第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于空间映射的蛋白质相互作用网络链接预测算法

洪海燕[1] 刘 维[1,2]

(扬州大学信息工程学院 扬州 225127)[1]

(江苏省动物重要疫病与人畜共患病防控协同创新中心 扬州 225127)[2]

摘 要 蛋白质间的相互作用预测问题本质上是复杂网络的链接预测问题。到目前为止,已经有很多方法用于链接预测,这些方法要么只考虑拓扑信息,要么只考虑蛋白质相互作用网络内部的交互信息,但是仅考虑一种信息来预测蛋白质的交互信息是远远不够的。因此提出了一种新方法:将蛋白质相互作用网络看作是一个有权图,根据网络中两节点的拓扑结构和属性信息,分别计算它们的拓扑相似度和属性相似度来预测它们之间是否存在链接关系。在两种相似度平衡方面,考虑基于空间映射的方法,将它们独立地映射到另一空间,并且使它们分别映射的空间尽量相近,从而使得拓扑信息、属性信息有机融合。实验结果表明,提出的算法具有较好的准确率和良好的生物统计特性。

关键词 蛋白质相互作用网络,链接预测,空间映射

中图法分类号 TP311 文献标识码 A

Link Prediction Algorithm in Protein-Protein Interaction Network Based on Spatial Mapping

HONG Hai-yan[1] LIU Wei[1,2]

(College of Information Engineering, Yangzhou University, Yangzhou 225127, China)[1]

(Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou 225127, China)[2]

Abstract Protein-protein interaction(PPI) prediction is essentially the link prediction problem in the complex network. So far, many of the proposed link prediction methods either only consider topological information, or only consider the PPI interaction information within the network, but it is not enough. Therefore, this paper proposed a new method where the PPI network is represented as a weighted graph. In the graph, according to the two nodes' topology information and attribute information, the topology similarity and attribute similarity can be calculated so as to predict whether there are links between the two nodes. In order to balance the two similarities, we considered the method based on spatial mapping, that is, the similarities are independently mapped to another space, and the spaces are made as close as possible, so as to fuse the topology information and attribute information fusion. The results show that the proposed algorithm has better accuracy and good biometric characteristic.

Keywords PPI network, Link prediction, Spatial mapping

1 引言

在后基因组时代,功能基因组学的一个主要目标是识别与分析细胞环境中生物分子的相互作用,以便深入地理解生物分子相互作用与执行功能的机制。现在大量蛋白质-蛋白质相互作用数据给理解细胞功能带来了机遇与挑战[1]。蛋白质的相互作用与其本身的序列、结构等属性信息有着密切关系,因此,整合蛋白质本身属性信息并分析其规律,对蛋白质相互作用的预测以及深入理解蛋白质功能有着重要的意义。

在很多网络中,顶点都具有自己的信息,这些信息可能是所存的文本内容,如网页等,也可能是一个记录的属性值集合,这样就使得顶点之间有一定的相似度。PPI网络也不例外,除了具有拓扑信息之外,每个顶点还拥有反映蛋白质自身生物学特征的属性信息。由生物知识可知,生物体中的蛋白质是通过细胞中基因的转录和翻译得到的,也就是说,一个蛋白质本身的功能与它的氨基酸序列有很大的关系。如果两个蛋白质的氨基酸序列的属性信息的相似度很高,那么就认为它们有很大的可能进行相互作用来完成某个生物功能。因此,蛋白质自身的属性信息对于蛋白质间的相互作用预测显得尤为重要。

蛋白质间的相互作用预测问题本质上是复杂网络的链接预测问题,即:通过网络的已知拓扑结构或节点属性等信息来估计在两个尚未连接的节点之间产生一条连边的可能性。链接预测在计算机领域主要有基于马尔科夫链或者机器学习的算法,往往要考虑节点的属性特征。该类方法虽然能够得到较高的预测精度,但是计算的复杂度以及非普适性的参数使

本文受国家自然科学基金(61379066,61070047,61379064,61472344),江苏省自然科学基金(BK20130452,BK2012672,BK2012128),江苏省高校自然科学基金(12KJB520019,13KJB520026)资助。

洪海燕(1991—),女,硕士生,主要研究方向为生物数据挖掘,E-mail:1249055464@qq.com;刘 维(1982—),女,博士,副教授,CCF会员,主要研究方向为数据挖掘、生物信息学等,E-mail:yzliuwei@126.com。

