# 不确定性PPI网络链接预测

章月阳 刘 维
(扬州大学信息工程学院计算机系 扬州 225127)

**摘 要** 蛋白质交互网络预测是后基因组时代生物学中很重要的研究内容。到目前为止,对蛋白质交互网络相互作用的预测都是假设相互作用是确定的。但是,蛋白质交互网络和其它的一些生物数据会因为实验检测方法的局限性而呈现出不确定性。提出了一种基于信息传播的不确定性PPI网络的链接预测算法。在每个顶点对上按其出现链接的概率定义了链接信息量,该算法将边上的链接信息量在图上以一定的概率来传播。利用标准数据集进行测试,实验结果表明,所提出的算法具有很好的准确率和良好的生物统计特性。

**关键词** 蛋白质交互网络,不确定性PPI网络,信息传播,链接信息量
**中图法分类号** TP311 **文献标识码** A

## Link Prediction in Uncertain Protein-Protein Interaction Network

ZHANG Yue-yang LIU Wei
(Department of Computer Science, Colleage of Information Engineering, Yangzhou University, Yangzhou 225127, China)

**Abstract** Prediction of protein-protein interaction network is an important research content in post-genomic era. So far, the forecast for the PPI network interactions are assuming that the interaction is determined. However, protein-protein interaction networks and other biological data because of the limitations of the experiment test and presents the uncertainty. Put forward a kind of based on the uncertainty of information dissemination PPI network link prediction algorithm. We according to their appearance on each vertex to link the probability that defines the link information, the algorithm will be on the edge of the link information to spread at a certain probability on the diagram. We set for testing using the standard data, the experimental results show that the proposed algorithm, has good accuracy and good biometric features.

**Keywords** Protein-protein interaction network, Uncertain PPI network, Information dissemination, Link information

生命的分子基础在于生物分子之间的相互作用,蛋白质通过相互作用来完成生命活动,因此只有对蛋白质进行整体的、网络水平上的研究,才能彻底解释生命活动的本质与规律,真正揭示生命现象的分子机制。如某些疾病并非是因为单个蛋白质缺失或变异,而是交互网络出现了问题,这就导致了蛋白质组学[3]的出现。蛋白质组学研究一个细胞或生物组织在一定条件下所有蛋白质的结构与性质,以及这些蛋白质与其它分子之间的相互作用关系。其所研究的相互关系包括蛋白质组内各个蛋白质分子之间的相互作用关系,以及蛋白质组内各个分子与其他类型分子之间的关系,因此检测蛋白质之间的相互作用(protein-protein interaction, PPI)成为蛋白质组学的重要研究课题之一,其最终目标是建立细胞内全部蛋白质之间的交互网络。

近年来,生物学家基于生物原理,利用物理化学技术,提出了很多检测蛋白质相互作用的生物实验方法[4-6],如免疫共沉淀、酵母双杂交系统、串联亲和纯化等。但是新近发展起来的高通量实验方法,正如vonMering 2002年在Nature上发表的关于评价蛋白质相互作用正确性的论文所指出的,在80000对相互作用中,只有2400对能够被两种或两种以上的高通量方法检测到。实验方法在提供大量数据的同时,会带来大量的假阳性和假阴性数据。因此,有些学者开始利用计算方法对蛋白质相互作用进行预测,其本质均利用了某种生物现象与蛋白质之间的相互作用具有的统计上的相关性,如基于基因组信息的蛋白质相互作用预测方法[5-8]、基于变异进化信息的蛋白质相互作用预测方法[9,10]、基于序列信息的蛋白质相互作用预测方法[11-13]和基于蛋白质结构信息的蛋白质相互作用预测方法[14-16]等。相比于传统的生物实验方法,它具有成本低、速度快的优点。近几年,网络链路预测算法受到了研究者的广泛关注。而蛋白质相互作用预测问题,本质上是复杂网络的链接预测问题。网络中的链路预测是指根据网络中节点的特征或已经存在的边(结构特征),预测两个节点间边的存在性[17-19]。这种预测既包含了对未知链接(existent yet unknown links, missing links)的预测,也包含了对未来链接(future links)的预测。由于对于生物网络中隐而未知链接的揭示是需要耗费高额实验成本的,如果可以预测,而并非盲目地检测所有链接,并以此指导实验,就可节约实验开销。

