基于关联基因本体论注释的蛋白质相互作用预测 *

张 茜[1,2] 王敬泽[1]**

([1]中国科学院动物研究所，生物膜与膜生物工程国家重点实验室，北京 100080；
[2]中国科学院研究生院，北京 100039)

摘要 细胞中的生理活动主要是通过蛋白质 - 蛋白质之间的相互作用来调控完成. 详尽细致的蛋白质 - 蛋白质相互作用网络的解析对于理解细胞中复杂的调控、代谢和信号通路有重要的意义. 近年来，关于新的蛋白质 - 蛋白质相互作用预测领域进展快速，这里，利用贝叶斯算法结合关联的 GO (Gene Ontology)，来预测蛋白质的相互作用. 利用非冗余的蛋白质相互作用数据来观察 GO 对的特性，得到 GO 关联的概率. 通过阳性的和阴性的标准对照数据证实这个新方法可以很好地区别这两类不同的数据，显示出较好的灵敏度和非常低的假阳性预测率. 通过与已知的高通量的实验数据比较，这个方法具有灵敏度高、速度快的优点. 而且，运用这个新方法可以提供一些新的关于细胞内蛋白质之间相互作用的信息，为进一步的实验提供理论依据.

关键词 蛋白质 - 蛋白质相互作用，关联的，基因本体论, 贝叶斯(Bayesian)
学科分类号 Q811.4

蛋白质 - 蛋白质相互作用 (protein-protein interaction，PPI) 是大多数生物细胞进行各种生化过程的基础，例如通过物理性的结合，暂时性的磷酸化，SUMO 化[1]等机制组成多蛋白质的复合物. 随着基因组和蛋白质组计划的不断发展，大量关于基因和蛋白质的数据可以被检索，从而通过计算的方法绘制 PPI 网络，详尽的 PPI 网络反过来又可以揭示生物基因组和蛋白质组中包含的多方面、复杂的细胞内功能的关联[2].

多年来的研究表明，PPI 网络可以通过很多实验方法来验证，这些方法大多需要费时费力的试验工作而且准确性不高[3]，比如芽殖酵母的 PPI 网络已经得到了很大程度的解析 [4,5].

为了提高效率和方便检索, 许多预测 PPI 的计算方法已陆续被开发，例如基于基因的融合和分裂 [6]，基因在基因组上顺序的保守或基因邻居关系[7,8]，进化谱[9]，共表达或者相关联的 mRNA 表达[10]而研究开发出的预测方法. 近年来有研究结果表明关联的序列特征，比如相互作用的蛋白质结构域可以被用来作为预测蛋白质相互作用的一种策略[11]，另外，潜在的蛋白质结构域互相之间的物理结合也可以通过 PPI 的数据进行推测[12].

随着计算技术的飞速发展，在计算的指导下进行实验验证是目前 PPI 研究中的一个重要趋势. 例如，蛋白关联谱算法(protein correlation profiling，PCP algorithm) 被用来开发以帮助中心体蛋白质组 (centrosome proteomics) 的研究并已经得出一些有价值的猜想[13]. 在 PPI 预测研究中，整合各类相关数据对进一步的分析非常重要[14]. 不同的实验数据通过整合得出的 PPI 网络可以得到更多精确度高的有价值的结果. 以前的结果证实，不仅全基因组可以通过整合的功能关联网络进行注释 [15]，即使是“纯粹的预测”也可以为进一步的实际试验提供更多的信息[16].

以前的研究表明，显著关联的序列特征 (潜在的相互作用结构域) 可以用来预测 PPI[11]，反之，利用 PPI 数据也可以推测潜在的结构域 - 结构域之间的相互作用 [12]. 既然 PPI 可以通过 GO 注释被用来预测蛋白质的功能 [15,17]，我们猜测是否可以通过蛋白质之间的功能关联来预测 PPI. 为了验证这个想法，我们使用一种被报道过的 Naïve Bayesian 算法[18] 来表明关联的 GO 对可以作为预测 PPI 的一种手段. 通过处理公用的数据库，我们得到了芽殖酵母的 GO 注释, 然后计算这些 GO 对的先验概率 (pre-test probability). 利用已有的被称为“黄金标准集”(gold standard)的阳性对照数据 (positive control data) 和阴性对照数据 (negative control data)，这项工作表明关联的 GO 可以作为一种新颖的方法来预

*国家自然科学基金资助项目(30171071).
** 通讯联系人.
Tel: 010-62551668, E-mail: wangjz@ioz.ac.cn
收稿日期：2004-12-01, 接受日期：2005-01-31

测 PPI，而且它的灵敏度可以与已知的大规模实验数据相媲美. 另一方面, 这种方法可以很好地区别阳性数据和阴性数据从而得到很低的假阳性率.