其应用范围受到限制。另一类方法是基于网络结构的最大似然估计,该类方法也有计算复杂度高的问题。相比上述两种方法,基于网络拓扑结构相似性的方法更加简单,且具有普适性,迄今为止,已经有很多相关的理论和研究方法被提出。例如,文献[2,3]提出了基于网络的局部特点进行预测,但该方法并不能准确地得到预测结果,因为它只考虑了网络的局部特征,而没有捕捉和跟踪网络的全局特征,而 RWR 算法[4]通过搜索整个网络来发掘潜在的信息,其问题在于忽略了网络局部特征的重要性。基于上述两种方法的缺点,Iakovidou 等人[5]将网络的局部和全局特征结合起来,利用多路谱聚类来研究链接预测。Popescul 等人[6]利用网络的关联性特征构建了一种结构逻辑回归模型来预测网络中的未知链接。虽然这些方法对链接预测研究都很有效,但是它们在构建加权网络时大多都只考虑了部分信息。就 PPI 网络而言,现有的方法要么只考虑拓扑信息,要么只考虑 PPI 网络内部的交互信息,仅考虑一种信息来预测蛋白质的交互信息是远远不够的。还有一些方法只是简单地在无权网络中进行研究,也就是说,它们都只考虑了蛋白质对及其与邻居点的拓扑关系,并没有考虑到网络中物体内部自身的特征,如:Saito 等人[7,8]首次提出利用一对相互作用的蛋白质对与其邻节点的拓扑关系来评估这对蛋白质相互作用的可靠性,Chen 等人[9]也提出了一种基于网络拓扑结构的方法 IRAP,Goldberg 等人根据实际网络的特点确定了一个阈值,在此阈值之上的蛋白质相互作用被认为是可靠的,此外还有 CD-Dist[10]和 FSweight[11]等方法。由于蛋白质与蛋白质之间的交互主要依赖于生物过程的内部机理,它们之间共同所具有的某些生物学属性使得它们相似程度较高。而生物学属性越相似的蛋白质之间越有可能发生相互作用。所以,不能有效地利用这些基于生物学属性相似度的 PPI 网络的预测方法是不完整的,很难得到精度较高的预测结果。基于上述的不足,本文提出了一种基于蛋白质之间的拓扑和属性相似度来预测潜在相互作用的新方法,分别给出了反映拓扑信息的相似度矩阵和蛋白质属性信息的相似度矩阵的定义,并且给出了平衡属性相似度与拓扑相似度的方法,即将拓扑信息、属性信息用不同的线性变换分别映射到不同的空间,并且使它们分别所映射的空间尽量相近,从而使得拓扑信息、属性信息有机融合。这种方法改变了以往的方法以某一种信息为主导的做法,使得两种信息均衡地对预测结果产生影响。通过酵母数据对本文提出的方法进行了验证,并对实验结果的准确率进行了分析,其充分证明了融合拓扑信息和蛋白质自身属性信息的 PPI 网络的相互作用的预测结果更准确。

2 反映拓扑信息的相似度矩阵

注意到这样一个事实:在网络中,如果两个节点拥有较多的共同邻居,在它们之间更有可能存在一条连边。基于这个事实来计算两个蛋白质在拓扑结构上的相似度。用 $s_t \in [0,1]$ 表示两点在拓扑结构上的相似度。为了能准确地描述这种拓扑相似度,用网络中两节点拥有的共同邻居数目与两点度的和的比率来计算 s_t。在图 G 中,对于一条边 $e=\langle v_i, v_j\rangle \in E$,$e$ 连接的两个顶点 v_i 和 v_j 的拓扑相似度 $s_t(v_i, v_j)$ 定义为:

$$s_t(v_i, v_j)=2p/(d(v_i)+d(v_j)) \tag{1}$$

其中,p 代表顶点 v_i 和 v_j 所拥有的共同邻居节点的数目;$d(v_i)$ 表示顶点 v_i 的度数;$d(v_j)$ 表示顶点 v_j 的度数。式(1)反映出两个顶点 v_i 和 v_j 在网络中拓扑结构的交叉程度,在它们所连接的顶点之中,重叠部分所占比例越大,反映出两节点在网络拓扑结构上的相似程度越高。