到目前为止,对PPI网络的相互作用的预测都是假设相

章月阳(1989—),男,硕士生,主要研究方向为生物数据挖掘,E-mail:zhangyueyang520@126.com;刘 维(1982—),女,博士,副教授,主要研究方向为数据挖掘、生物信息学等。

互作用是确定的，即在确定性的网络上进行链接预测。但是，由于蛋白质相互作用的高通量生物实验技术，如酵母双杂交技术等，存在着固有的误差，因此实验测得的蛋白质相互作用是否真实地反映了实际相互作用是不确定的，只是给出了一个可能存在作用的概率[20-22]。著名的生物数据库STRING[23]已经将这种不确定性信息量化存储于数据库中。图1给出了一个蛋白质交互网络的片段，其中，顶点上的文字是蛋白质的名称，边上的数值表示蛋白质间相互作用真实存在的可能性。该数据取自STRING生物数据库。

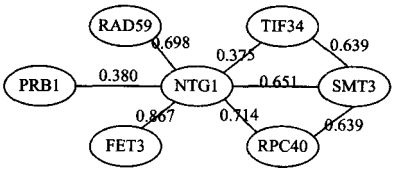


图1 蛋白质交互网络(数据来自STRING数据库)

综上所述，基于实验或计算方法构建的蛋白质交互网络实际上应当被建模为一个不确定图，其中，顶点代表唯一的蛋白质，边代表两个蛋白质的相互作用，并且把发生相互作用的概率标记在边上。

对于不确定网络的研究，主要集中在对其建模、聚类、频繁子图挖掘、Top$k$-极大团、频繁路径发现、紧密顶点子集挖掘、聚类分析等方面。1960年数学家Erdos和Renyi提出了随机图理论[24]来研究复杂网络中随机拓扑模型(ER)，自此ER模型一直是研究复杂网络的基本模型。由于不确定图的挖掘既要考虑到图的结构，又要考虑到图的不确定性，因此对于不确定图的挖掘问题的计算复杂度至少不会低于同类确定图数据挖掘的计算复杂性。Zou等人[25]证明了在期望语义下挖掘不确定图数据中期望频繁子图模式是一个NP问题，换句话说，它至少不比NP完全问题简单。Zou等人[26]证明了在概率语义下挖掘不确定图数据中概率频繁子图模式也是一个NP问题。Li等人[27]证明了挖掘不确定图中的Top$k$-极大团是一个NP完全问题。Zou等人[25]提出了一种近似期望频繁子图模式挖掘算法，采用近似挖掘技术，允许挖掘结果中有少量的非频繁子图模式，但其期望支持度可以不小于一个阈值。Parapetrou等人[28]在该算法的基础上提出了一个索引结构，进一步提高了该算法的效率。Jin[29]等人使用数据挖掘的方法研究了如何从不确定图中挖掘联通可靠性高于某一个阈值的全部导出子图。该问题在蛋白质复合体发现中有重要作用。Kollios等人[30]研究了不确定图的聚类问题，即将不确定图的顶点集合快速划分成若干互不相交的子集，使得该聚类在不确定图的所有可能子图上的聚类目标函数的期望值最大。Asthana等人[21]将蛋白质相互作用网络视作不确定图，研究了如何从蛋白质相互作用网络中预测最有可能属于某一给定蛋白质复合体的其他蛋白质。不确定性蛋白质相互作用网络中的相互作用预测问题，实际上是不确定图上的链接预测问题，是根据不确定图的拓扑特征和链接的概率信息来预测顶点间潜在的链接。对于这一问题，计算复杂度至少不会低于同类确定图上的链接预测问题，因此至今尚未发现有效的算法，所以很有必要针对高通量数据集的特点和相关蛋白质相互作用预测算法本身存在的缺陷，设计出更有效的基于不确定的PPI网络的相互作用预测算法。