1 材料和方法

1.1 材料

1.1.1 训练数据集. 用做训练数据集的 PPI 数据从两个公用数据库中得到, Database of Interacting Proteins (DIP) (Oct. 2003) 和 MIPS Comprehensive Yeast Genome Database (CYGD) PPI (Oct. 2003). 在本工作中我们只使用了酵母的数据. DIP 包含了 15 164 对相互作用, 而 CYGD 包含 11 862 对. 非冗余的蛋白质相互作用对总共是 17 013 对.

1.1.2 “黄金标准集”和检验数据.

为了验证方法，我们使用已经报道的被称为“黄金标准集”的阳性和阴性的对照数据 [19]. 阳性对照数据包含从同样 MIPS 复合物中抽提出来的对, 阴性对照数据包含从不同亚细胞定位的蛋白质对.

为了比较，我们也随机产生 100 000 伪相互作用对 (pseudo-PPI) 作为检验数据, 这其中可能包含一些真的相互作用, 但占较小的比例.

1.1.3 高通量实验数据集. 我们把结果与高通量的数据相对比，包括两组体内(*in vivo*) pull-down 的数据[20, 21]，和两组体外酵母双杂交的数据[22, 23]. 表 1 中列出了这项工作中所有的 PPI 的相关信息.

1.1.4 蛋白质的基因本体论 GO 注释.

基因本体论联合会 [24] (Gene Ontology Consortium) 所建立的数据库，旨在建立一个适用于各个物种，对基因和蛋白质功能进行限定和描述的语言词汇标准，这个标准可以随着研究的不断深入而更新. 这里我们将 GO 注释映射到我们的非冗余 PPI 数据集上.

首先将从 DIP 和 MIPS 中得到的芽殖酵母总蛋白的数据映射到 SWISS-PROT 和 TrEMBL 数据库，从而得到蛋白质的标准标识符. 然后又参照 SGD External Links 手工检验了每个蛋白质的信息，以保证所得到的蛋白质标识符正确无误.

接下来，我们从 European Bioinformatics Institute (EBI)的 GOA[25]的 UniProt 数据库下载(ftp://ftp.ebi.ac.uk/pub/databases/GO/)了包含所有 GO 注释的数据. 另外，将这些数据中的 GO 注释映射到我们的非冗余 PPI 数据集上.

在分析中去除了 3 个 GO：GO:0000004 (biological process unknown)，GO:0008372 (cellular component unknown)，GO:0005554 (molecular function unknown)，因为这些未知的和不明确的注释将会干扰我们的分析. 那些没有 GO 详细注释过的蛋白质及其作用对被去除掉以后，我们得到一套含有 14 565 PPI 对和 3 992 个蛋白质的数据，同时在所用的数据中总的 GO 注释的数目是 2 425 条.

1.2 算法

这项工作采用了 Naïve Bayesian 算法. 因为根据推测，如果两个蛋白质 P_i，P_j 可以产生相互作用，它们分别被 m，n GO 条目注释，那么它们应该功能上相互关联，也就是两个 GO 条目对相同的功能有注释. 那么，P_i 和 P_j 是相互作用对的概率为：

$$P(I_{ij}=1)=1-\prod_{(G_{m'},G_{n'})\in(P_i\times P_j)}(1-P(G_{m'n'}=1))$$

I_{ij} =1: 蛋白质 P_i 和 P_j 是相互作用对；$G_{m'n'}$= 1: Gene Ontology (GO) 条目 $G_{m'}$ 和 $G_{n'}$是功能相关的；m'，n': $m' \in (1 \sim m)$, $n' \in (1 \sim n)$；$(G_{m'},G_{n'}) \in (P_i \times P_j)$: GO 对 $(G_{m'},G_{n'})$ 包含于 $P_i \times P_j$.

$P(G_{m'n'}=1)$ 被认为是先验概率 (pre-test *probability*)，它可以从训练现有的实验数据中得到. 计算先验概率的公式如下：

$$P(G_{mn}=1)=\frac{Int_{mn}}{N_{mn}}$$

Int_{mn}：相互作用蛋白对中包含 (G_m, G_n)的数目；N_{mn}：可能的蛋白对中包含 (G_m, G_n)的数目.

