYI I, FARKAS I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818.

[17] SAMANTA M, LIANG S. Predicting protein functions from redundancies in large-scale protein interaction networks [J]. Proceedings of National Academy of Sciences, 2003, 100(22): 12579-12583.

[18] SHARAN R, SUTHRAM S, KELLEY R M, *et al*. Conserved patterns of protein interaction in multiple species [J]. Proceedings of National Academy of Sciences, 2005, 102(6): 1974-1979.

[19] Database of interacting proteins [DB/OL]. http://dip.doe-mbi.ucla.edu/dip/Download.cgi? SM=7/.

[20] TU B P, KUDLICKI A, ROWICKA M, *et al*. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes [J]. Science, 2005, 310(5751): 1152-1158.

[21] PU S, WONG J, TURNER B, *et al*. Up-to-date catalogues of yeast protein complexes [J]. Nucleic Acids Research, 2009, 37(3): 825-831.

(上接第 4481 页)验证了该方法的可行性和有效性，并与基于距离的异常挖掘方法作出对比，突出了该方法保证结果正确全面的前提下，时间复杂度大大降低的优势。目前，符号化方法在水文时间序列异常挖掘领域的研究还不多，今后的研究方向包括对算法的改进，以及该类方法在多维水文时间序列异常挖掘中的运用。

参考文献：

[1] 艾萍，倪伟新. 我国水文数据挖掘技术研究的回顾与展望 [J]. 计算机工程与应用，2003，39(28)：13-17.

[2] LENG Ming-wei, CHEN Xiao-yun, LI Long-jie. Variable length methods for detecting anomaly patterns in time-series [C]//Proc of International Symposium on Computational Intelligence and Design. 2008.

[3] 林森. 时间序列异常检测的研究与应用 [D]. 南京：河海大学，2008.

[4] 余跃. 水文时间序列异常模式挖掘方法的研究与应用 [D]. 南京：河海大学，2011.

[5] DASGUPTA D, FORREST S. Novelty detection in time series data using ideas from immunology [C]//Proc of the 5th International Conference on Intelligent Systems. 1996: 82-87.

[6] KEOGH E, LONARDI S, CHIU B. Finding surprising patterns in a time series database in linear time and space [C]//Proc of SIGKDD'02. 2002: 23-26.

[7] MA Jun-shui, PERKINS S. Time-series novelty detection using one-class support vector machines [C]//Proc of International Joint Conference on Neural Networks. 2003: 1741-1745.

[8] KEOGH E, LIN J, FU A. HOT SAX: finding the most unusual time series subsequence: algorithms and applications [C]//Proc of the 5th IEEE International Conference on Data Mining. 2005: 226-233.

[9] SHAHABI C, TIAN Xiao-ming, ZHAO Wu-gang. TSA-tree: a wavelet-based approach to improve the efficiency of multi-level surprise and trend queries [C]//Proc of the 12th International Conference on Scientific and Statistical Database Management. 2000.

[10] LKHAGVA B, SUZUKI Y, KAWAGOE K. New time series data representation ESAX for financial applications [C]//Proc of the 22nd International Conference on Data Engineering Workshops. 2006: 17-22.

[11] LIN J, KEOGH E, LONARDI S, *et al*. A symbolic representation of time series, with implications for streaming algorithms [C]//Proc of DMKD'03. 2003: 2-11.

[12] HUYNH T, DUONG T. Time series discord discovery based on iSAX symbolic representation [C]//Proc of the 3rd International Conference on Knowledge and Systems Engineering. 2011: 11-18.