本文提出了一种基于信息传播的不确定性PPI网络的链接预测算法。我们在每个顶点对上按其出现链接的概率定义了链接信息量，该算法将边上的链接信息量在图上以一定的概率来传播。在图上每个顶点对之间设定一个边权重，来衡量顶点对之间能够传播信息的能力。在边权重值较大的顶点对之间，链接信息量有较大的概率被传播。当某一个顶点对接收到来自相邻顶点对上的链接信息量时，它们原来的链接信息量以一定的比例被保留，也有一定的比例受到传送来的相邻顶点对上的链接信息量的影响。这个比例取决于顶点对之间的边权重，以及相邻顶点对上的边权重。在传播过程迭代到收敛时，各个顶点对之间的链接信息量即为它们之间存在链接的概率。我们利用标准数据集进行测试，实验结果表明，所提出的算法具有很好的准确率，算法识别的相互作用具有良好的生物统计特性。

## 1 链接信息量及其传送模型

蛋白质交互网络数据常用一个关系图来表示[31]，在这个关系图中，顶点代表唯一的蛋白质，边代表两个蛋白质之间的唯一的相互作用关系，即关系图中有唯一的不重复的顶点标志，每两个顶点之间有唯一的边。对于具有不确定性的蛋白质交互网络数据，每条边上有一个概率值，表示所连接的两个蛋白质之间存在链接的可能性。因此，我们对具有不确定性的蛋白质交互网络用以下的不确定图来描述。

**定义1**(不确定图) 一个不确定性的图由一个三元组$G=(V,E,P)$描述，其中$V$代表图中的顶点集合，$E$代表边的集合，$P=[p_{ij}]$为$|V|\times|V|$阶链接概率矩阵。$p_{ij}$表示$E$中的边$(i,j)$上附带的链接概率值，表示蛋白质$i$与蛋白质$j$之间存在真实相互作用的概率。当它们之间不存在相互作用时，即边$(i,j)$不存在时，$p_{ij}=0$。即：

$$p_{ij}=\begin{cases}\text{蛋白质 } i,j \text{ 之间存在相互作用的概率}, & (i,j)\in E\\ 0, & (i,j)\notin E\end{cases} \tag{1}$$

**定义2**(链接信息量) 我们在每一个顶点对$v_i$和$v_j$之间设立一个链接信息量$q(i,j)$，其初始值定义为$q^{(0)}(i,j)=p_{ij}$，它会以一定的方式在图上传播，因而使得各条边上的链接信息量$q(i,j)$不断改变。传播结束时，$q(i,j)$反映顶点对$v_i$和$v_j$之间有潜在链接的可能性。

由于链接信息量在图上传播后，各条边上的链接信息量会互相影响、互相修改，$q(i,j)$的值在不断改变。原来具有链接的顶点对$v_i$和$v_j$之间的$q(i,j)$不一定再等于初始值$p_{ij}$，而原来不具有链接的顶点对$v_i$和$v_j$之间的$q(i,j)$也不一定仍然等于零。这时，$q(i,j)$可以反映顶点对$v_i$和$v_j$之间有潜在链接的可能性。

**定义3**(边权重) 我们在每个顶点对$v_i$和$v_j$之间设定一个边权重$w_{ij}$来衡量顶点对$v_i$和$v_j$之间能够传播信息的能力，$w_{ij}$的值由下式来定义：

$$w_{ij}=\begin{cases}p_{ij}, & (i,j)\in E\\ \dfrac{\sum\limits_{k=1,k\neq i,j}^{n}p_{ik}\cdot p_{kj}}{\left[\sum\limits_{k=1,k\neq i}^{n}p_{ik}\right]\left[\sum\limits_{k=1,k\neq j}^{n}p_{kj}\right]}, & (i,j)\in E\end{cases} \tag{2}$$

由式(2)可以看出，若具有直接链接的顶点对$v_i$和$v_j$之间的链接概率较高，则其边权重$w_{ij}$值也较大；没有直接链接的顶点对$v_i$和$v_j$之间的边权重$w_{ij}$的值等于它们有第3个顶