下面将列举实例来说明先验概率的计算方法. 例如在计算结果中，两个 GO 标注 GO:0005685 (snRNP U1) 和 GO:0005686 (snRNP U2)，在芽殖酵母中共有 10 个蛋白质被 GO:0005685 注释，11 个蛋白质被 GO:0005686 注释. 因此 N_{mn} = 10×11=110. 在我们的非冗余 PPI 数据中，共有 16 对蛋白质分别被 GO:0005685 和 GO:0005686 注释，因此 Int_{mn}= 16. 因此在本例中，可以计算，$P(G_{m'n'}= 1)= 16/110$.

2 结　果

2.1 关联的 GO 可以被用作预测蛋白质相互作用的一种方法

工作中所应用到的总 PPI 数据的相关信息已在表 1 中列出. 将两个公共 PPI 数据库合成一个非冗余数据集用来做为训练数据集，通过计算每个 GO 对的先验概率. 事实上很多 GO 对的先验概率为 0，所以，GO 关联性的空间非常稀疏.

Table 1 Information of the total data set we use

Data type	Data set	# Protein pairs	Used for
Training data	CYGD (MIPS)	11 862	Training
Training data	DIP	15 164	Training
Gold standards	(Positive) Protein pairs in the same MIPS complex	8 617	Testing
Gold standards	(Negative) Protein pairs with different subcellular location	2 705 844	Testing
Testing data	Random PPI	100 000	Testing
High-throughput interaction data	*In vivo* pull-down (Gavin, *et al.*)	31 304	Comparison
High-throughput interaction data	*In vivo* pull-down (Ho, *et al.*)	25 333	Comparison
High-throughput interaction data	Yeast two-hybrid (Uetz, *et al.*)	981	Comparison
High-throughput interaction data	Yeast two-hybrid (Ito, *et al.*)	4 393	Comparison

From public PPI databases CYGD and DIP, we generate a non-redundant PPI data set with the number of 17 013. Both positive and negative control data sets are used for testing. And we also generate a random data set containing 100 000 PPI pairs. Four high-throughput PPI data sets are chosen for comparison.

接下来用四组数据来检验这个方法，包括训练使用的数据集，随机伪相互作用对，阳性对照和阴性对照. 至于随机伪相互作用对，是指从随机挑出2 个芽殖酵母蛋白形成的 100 000 对组合. 很显然，这些数据中包含极少量真正的 PPI， 但是它的预测结果预期应当不同于真正的 PPI. 我们采用了以前报道过的[19]阳性和阴性对照，阳性对照包括在同一个 MIPS 复合物中的蛋白质对，而阴性对照中的蛋白质则是有着不同的亚细胞定位的蛋白质对. 因为应用真正的 PPI 作为训练的数据，阳性对照的预期结果与训练的数据相似，而阴性对照和随机伪相互作用对的数据结果和阳性对照和训练的数据显著不同.

图 1 中列出了结果的对比. 很明显，训练数据集准确性最高，当概率值 (probability score) 大于 0.5 的时候，阳性对照的曲线与训练数据的很相似. 有趣的是，阴性对照的曲线位于最近底部， 这表明用这种方法不会产生太多的假阳性结果. 因为随机伪相互作用对中包含极少的真正的 PPI，我们发现当概率值低于 0.2 的时候，很难区分阳性对照和我们的随机 PPI，这表明，即使在黄金标准集阳性对照中，仍然存在相当的假阳性，当概率值大于 0.3 的时候，阳性对照的曲线趋于平滑，即使是概率值大于 0.9 的时候，本方法仍能够准确地预测约 20% 相互作用.