例如:有一 PPI 网络如图 1 所示,根据上述定义,可以构建出反映该 PPI 网络拓扑结构的相似度矩阵 A 如下。

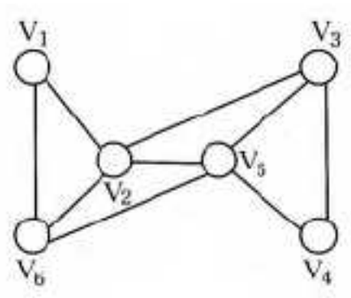

图 1 PPI 网络

$$A=\begin{bmatrix} 0 & \frac{1}{3} & \frac{2}{5} & 0 & \frac{2}{3} & \frac{2}{5} \\ \frac{1}{3} & 0 & \frac{2}{7} & \frac{2}{3} & \frac{1}{2} & \frac{4}{7} \\ \frac{2}{5} & \frac{2}{7} & 0 & \frac{2}{5} & \frac{4}{7} & \frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{5} & 0 & \frac{1}{3} & \frac{2}{5} \\ \frac{2}{3} & \frac{1}{2} & \frac{4}{7} & \frac{1}{3} & 0 & \frac{2}{7} \\ \frac{2}{5} & \frac{4}{7} & \frac{2}{3} & \frac{2}{5} & \frac{2}{7} & 0 \end{bmatrix}$$

例如,矩阵 A 中第一行第五列的元素 $A(1,5)$ 的值为 2/3,在矩阵中所有元素中是最高的,反映了顶点 v_1 和 v_5 的拓扑相似度较高,极有可能存在潜在的链接。

3 反映属性信息的相似度矩阵

由生物知识可知,生物体中的蛋白质是通过细胞中基因的转录和翻译得到的,也就是说,一个蛋白质本身的功能与它的氨基酸序列有很大的关系。如果两个蛋白质的氨基酸序列的属性信息的相似度很高,那么就认为它们有很大的可能相互作用来完成某个生物功能。所以,蛋白质序列的属性信息提取是决定预测结果的关键所在。如何从氨基酸的一级序列(字母序列)中提取出能够代表序列属性的信息,然后用适当的数学方法将这些属性参数描述或表示出来,使之能正确反映蛋白质的序列信息与结构或功能之间的关系,这就是对蛋白质序列信息进行属性信息提取的过程。这个过程实现了从字母序列到数值序列的转换,对于蛋白质结构类研究、相互作用及其功能预测是非常重要的,所以受到了越来越多生物学者的重视。

3.1 伪氨基酸组成(PseAAC)

在早期应用较广泛的属性信息提取方法中,大多是基于蛋白质序列中氨基酸组成成分的(AAC)。但是这种方法信息量过于单一,忽略了各氨基酸之间的联系,不可避免地丢失了一些非常重要的信息,从而影响预测的精度。本文选取近年来比较有代表性的伪氨基酸组成(PseAAC)[12]来提取蛋白质属性信息,PseAAC 能够反映出除了氨基酸组成以外的序列排序信息,因而被广泛使用。借助该方法,一个蛋白质 P 可以表示为如下形式:

$$P=[p_1, p_2, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+k}, \cdots, p_{20+\lambda}] \tag{2}$$

该方法构造了一个 $(20+\lambda)$ 维的向量,其中 $p_1, p_2, \cdots, p_{20}$

为 20 个氨基酸的组成成分。除此以外，还包含 λ 维由一系列序列次序相关因子组成的离散数据向量 $p_{20+1}, p_{20+2}, \cdots, p_{20+\lambda}$，表示了融合氨基酸的疏水指数、亲水指数以及侧链分子量的序列次序等信息。

使用 PseAAC 来分别表示第 i 个及第 j 个蛋白质：P^i 和 P^j。这里的 P^i 和 P^j 都写成形如式(2)的向量，因此不论是 P^i 还是 P^j，其维数都是相同的。形式如下：

$$P^i=[p_1^i, p_2^i, \cdots, p_{20+\lambda}^i] \tag{3}$$

$$P^j=[p_1^j, p_2^j, \cdots, p_{20+\lambda}^j] \tag{4}$$

3.2 属性信息的选择

在上述属性向量 $p_{20+1}, p_{20+2}, \cdots, p_{20+\lambda}$ 的选择中，为了更好地预测蛋白质间的相互作用及空间结构，一些研究者越来越倾向于用更多的属性信息来表示被预测的蛋白质。但如果属性信息过多，同样会造成冗余信息增加，引起特征空间维数过高，反而引起预测准确率不升反降的效果。为了解决这个问题，林智仁等人[13]开发了名为 fselect.py 的程序。该程序可以对蛋白质的全部属性信息进行打分，根据各个属性信息对应的分值进行排序，然后选择分值较高的作为蛋白质的最终属性。这样就能提取出最为合适且最为有效的属性信息，并避免维数灾难。