点相互游走的概率。因此，这样定义的边权重值 $w_{ij}$ 能够反映顶点对 $v_i$ 和 $v_j$ 之间能够传播信息的能力。

算法让链接信息量在图上按下面的规则传播：在边权重值 $w_{ij}$ 较大的顶点对 $v_i$ 和 $v_j$ 之间，链接信息量有较大的概率被传播，在某一个顶点对 $v_i$ 和 $v_j$ 接收到来自相邻顶点对上的链接信息量时，它们原来的链接信息量 $q(i,j)$ 以一定的比例被保留，也以一定的比例受到传送来的相邻顶点对上的链接信息量的影响。这个比例取决于顶点对 $v_i$ 和 $v_j$ 之间的边权重，以及它们相邻顶点对上的边权重。

链接信息量 $q(i,k)$ 在相邻顶点对之间传送到顶点对 $v_i$ 和 $v_j$ 的方式如图 2 所示。$q(i,k)$ 经由顶点 $v_k$，再通过边 $(v_k, v_j)$ 以边权重 $w_{kj}$ 传送到顶点对 $v_i$ 和 $v_j$，来影响顶点对 $v_i$ 和 $v_j$ 的链接信息量 $q(i,j)$。对称地，顶点对 $v_k$ 和 $v_j$ 上的链接信息量 $q(j,k)$ 也经由顶点 $v_k$，再通过边 $(v_i, v_k)$ 以边权重 $w_{ik}$ 传送到顶点对 $v_i$ 和 $v_j$，来影响链接信息量 $q(i,j)$。

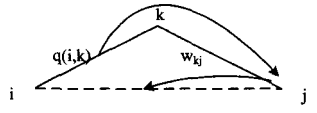


图 2　不确定性 PPI 网络上的信息传播过程示意图

通过以上阐述，我们发现若顶点 $v_k$ 对 $v_j$ 有强的影响，即 $w_{kj}$ 较大，同时 $v_i$、$v_k$ 之间存在边，则 $v_i$，$v_j$ 之间很可能有边存在。因此通过计算 $q(i,j)$ 就可以在不确定性的 PPI 网络中预测链接信息，即蛋白质间是否真实存在相互作用。

以上是顶点对 $v_i$ 和 $v_j$ 通过第 3 个顶点 $v_k$ 传送来的链接信息量来修改 $q(i,j)$ 的情形。事实上，除了 $v_i$ 和 $v_j$ 以外的其它所有顶点都可以作为第 3 个顶点 $v_k$ 传送来的链接信息量。因此，若设初始值 $q^{(0)}(i,j)=p_{ij}$，则 $q(i,j)$ 的迭代公式如下：

$$q^{t+1}(i,j)=\frac{1}{2(n-2)}\sum_{k\neq i,j}^{n}[w_{kj}q^t(k,i)+w_{ki}q^t(k,j)]+[1-\frac{1}{2(n-2)}\sum_{k\neq i,j}^{n}[w_{kj}+w_{ki}]]q^t(i,j) \quad (3)$$

在上述公式中，$n$ 为网络中的顶点总数，由于考虑了除 $i$、$j$ 以外的所有顶点与边 $(i,j)$ 的关系，因此一共考虑了 $n-2$ 个顶点。式(3)中的第一部分表示通过第 3 个顶点 $v_k$ 传送来的链接信息量来修改 $q(i,j)$ 的情况，其中求和项的第一部分表示 $q(k,i)$ 经由顶点 $k$，边 $(k,j)$ 对 $q(i,j)$ 的影响；第二部分表示 $q(k,j)$ 经由顶点 $k$，边 $(k,i)$ 对 $q(i,j)$ 的影响。式(3)中的第二部分表示各边上的信息不被传播时的情形。$[1-\frac{1}{2(n-2)}\sum_{k\neq i,j}^{n}[w_{kj}+w_{ki}]]$ 表示的是相关标号不被传播、$q^{(t)}(i,j)$ 的值保持不变的概率，而 $q^{(t)}(i,j)$ 则表示顶点对 $v_i$ 和 $v_j$ 之间的链接信息量。