图 1 显示本实验的方法能够将真的 PPI 和假的 PPI 区分开来，当概率值大于 0.5 的时候，这个方法将阴性对照预测为阳性得分的可能性为 0. 所以

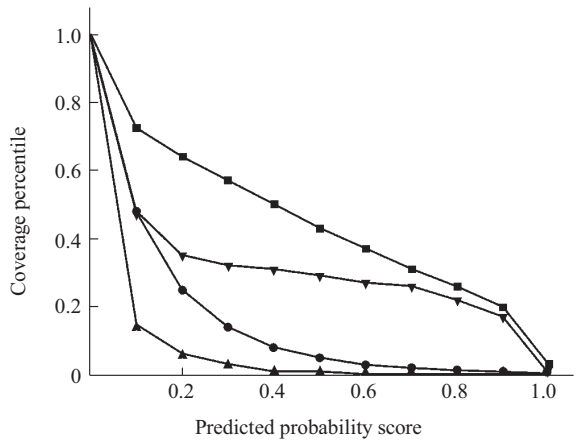

Fig.1 The curves of predicted accuracy

■—■: DIP & CYGD; ●—●: Random 10K; ▲—▲: L_neg; ▼—▼: MIPS_complex. We use four data sets to compare the predicted accuracy: our training data set, positive control (MIPS complex), random PPI and negative control. It's not strange that our training data get the highest rank. But it's clear that the positive control is much like our training data when the Predicted Probability Score is greater than 0.5. And our method predicts both the negative and random PPI with poor hits. So our method has a much satisfying sensitivity with a much low false positive rate.

认为相关的 GO 可以作为 PPI 预测的一种方法，而且可以得到令人满意的敏感度和很低的假阳性率. 在这篇文章中，我们任意选择了 0.5 作为概率值的一个界限，来做进一步的讨论. 以这个界限来说，能够对我们的训练数据集和阳性对照数据分别有约 50% 和 30%的阳性预测率. 然而对与随机伪相互作用对和阴性对照，仅仅有 5% 和 1%的假阳性率.

2.2 与高通量实验数据作比较

为了对比预测准确性，一些高通量实验产生的试验数据被用来做比较，同时阳性对照被选择用来作为黄金标准集. 两组数据从基因组层面的 *in vivo* pull-down[20,21]得来，也就是用已深刻研究过的蛋白质作为饵 (bait) 来钓出与之结合的蛋白质. 另外两组数据[22,23]是不同条件下大规模的酵母双杂交实验产生的.

表 2 中大规模实验数据的灵敏度计算是根据之前报道的方法[19]，将大规模实验的数据对应到阳性对照数据上，能与阳性对照数据对应的认为是对的，反之认为是大规模实验的疏漏或者假阳性结果. 因为一般认为理想情况下，基因组范围的大规模的 PPI 实验是能够充分地发现所有可能的相互作用，但是因为受实验条件限制，比如灵敏度不高，假阳性结果太多，因此高通量实验得到的结果中，常常含有大量的假阳性数据. 对于表 2 中列举的假阳性率 (false positive rate，FPR) 结果是根据阴性对照数据计算的. 而灵敏度数据则是根据阳性对照数据计算，同时列出了四组高通量实验数据的灵敏度数据作为比较.

Table 2 Comparison to high-throughput data

Testing & comparison	Cut-off	Sensitivity	FPR
Correlated GO	0.1	0.47	0.14
Correlated GO	0.2	0.35	0.06
Correlated GO	0.3	0.32	0.03
Correlated GO	0.4	0.31	0.01
Correlated GO	0.5	0.29	0.01
Gavin, *et al*.	—	0.21	—
Ho, *et al*.	—	0.05	—
Uetz, *et al*.	—	0.01	—
Ito, *et al*.	—	0.006	—

In order to test our method, the results predicted were in comparison to the four high-throughput data sets published, and the accuracy was evaluated using the positive control sets. When the probability score is greater than 0.1, a higher sensitivity of 47% was got with our method. And when we adopted a much stringent threshold score at 0.5, the false positive rate is about equal to zero, while the sensitivity is still 29%, which greater than the experimental hits.

在 MIPS 复合物中含有 8 617 个蛋白质，我们预测这些 PPI 对从而得到表 2 中的结果. 同时，这四组大规模的 PPI 数据与阳性控制数据也做了对比，这些比较在表 2 中列出. 当概率值大于 0.1 的时候，灵敏度为 47%，当 0.5 被作为一个起始值 (Cut-off) 的时候，灵敏度约为 30%，这些都高于高通量的数据得出的灵敏度. Gavin 等[20] 得到的最高的灵敏度是 21%，这表明本实验采用的方法在准确率和效率方面都非常有效. 而且，本方法能够通过结合越来越多的高质量的实验证实的 PPI 数据来增强自己的预测能力. 然而在所有芽殖酵母的 PPI 被准确阐明之前，比较我们这种方法与高通量的数据之间特异性的大小还难以进行.