3.3 属性相似度的计算

目前，无论是社交网络还是生物网络，对于顶点属性向量的相似度的计算，都已经有很多有效的计算方法被提出，最常见的是用欧氏距离和 cosine 来计算。但这两种方法都有很大的缺点，它们只是简单地比较两个向量中的元素，因而这两种方法只能反映两个向量之间的距离，而不能反映出它们对应分量之间的变化趋势的相似度。鉴于上述缺点，本文采用 Pearson 相关系数(Pearson Correlation Coefficient，PCC)来计算两个向量之间的相似度。由于 PCC 相关系数能够很好地反映出两个变量变化趋势，它被广泛地用来衡量两个向量的相似度。PCC 的计算如式(5)所示：

$$r(P^i, P^j)=\frac{N\sum_{k=1}^{N}p_k^i p_k^j-\sum_{k=1}^{N}p_k^i\sum_{k=1}^{N}p_k^j}{\sqrt{(N\sum_{k=1}^{N}(p_k^i)^2-(\sum_{k=1}^{N}p_k^i)^2)(N\sum_{k=1}^{N}(p_k^j)^2-(\sum_{k=1}^{N}p_k^j)^2)}} \tag{5}$$

其中，$r(p^i, p^j)$ 代表两个相似度的计算结果，p_k^i 代表向量 P^i 中的第 k 个元素；p_k^j 代表向量 P^j 中的第 k 个元素；N 为向量 P^i 与向量 P^j 的维数。计算完两个蛋白质间的属性相似度后，可以得到相应的属性相似度矩阵 $B=[B_{ij}]_{n\times n}$。其中元素 $B_{ij}=r(p^i, p^j)$ 由式(5)计算，表示第 i 个及第 j 个蛋白质间的属性相似度值

4 算法描述

4.1 拓扑相似度与属性相似度间的平衡

目前多数对蛋白质相互作用网络的预测算法要么只考虑拓扑信息，要么只考虑蛋白质的属性信息。这种仅考虑一种信息来预测蛋白质间的相互作用的算法所使用的信息是不完整的，其结果的精确度会受到影响。基于上述不足，已有不少研究者提出了将蛋白质间属性相似度与它们交互网络的拓扑相似度结合起来预测 PPI 网络中潜在的蛋白质相互作用的方法。就如何平衡属性相似度与拓扑相似度间的关系而言，最常用的方法如下。

(1)目前绝大多数的、最简单的方法是将属性相似度与拓扑相似度加权求和得到总相似度。设拓扑相似度矩阵为 A，属性相似度矩阵为 B，总相似度则为 $c=w_1A+w_2B$，w_1，w_2 为权重，如图 2 所示。此类方法的缺点是：1)造成信息的损失，仅仅简单地加权求和不能反映出原有矩阵的分布规律；2)权重 w_1，w_2 的值难以确定。

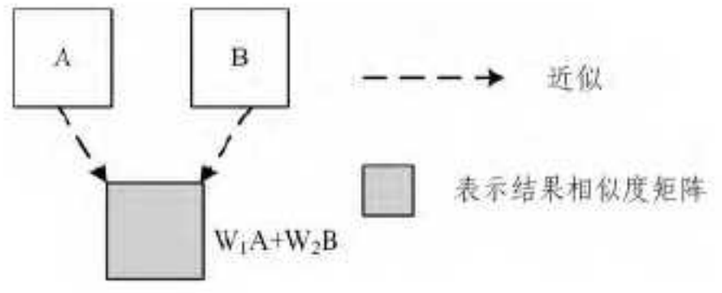

图 2 将两种相似度加权平均

(2)另一种方法是应用矩阵比对的方法，该方法是将 G 的拓扑相似度矩阵 A 和蛋白质的属性矩阵 B 进行矩阵比对，目标是使得下式最小：

$$L=\| A-BWB^{\mathrm{T}} \|_F^2 \tag{6}$$

由式(6)不难发现该方法的步骤是寻找一个权重向量 W，对不同属性赋予不同权重从而使得与 A 十分接近，最后取 BWB^{T} 作为所要的相似度矩阵，如图 3 所示。这种做法实际上是忽视了属性信息，服从了拓扑信息。