## 2　算法描述

综上所述，对于给定的不确定性蛋白质交互网络 $G=(V,E,P)$，我们提出了一个预测其潜在蛋白质间的相互作用信息的高效算法 PPI_BIP(Protein-Protein Interaction Prediction Based on Information Propagation)。算法的描述具体如下：

**算法 1**　PPI_BIP

输入：G=(V,E,P)：不确定图；

　　ε：迭代误差的阈值；

输出：Q=[q^(t)(i,j)]：蛋白质间是否存在相互作用的链接预测信息；

Begin

初始化：

1. For i=1 to n do

　For j=1 to n do

　　{q^(0)(i,j)=p_ij；t=0；}

　迭代计算 Q 矩阵

2. Do

　For i=1 to n do

　　For j=1 to n do

　　{根据式(3)由 q^(t)(i,j)计算得到 q^(t+1)(i,j)；

　　t=t+1；}

　until $\max_{1\leq i,j\leq n}|q^{(t+1)}(i,j)-q^{(t)}(i,j)|<\varepsilon$；

输出 Q 矩阵

3. For i=1 to n do

　For j=1 to n do

输出链接预测结果值 q^(t+1)(i,j)；

End

算法的主要计算量在第 2 步的 $Q$ 矩阵迭代计算上。设网络中有 $n$ 个顶点，算法计算 $Q$ 矩阵需要 $O(n^3)$ 时间，因此，算法的时间复杂度为 $O(n^3)$。

## 3　实验结果与分析

### 3.1　实验环境

我们通过程序来验证本文所提出的基于信号传播的不确定性蛋白质交互网络链接预测算法 PPI_BIP 的有效性，实验程序的运行计算机系统配置为 Intel Core i5 1.8GHz CPU、内存 6GB、Windows 7 操作系统、Visual C++6.0 的程序编辑、编译链接环境。所有的算法均采用 C 语言实现。

### 3.2　标准数据集的选取

本文实验以酵母蛋白质交互网络作为研究对象，因为酵母是所有物种中蛋白质相互作用数据最为完备的。考虑到不确定性，本文使用文献[32]中的核心数据集作为测试数据，该数据集中包含 3672 个蛋白质，14317 对相互作用，每对相互作用的蛋白质之间还赋予了发生相互作用的概率值。我们从该数据集中收集的相互作用的蛋白质对中挑选对应的氨基酸序列长度＞50 的蛋白质对作为正的标准集，对于负的标准集，随机选取一对不在同一个亚细胞位置的蛋白质对。因为不在一个位置的蛋白质对一般来说不存在相互作用关系。这样最后总共得到 1025 对 PPI 作为正标准数据集，1077 对 PPI 作为负的标准数据集。

### 3.3　结果分析

#### 3.3.1　ROC 曲线

为了验证 PPI_BIP 算法在预测蛋白质相互作用这一问题上的有效性，我们利用标准数据集作为测试集验证了其准确性。测试方式为十折交叉验证(10-fold cross validation)，具体做法如下：将数据集分成 10 份，轮流将其中 9 份作为训练数据，1 份作为测试数据。这样进行 10 次试验，取其平均性能作为结果的性能指标。

我们用 ROC(Receiver Operating Characteristics)曲线下的面积(Hanley and McNeil 1982；Fawcett 2006)来衡量预测结果的好坏。给定一个预测算法和一个实例，可能有 4 种可能的结果(假设所有实例都只能属于两种可能的类别之一，即

正例或负例)：一个实例如果是正的而且被算法预测成正例，则称为真阳性(True Positive，TP)；若被预测成负例则称为假阴性(False Negative，FN)；一个实例如果是负的而且被算法预测成负例，则称为真阴性(True Negative，TN)；若被预测成正例则称为假阳性(False Positive，FP)。表 1 为 4 种情况的依赖表。假阳性率 $FP\ rate = FP/(FP+TN)$，真阳性率 $TP\ rate=TP/(TP+FN)$。把预测得到的结果按照分值从大到小排列，然后设定一系列不同的阈值，会得到一系列假阳性率和真阳性率。以假阳性率为横坐标，以真阳性率为纵坐标，描绘出的曲线就是 ROC 曲线。我们尽量希望 ROC 曲线偏向左上方，即在保持较低的假阳性率的同时得到较高的真阳性率。曲线下的面积 *AUC*(Area Under Curve)是对这一直观感觉的定量化描述。*AUC* 可通过式(4)计算得到。