2.3 推导芽殖酵母蛋白的蛋白质相互作用信息

既然关联的 GO 注释能被用来预测 PPI，那么这种方法能够对细胞 PPI 的关系提供更多的，尚未被阐明的有洞察力的和新颖的知识吗？

为了解决这个问题，我们查阅了最近的关于端粒 PPI 和蛋白质复合物的文献. 端粒是指在染色体末端的端粒 DNA 和与它结合的蛋白质复合物[26]，端粒 DNA 和端粒蛋白复合物之间在 DNA 复制和一条端粒 DNA 链加长反应中起重要作用，这一功能过程是维护基因组所必需的，退行性的过程可以使端粒 DNA 变短. 有报道称端粒蛋白复合物可以作为抗癌药物的靶向[27].

有些药物已经被开发设计用来阻止端粒复合物减短端粒 DNA. 但是，对于端粒复合物个体成员的数目以及它们之间的相互作用还不是很清楚. 参照文献[28]，我们找到了 12 个被详细研究过的蛋白质，它们很有可能组成蛋白质复合体或者分子机器来执行特定的过程. 通过在公共的 PPI 数据库中搜索，我们只找出了 14 对实验证实过的相互作用对，在图 2 中列出. 以 0.5 作为限制值，用我们的方法预测在这 12 个蛋白质之间有 48 对新的相互作用，用这种方法只漏掉了 2 对真正的相互作用. 有趣的是，实验性的相互作用显示这些蛋白质形成两个独立的亚复合体，其中一个包括 STN1，TEN1，TLC1，CDC13，EST1 和 EST2，另外一个包括 RIF1，RAP1，RIF2，SIR4，SIR2 和 SIR3. 通过预测，我们认为实际上这两个亚复合体之间存在相互的对话来形成一个独立的蛋白质复合体，这对实验室工作者有一定的启发意义.

另外，芽殖酵母中的 IPL1 是一个很重要的有丝分裂激酶[29]，在动点 (kinetochore) 和微管之间的动力态链接中起主导作用[30]. 失去生物极性的突变型 IPL1 将不能定位到动点上. IPL1 在胞质分裂期也很重要，在线虫和果蝇的 IPL1 同源物被抑止后将导致有丝分裂后期和末期的异常. DAM1 蛋白是

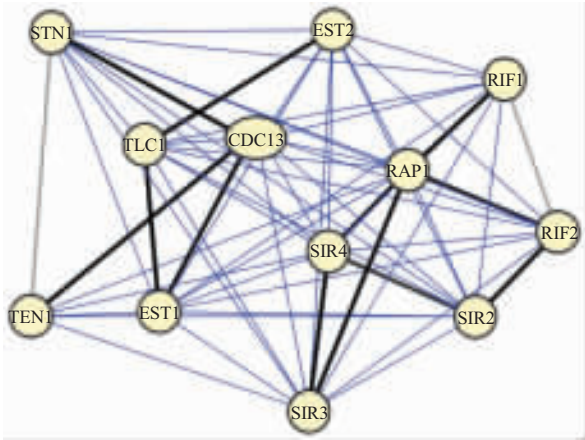

Fig.2 The PPI network of telomerase complex

From literature and public PPI database, we find that there exist twelve telomeric proteins which will form two separate sub-complexes with 14 verified PPI. And our prediction results provided 48 additional potential PPI, with only two missed. From prediction results, we suppose that these proteins may crosstalk with each other and form an integrated protein complex.

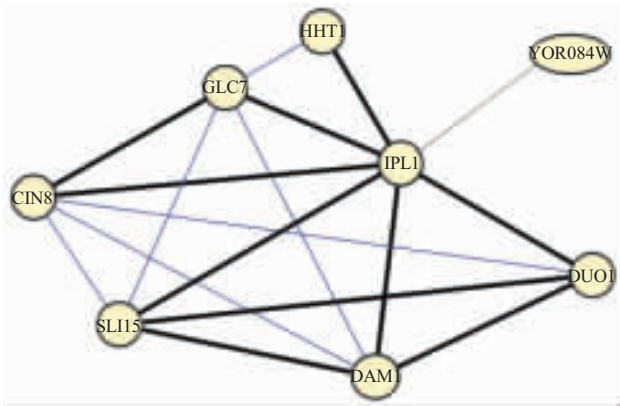

Fig.3 The PPI network of mitotic kinase IPL1 with its binding partners

IPL1 can interact with seven other proteins with eleven verified PPI. Our method can prediction ten of eleven PPI, furthermore, six probable new PPI were predicted using our approach. Since CIN8 can interact with IPL1 during mitosis process, while IPL1 and SLI15 can form a protein complex, it's quite possible that CIN8 interacts with SLI15.