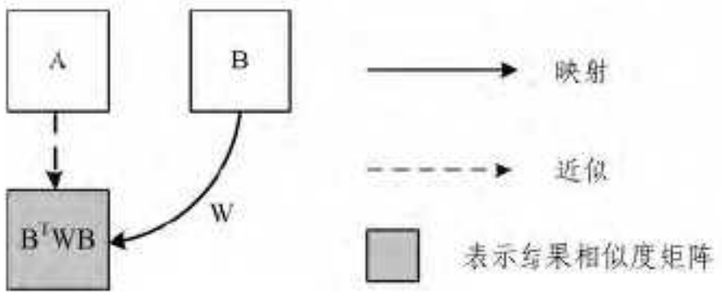

图 3 将属性相似度单向映射到拓扑相似度

(3)第 3 种方法如图 4 所示[14]，即将矩阵 A、B 分别进行映射的方法。该方法用同一种线性变换将 A、B 分别映射到不同的空间，取 A 映射后的空间内的度量作为最终的相似度矩阵。即最小化下式：

$$L=\| A-USU^{\mathrm{T}} \|_F^2+\| B-UV^{\mathrm{T}} \|_F^2 \tag{7}$$

尽管该方法使用同一种映射 U，但是在不同空间，其结果取 USU^{T} 作为所要的相似度矩阵。由于 USU^{T} 与 A 接近，因此该方法仍然以 A 为标准，即以拓扑信息为主。

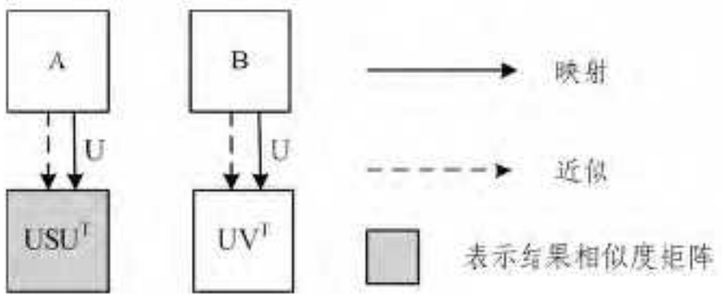

图 4 将两种相似度用同一映射映射到不同的空间

通过以上阐述，我们发现现有的方法基本上都以 A 的近似作为最终的相似度度量，从而使得蛋白质间的属性信息失去作用。

我们提出一种新的方法，将拓扑信息、属性信息分别独立地映射到另一空间，并且使它们分别映射的空间尽量相近，从而使得拓扑信息、属性信息有机融合，而不是以某一种信息为主导。

为此，设 A 为反映拓扑信息的相似度矩阵，B 为反映属性信息的相似度矩阵，要求映射 U、V 使得下式极小：

$$L(U,V)=\|AU-BV\|_F^2+\lambda_1\|U\|_F^2+\lambda_2\|V\|_F^2 \quad (8)$$

其中，U、V 为 $n\times r$ 矩阵（$r\leqslant n$），$\|\cdot\|_F$ 为 Frobenius 范数，式(8)中的后两项为惩罚项，目的是防止过分拟合。该做法的映射过程如图5所示。

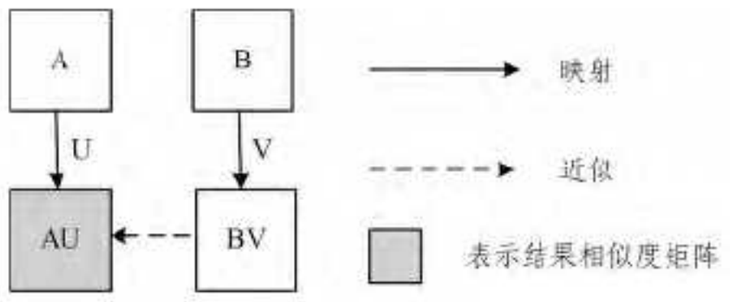

图5 将两种相似度用不同映射映射到相近的空间

由图5可以看出，当 $L(U,V)$ 取得最小值时，AU 与 BV 将十分接近，AU 为反映拓扑信息的相似度矩阵 A 的映射结果，而 BV 为属性相似度矩阵 B 的映射结果，由于二者相近，因此得到的相似度矩阵 AU 同时反映了网络的拓扑结构信息和蛋白质自身的属性信息，是二者融合的结果，它并不偏向于二者中的某一个，可以将其作为最后的结果。