$$AUC=\frac{\sum_{i=1}^{n_T} rank(r_i)-\frac{n_T(n_T+1)}{2}}{n_T n_N} \tag{4}$$

表 1　联合依赖表

| | |
|---|---|
| True Positives | False Positives |
| False Negatives | True Negatives |

在式(4)中，$n_T$ 是测试集中正例的个数，$n_N$ 是测试集中负例的个数，$r_i$ 是所有测试用例按照分值从小到大排列后第 $i$ 个正例所排的序号。最理想的情况是所有正例都排在后面，所有负例都排在前面，此时 $AUC=1$；最差的情况是所有负例都排在后面，所有正例都排在前面，此时 $AUC=0$。

图 3 是在一组测试集上得到的 ROC 曲线，其 *AUC* 值是 0.965，说明只有极少部分的正例没有被预测出来，而且大部分的负例都如实地被排除在外。图中阴影部分的面积就是 *AUC*。

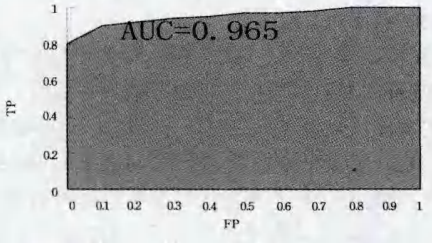


图 3　用一组测试集测试 PPI_BIP 算法的预测结果

图 4 是十折交叉验证得到的 ROC 曲线中的 5 条(为了清楚起见，只给出了其中的 5 条)，从图上可以看出，几乎每一次的结果都与图 3 的情况类似。这说明 PPI_BIP 算法在对预测蛋白质相互作用这一特定的问题上不仅准确度高，而且预测性能比较稳定。

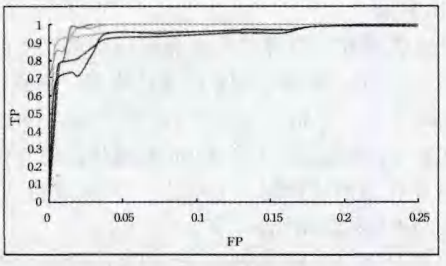


图 4　十折交叉验证中每一次验证得到的 ROC 曲线图

### 3.3.2　独立测试集性能评估

为了进一步检测本文算法的有效性，我们从 DIP 数据库中选择了其它 7 个物种的蛋白质相互作用的数据进行预测，实验结果如表 2 所列。在 7 个物种中有 5 个物种的预测精度超过了 80%，只有一个物种的预测精度低于 60%，为 58.38%，这 7 个物种的平均预测精度达到了 79.55%。这些结果进一步地说明了该方法是有效的，是能够用来表达蛋白质相互作用的比较重要的信息，具有良好的生物统计特性。

表 2　独立测试集的性能

| 独立测试集 | 相互作用的蛋白质对个数 | 正确的预测结果 |
|---|---|---|
| D. melanogaster(fruit fly) | 21975 | 18220(82.91%) |
| E. coli | 6954 | 4940(71.03%) |
| C. elegans | 4013 | 3260(81.24%) |
| H. saPiens(Human) | 1439 | 1189(82.63%) |
| H. Pylori | 1420 | 829(58.38%) |
| M. musculus(House mouse) | 319 | 281(88.09%) |
| R. norvegieus(Norway rat) | 114 | 105(92.11%) |
| Total average | 36234 | 28824(79.55%) |