IPL1的一个关键底物，在细胞周期的中期与之相互作用并被其磷酸化. IPL1 还与 SLI15 相互作用，结合形成一个暂时性的染色体过客复合物 (chromosome passenger complex) 来协助染色体的两极定向. 我们搜索了已存在的相互作用数据，发现 IPL1 可能与其他 7 个蛋白质相互作用. 这 8 个蛋白质之间的已经实验验证的相互作用数是 11. 用我们的方法能够证实其中 10 个为阳性的，只漏掉了其中之一，如图 3 所示，概率值的阈值是 0.5. 而且我们还预测了 6 个新的相互作用. 在图 3 中，我们发现动点蛋白 CIN8 在细胞分裂中能够与 IPL1 相互作用，IPL1 与 SLI15 形成一个复合物，所以很有可能的是 SLI15 也与 CIN8 相互作用. 这项预测对于进一步的实验很有指导意义.

3 讨 论

在本文中提出关联的 GO 注释可以作为预测蛋白质之间相互作用的一种方法. 从公共的 PPI 数据库中，得到了包含 17 013 对非冗余相互作用训练数据集. 将这些数据映射到 GO 注释上，并去除无用的信息. 在这项工作中， Naïve Bayesian 算法被用来计算每个 GO 对的先验概率 (pre-test probability)，以前报道过的[19]阳性对照数据和阴性对照数据也被当作“黄金标准集”来检验. 同时，一组随机伪相互作用数据被作为检验比较数据. 从图 1 中可以看出本实验采用的方法可以正确地预测阳性对照数据和训练数据集. 而随机伪相互作用数据和阴性对照数据的阳性预测率则很低. 所以这种方法有令人满意的灵敏度和相当低的假阳 性率.

把本实验的 *in silico* 预测方法和最近的高通量的相互作用数据相对比，即使在概率值 =0.5 这样严格的阈值条件下，这个方法仍然有 29% 的灵敏度，比 Gavin 等[20]采用的方法要高出 21%. 可以相信，当更多、更全面、更可靠的相互作用数据加入到训练数据集中后，准确性将得到相应的提高. 而且，因为本实验方法采用的是贝叶斯算法，因此很容易将这个方法与其他方法结合起来.

本文的工作也对细胞内蛋白质的相互作用和关系提供了有潜在指导意义的信息，文章列举两个例子加以说明. 第一是端粒蛋白相互作用预测，从现有的相互作用数据来看，端粒蛋白相互作用形成两个独立的蛋白质复合体，但是我们的工作表明，这些蛋白质可能相互之间密切作用而形成单一的复合物. 另外一个例子就是有丝分裂激酶 IPL1 在有丝分裂过程中与 CIN8 相互作用，同时 IPL1 与 SLI15 也形成复合体. 通过计算，我们推测 SLI15 很有可能与 CIN8 相互作用. 这些预测对于未来的具体实验室研究具有重要的指导意义.

综上，本文提出了一种新的可以用来作为 PPI 预测的方法，由于这种方法的灵敏度高，假阳性率低，而且可以与其他的预测方法相结合，能够随着数据库的数据增加而增强效率和准确率. 可以相信

这种算法必然能够加强具体 PPI 实验工作的顺利开展.

参 考 文 献

1 Seeler J S, Dejean A. Nuclear and unclear functions of SUMO. Nat Rev Mol Cell Biol, 2003, **4** (9): 690~699

2 von Mering C, Krause R, Snel B, *et al*. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 2002, **417** (6887): 399~403

3 Lakey J H, Raggett E M. Measuring protein-protein interactions. Curr Opin Struct Biol, 1998, **8** (1): 119~123

4 Gavin A C, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 2002, **415** (6868): 141~147

5 Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature, 2002, **415** (6868): 180~183

6 Marcotte E M, Pellegrini M, Ng H L, *et al*. Detecting protein function and protein-protein interactions from genome sequences. Science, 1999, **285** (5428): 751~753

7 Overbeek R, Fonstein M, D'Souza M, *et al*. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA, 1999, **96** (6): 2896~2901

8 Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol, 2003, **7** (2): 238~251