为了求得 $\min_{U,V} L(U,V)$，要对 $L(U,V)$ 中的 U、V 元素求偏导数，具体过程如下：

$$\frac{\partial L}{\partial U}=(A^TA+AA^T)U-2V^TB^TA+2\lambda_1U$$

考虑到 A 的对称性，得到：

$$\frac{\partial L}{\partial U}=2A^2U-2V^TB^TA+2\lambda_1U$$

令 $\frac{\partial L}{\partial U}=0$，得到

$$(A^2+\lambda_1 I)U=V^TB^TA \quad (9)$$

同理可得：

$$\frac{\partial L}{\partial V}=(B^TB+BB^T)V-2B^TAU+2\lambda_2V$$

考虑到 B 的对称性，得到：

$$\frac{\partial L}{\partial V}=2B^2V-2B^TAU+2\lambda_2V$$

令 $\frac{\partial L}{\partial V}=0$，得到

$$(B^2+\lambda_2 I)V=B^TAU \quad (10)$$

根据式(9)、式(10)，可以用 EM 算法求解 U 和 V，过程如下。

算法1 EM-PPI

输入：相似度矩阵 A,B；相应的参数 λ_1,λ_2；最大迭代次数 N_{max}

输出：映射矩阵 U,V

Begin

1. 取 V 的一组初值 $V_{(0)}$；i=1；
2. Repeat
3. $U_{(i+1)}=(A^2+\lambda_1 I)^{-1}V_{(i)}^TB^TA$
4. $V_{(i+1)}=(B^2+\lambda_2 I)^{-1}B^TAU_{(i)}$；
5. i=i+1；
6. Until convergence or $i>N_{max}$。

End

算法第3、4行的迭代公式分别由式(9)、式(10)得到。N_{max} 是最大迭代次数，算法迭代到收敛或者在迭代次数达到 N_{max} 时结束。

5 算法框架

通过以上分析，可以得到基于空间映射的蛋白质相互作用预测算法 PPIP-BSM(Protein-Protein Interaction Prediction Based on Space Mapping)的框架，具体如下。

算法2 PPIP-BSM

输入：图 G 中点的集合 V，图 G 中边的集合 E，相应的参数 λ_k，k=1,2

输出：所有预测的链接(蛋白质相互作用)：E′

Begin

1. 根据式(1)计算拓扑相似度矩阵 A；
2. 给定一蛋白质，将蛋白质表示为如式(2)的形式；
3. 利用 fselect.py 进行属性信息选择；
4. 根据式(5)计算蛋白质之间的属性相似度矩阵 B；
5. 利用 EM-PPI 算法求得 U,V 的值，使得式(8)最小；
4. 取 AU 为最终的相似度矩阵，进而预测链接(蛋白质间的相互作用)的存在。

End

6 实验结果及分析

6.1 实验环境及数据集描述

我们用实验来验证本文所提出的基于空间映射的蛋白质相互作用预测算法 PPIP-BSM 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，Windows7 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。

本文的实验数据均采用了当前的标准测试数据来测试算法的有效性，即酵母蛋白质相互作用网络。因为酵母是所有物种中蛋白质相互作用数据最为完备的，以 1998 年 Cho 等人[15]发布的酵母基因表达数据作为微阵列数据。这些数据均可以从斯坦福基因数据库和其他生物数据库中得到。它们是通过观察酵母细胞在 α 条件下从分裂前期的 G1 期到分裂后期的 M 期表达收集的。

6.2 结果分析

6.2.1 属性个数对实验结果的影响

以 MIPS 数据库(http://mips.helmholtz-muenchen.de)中的酵母蛋白质交互网络[16]作为测试数据，随机从网络中剔除 500 条边，构成一个不完整的网络。在这个不完整网络中按照本文的算法 PPIP-BSM 进行预测，再用预测后的结果与原有的标准网络进行比较。考察在所选择的属性个数 λ 变化的情况下，实验结果在精确度上的变化。其中，精确度(Precision)定义如下：