**结束语**　由于蛋白质相互作用的高通量生物实验技术，如酵母双杂交技术等，存在着固有的误差，因此实验测得的蛋白质相互作用是否真实地反映了实际相互作用是不确定的，只是给出了一个可能存在作用的概率。因此，基于实验或计算方法构建的蛋白质交互网络实际上带有不确定性，应当被建模为一个不确定图。本文提出了一种基于信息传播的不确定性 PPI 网络的链接预测算法。我们在每个顶点对上按其出现链接的概率定义了链接信息量，该算法将边上的链接信息量在图上以一定的概率来传播。我们在图上的每个顶点对之间设定一个边权重，来衡量顶点对之间能够传播信息的能力。在边权重值较大的顶点对之间，链接信息量有较大的概率被传播。当某一个顶点对接收到来自相邻顶点对上的链接信息量时，它们原来的链接信息量以一定的比例被保留，也以一定的比例受到传送来的相邻顶点对上的链接信息量的影响。这个比例取决于顶点对之间的边权重，以及它们相邻顶点对上的边权重。在传播过程迭代到收敛时，各个顶点对之间的链接信息量即为它们之间存在链接的概率。我们在标准数据集上进行测试，实验结果表明，所提出的算法具有很好的准确率，算法识别的相互作用具有良好的生物统计特性。同时，对于 DIP 数据库中的 7 个物种的独立数据集的预测结果表明，本文算法有着优异和稳定的表现。考虑到蛋白质相互作用的复杂实质，我们提出的方法是对蛋白质相互作用预测手段的一个有效补充。

## 参 考 文 献


[1]　沈瑶瑶，严庆丰. 蛋白质相互作用研究进展[J]. 生命科学，2013，25(3)：269-274

[2]　Sehaerer M T，Kannenberg K，Hunziker P，et al. Interaction between GABA(A) receptor beta subunits and the multifunctional protein gC1q-R[J]. Biol. Chem，2001，276(28)：26597-2660

[3]　Gavin AC，Aloy P，Grandi P，et al. Superti-Furga G. Proteome Survey Reveals Modularity of the Yeast Cell Machinery[J]. Nature，2006，440(7084)：631-636

[4]　Tong A H，Drees B，et al. A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules[J]. Science，2002，295(5553)：321-324


(下转第 418 页)

[8] Pennaechiotti M, Pantel P. Entity Extraction via Ensemble Semantics[C]//Proc of EMNLP2009. Singapore: ACL, 2009: 238-247

[9] Tan Pang-ning, Kumar V. Introduction to Data Mining [M]. 2005

[10] 李贵,张淼,李征宇,等. 基于领域模型的 Web 数据抽取与集成[J]. 微电子学与计算机,2012,29(9):152-156

[11] 马安香,张斌,高克宁,等. 基于结果模式的 Deep Web 数据抽取[J]. 计算机研究,2009,46(2):280-288

[12] Probst K, Ghani R, Krema M, et al. Semi-supervised learning of at-tribute-value pairs from product descriptions[C]// Proceedings of the 20th International Joint Conference on Artifical Intelligence. 2007: 2838-2843

[13] Pasca M. Organizing and searching the world wide web of facts-step two: harnessing the isdom of the crowds[C]// Proceedings of the 16th International Conference on World Wide Web. 2007: 101-110

[14] Wick M, Culotta A, McCallum A. Learning Field Compatibilities to Extract Database Records from Unstructured Text[C]// EMNLP. 2006: 603-611

(上接第 402 页)

[5] Dandekar T, Snel B, Huynen M, et al. Conservation of Gene Order: A Fingerprint of Proteins that Physically Interact[J]. Science, 1998, 23(9): 324-328

[6] Marcotte E M, Pellegrini M, Ng H L, et al. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences [J]. Science, 1999, 285(5428): 751-753

[7] Enright A J, Iliopoulos I, Kyrpides N C, et al. Protein Interactions Maps for Complete Genomes Based on Gene Fusion Events [J]. Nature, 1999, 402(6747): 86-90

[8] Pellegrini M, Marcotte E M, Thompson M J, et al. Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles[J]. Proc. Natl. Acad. Sci. USA, 1999, 96 (8): 4285-4288

[9] Pazos F, Valencia A. In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs[J]. Proteins: Structure, Function and Genetics, 2002, 47(2): 219-227