9 Huynen M A, Bork P. Measuring genome evolution. Proc Natl Acad Sci USA, 1998, **95** (11): 5849~5856

10 Cai L, Xue H, Lu H C, *et al*. Analysis of correlations between protein complex and protein-protein interaction and mRNA expression. Chinese Science Bulletin, 2003, **48** (20): 2226~2230

11 Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol, 2001, **311** (4): 681~692

12 Deng M, Mehta S, Sun F, *et al*. Inferring domain-domain interactions from protein- protein interactions. Genome Res, 2002, **12** (10): 1540~1548

13 Andersen J S, Wilkinson C J, Mayor T, *et al*. Proteomic characterization of the human centrosome by protein correlation profiling. Nature, 2003, **426** (6966): 570~574

14 Bader G D, Heilbut A, Andrews B, *et al*. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. Trends Cell Biol, 2003, **13** (7): 344~356

15 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

16 McDermott J, Samudrala R. Enhanced functional information from predicted protein networks. Trends Biotechnol, 2004, **22** (2): 60~62

17 Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics, 2003, **19** (Suppl 1): i197~i204

18 Hayashida M, Ueda N, Akutsu T. Inferring strengths of protein-protein interactions from experimental data using linear programming. Bioinformatics. 2003, **19** (Suppl 2): II58~II65

19 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

20 Gavin A C, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002, **415** (6868):141~147

21 Ho Y, Gruhler A, Bader G D, *et al*. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature, 2002, **415** (6868): 180~183

22 Uetz P, Giot L, Cagney G, *et al*. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature. 2000, **403** (6770): 623~627

23 Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA, 2001, **98** (8): 4569~4574

24 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

25 Camon E, Barrell D, Brooksbank C, *et al*. The Gene Ontology Annotation (GOA) Project: Application of GO in SWISS-PROT, TrEMBL and InterPro. Comp Funct Genom, 2003, **4**: 71~74

26 McEachern M J, Krauskopf A, Blackburn E H. Telomeres and their control. Annu Rev Genet, 2000, **34**: 331~358

27 Rezler E M, Bearss D J, Hurley L H. Telomere inhibition and telomere disruption as processes for drug targeting. Annu Rev Pharmacol Toxicol, 2003, **43**: 359~379

28 Kanoh J, Ishikawa F. Composition and conservation of the telomeric complex. Cell Mol Life Sci, 2003, **60** (11): 2295~2302

29 Stern B M. Mitosis: aurora gives chromosomes a healthy stretch. Curr Biol, 2002, **12** (9): R316~318

30 Kallio M J, McCleland M L, Stukenberg P T, *et al*. Inhibition of aurora B kinase blocks chromosome segregation, overrides the spindle checkpoint, and perturbs microtubule dynamics in mitosis. Curr Biol, 2002, **12** (11): 900~905

Protein-protein Interaction Prediction With Correlated Gene Ontology*

ZHANG Qian[1,2], WANG Jing-Ze[1)**]

([1)]*State Key Laboratory of Bio-membrane and Membrane Biotechnology, Institute of Zoology, The Chinese Academy of Sciences, Beijing* 100080, China;

[2)]*Graduate School of The Chinese Academy of Sciences, Beijing* 100039, China)

Abstract The cellular processes in cells are controlled by protein-protein interactions (PPI), and comprehensive PPI maps are important to understand the complicated regulatory, metabolic and signaling pathways. Recently, new parameters for PPI prediction are under discovering. Here, a Naïve Bayesian algorithm with the correlated Gene Ontology (GO) was used for PPI prediction. The characteristic pairs of GO terms was demonstrated by training a non-redundant PPI data set from two online budding yeast databases, and the probability about this two correlated GO terms was also obtained. The accuracy of the prediction was tested by both positive and negative control data. The approach can distinguish them properly, with a satisfied sensitivity and low false positive rate. After comparing the prediction result to the data derived from high-throughput experiments, it is proved that the method is more sensitive and efficient than other means. Furthermore, some new insightful knowledge about interactions of the proteins will be found using this prediction approach, and the prediction is very helpful to the laboratory experiments.

Key words protein-protein interaction (PPI), correlated, Gene Ontology (GO), Naïve Bayesian

---

*This work was supported by a grant from The National Natural Sciences Foundation of China (30171071).

**Corresponding author . Tel: 86-10-62551668, E-mail: wangjz@ioz.ac.cn

Received: December 1, 2004 Accepted: January 31, 2005