$$Precision=\frac{m}{L} \quad (11)$$

其中，m 为在预测结果在标准网络中存在的链接数，L 为预测的总链接数。

表1为实验结果所反映的精确度随属性个数 λ 的变化情况。从表1中可以看出，随着属性个数 λ 的提高，预测结果的准确率也在不断提高，平均准确率达到 71.87%。当属性个数 $\lambda=37$ 时，预测结果的最高准确率可达 76.51%；随后，随着 λ 的增大，预测结果的准确率反而会下降，因此属性个数 λ 的最佳取值为 37，所以给定一个蛋白质和常数 $\lambda=37$，可以得到 $20+37=57$ 种伪氨基酸组成成分信息，则可将其表示为如式(2)所示的 57 维向量。同时，我们用 AAC 代替 PseAAC 进行属性提取，结果发现对于同样的数据集，该方法的平均准确率为 69.65%。显然，属性提取方法 PseAAC 更为优越。

表 1　属性个数对精度的影响

属性个数 λ	平均准确率(%)
11	65.90
24	70.34
37	76.51
50	74.72

6.2.2　实验结果随网络不同完整性水平的变化

通过不同完整性水平上的网络来观察实验结果准确率的变化。首先从原酵母蛋白质交互网络中随机剔除一些边，分别保留 40%、50%、60%、70%、80%、90% 条边，形成不同的不完整网络。图 6 为实验结果随观察点的变化。从图 6 中可以看出，观察点越多实验结果的准确率就越大。当网络中有 90% 的边是已知的时，预测结果的精确度最高；当只有 40% 的边是已知的时，预测结果的平均准确率也能达到 68%。从上述的分析结果可以看出，本文的算法能够进行有效的预测。

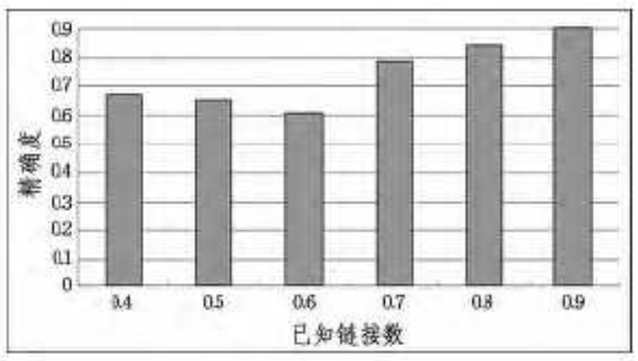

图 6　实验结果随已知链接数的变化

6.2.3　实验结果受不同权重信息的影响

通过不同权重信息构成的 PPI 网络来观察不同完整性水平下网络预测结果准确率的变化。在实验中，我们将 CD-Dist、带权重的 SRW 两种比较有代表性的算法与本文提出的算法进行比较，CD-Dist [10] 是基于网络拓扑结构的方法，带权重的 SRW[17] 是考虑了蛋白质交互网络内部的交互信息的方法，而本文算法则是同时考虑网络拓扑结构和网络中蛋白质内部自身的特征的方法。从原始酵母蛋白质交互网络中随机地剔除一些边，分别保留 40%、50%、60%、70%、80%、90% 条边，形成不同的不完整网络。运用上述 3 种不同方法分别在不同的已知链接数的情况下观察预测结果的精确度。图 7 为实验结果。从图 7 中可以看出，综合考虑网络拓扑结构以及网络内蛋白质自身属性信息的算法的整体精确度优于其它两种算法的精确度。当网络中有 90% 的边是已知的时，本文算法的预测精度可达 87%，仅考虑网络中物体内部自身的特征的带权重的 SRW 算法的精确度只能达到 66%，考虑网络拓扑结构的 CD-Dist 算法的精确度只能达到 46%。同样地，我们分别在不同的已知链接数的情况下观察预测结果的 AUC 指标。AUC(Area Under the Receiver Operating Characteristic Curve)[18] 是从整体上衡量算法的精确度的一个指标。由图 8 可以发现相比于其他两种算法，本文算法的预测结果的 AUC 指标值要远远优于其它算法，说明本文算法可以取得精度较高的预测结果。从上述的分析结果可以看出，单一信息在链接预测中存在局限性，因此将两种权重信息融合起来进行链接预测可以达到很好的效果。