[10] Goh C S, Bogan A A, Joachimiak M, et al. Co-evolution of Proteins with their interaction Partners[J]. J Mol Biol, 2000, 299 (2): 283-293

[11] Martin S, Roe D, Faulon J L. Predicting Protein-Protein Interactions Using Signature Products[J]. Bioinformatics, 2005, 21(2): 218-226

[12] Shen J, Zhang J, Luo X, et al. Predicting Protein-Protein Interactions Based Only on Sequences Information[J]. PNAS, 2007, 104(11): 4337-4341

[13] Guo Y, Yu L, Wen Z, et al. Using Support Vector Machine Combined with Auto Covariance to Predict Protein-Protein Interactions from Protein Sequences[J]. Nucleic. Acids. Res., 2008, 36 (9): 3025-3030

[14] Gomez S M, Lo S H, Rzhetsky A. Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks[J]. Genetics, 2001, 159(3): 1291-1298

[15] Gomez S M, Noble W S, Rzhetsky A. Learning to Predict Protein-Protein Interactions from Protein Sequences[J]. Bioinformaties, 2003, 19(15): 1875-1881

[16] Deng M, Mehta S, Sun F, et al. Inferring Domain-Domain Interactions from Protein-Protein Interactions[J]. Genome Research, 2002, 12(10): 1540-1548

[17] Ryan N. Lichtenwalter. New Precepts and Method in Link Prediction[C]//Proceedings of ACM KDD'10. 2010: 243-252

[18] Lv Lin-yuan, ZhouTao. Link Prediction in Complex Networks: A survey[J]. Physica A, 2011, 390: 1150-1170

[19] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39 (5): 651-661

[20] Bader J S, Chaudhuri A, Rithberg J M, et al. Gaining Confidence in High-Throughput Protein Interaction Networks [J]. Nature Biotechnology, 2003, 22(1): 78-75

[21] Asthana S, King O D, Gibbons F D, et al. Predicting Protein Complex Membership using Probabilistic Network Reliability [J]. Genome Research, 2004, 14(6): 1170-1175

[22] Suthram S, Shlomi T, Ruppin E, et al. A Direct Comparison of Protein Interaction Confidence Assignment Schemes [J]. BMC Bioinformatics, 2006, 7(1): 360

[23] Jensen L J, Kuhn M, Stark M, et al. STRING 8-a Global View on Proteins and Their Functional Interactions in 630 Organisms [J]. Nucleic Acids Research, 2009, 37: 412-416

[24] Erdos P, Renyi A. On the Evolution of Random Graphs [J]. Publ. Math. Inst. Hung. Acad. Sci., 1960, 5: 17-60

[25] Zou Z, Li J, Gao H, et al. Mining Frequent Subgraph Patterns from Uncertain Graph Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(9): 1203-1218

[26] Zou Z, Gao H, Li J. Discovering Frequent Subgraph over Uncertain Graph Database under Probabilistic Semantics[C]// ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). New York, USA, ACM, 2010: 633-642

[27] Li J, Zou Z, Gao H. Finding Top-k Maximum Cliques in an Uncertain Graph[C]// Proceedings of 26th International Conf. on Data Engineering. 2010: 649-652

[28] Parapetrou O, Ioannou E, Skoutas D. Efficient Discovery of Frequent Subgraph Patterns in Uncertain Graph[C]// Proceedings of the 14th International Conf. on Extending Database Technology. New York, USA, CAN, 2011, 355-366

[29] Jin R, Liu L, Aggarwal C C. Discovering Highly Reliable Subgraphs in Uncertain Graphs[C] // ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). New York, USA, ACM, 2011, 992-1000

[30] Kollios G, Potamias M, Terzi E. Clustering Large Probabilistic Graph[J]. IEEE Transactions on Knowledge and Data Engineering, 2012

[31] Yan Xi-feng, Zhou X J, Han Jia-wei. Mining Closed Relational Graphs with Connectivity Constraints[C]// Proc of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York: ACM Press, 2005, 324-333

[32] Krogan N J, Cagney G, et al. Global Landscape of Protein Complexes in theYeastSaccharomyces Cerevisiae[J]. Nature, 2006, 440(7084): 637-643