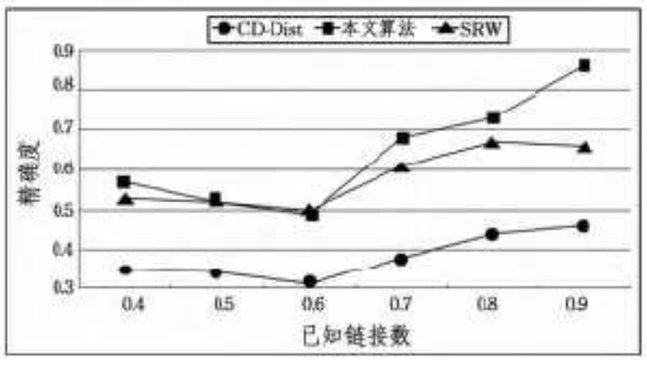

图 7　精确度随已知链接数的变化

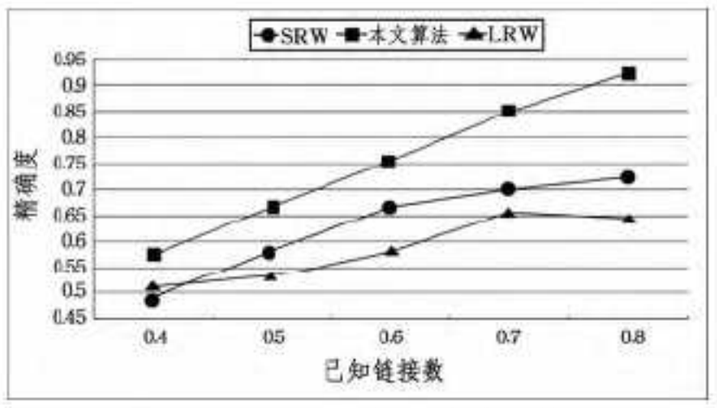

图 8　AUC 随已知链接数的变化

结束语　本文提出了一种基于空间映射的蛋白质相互作用网络链接预测方法来预测 PPI 网络中潜在的、隐藏的链接。将 PPI 网络看作是一个有权图，根据网络中两节点的拓扑结构和属性信息，分别计算它们的拓扑相似度和属性相似度来预测它们之间是否存在链接关系。在两种相似度平衡方面，考虑基于空间映射的方法，将它们独立地映射到另一空间，并且使它们分别所映射的空间尽量相近，从而使得拓扑信息、属性信息有机融合。通过真实的酵母蛋白质相互作用网络观察预测结果，表明本文算法是有效的。但是本文仍有不足之处，如针对该方法只在一个数据集上进行了实验验证。在多个数据集上验证本文的算法和将该方法拓展到动态的 PPI 网络进行链接预测是下一步研究的任务。

参 考 文 献

[1] Birlutiu A, D'Alche-Buc F, Heskes T. A Bayesian Framework for Combining Protein and Network Topology Information for Predicting Protein-Protein Interactions [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 12(3): 538-550

[2] Chen Ji-lin, Geyer W, Dugan C, et al. Make New Friends, but Keep the Old Recommending People on Social Networking Sites [C]// Proc of the 27th International Conference on Human Factors in Computing Systems. 2009: 201-210

[3] Wang Peng, Xu Bao-wen, Wu Yu-rong, et al. Link prediction in social networks: the state-of-the-art [J]. Science China Information Sciences, 2015, 58(1): 1-38

[4] Pan Jia-yu, Yang H I, Faloutsos C, et al. Automatic Multimedia Cross-Modal Correlation Discovery[C]// Proc of the 10th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004: 653-658

[5] Iakovidou N, Symeonidis P, Manolopoulos Y. Multiway Spectral Clustering Link Prediction in Protein-protein Interaction Networks[C]// Proc of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine. 2010: 1-4

[6] Popescul A, Ungar L H. Statistical Relational Learning for Link Prediction[C]// Proc of at IJCAI Workshop on Learning Statistical Models from Relational Data. 2003

[7] Saito R, Suzuki H, Hayashizaki Y. Interaction Generality, a Measurement to Assess the Reliability of a Protein-Protein Interaction[J]. Nucleic Acids Researeh, 2002, 30: 1163-1168

[8] Saito R, Suzuki H, Hayashizaki Y. Construction of Reliable Protein-Protein Interaction Networks with a New Interaetion Generality Measure[J]. Bioinformatics, 2003, 19: 756-763

(下转第 434 页)

根据词语在 VDEA 词典中所处位置可以得到与特征词的语义距离和情感距离，如表 2 所列。

表 2 语义情感距离表

语义距离	褒义	17,13,3,13,13
	贬义	15,12,13,12,15
情感距离	褒义	3,1,7,3,3
	贬义	7,8,7,6,6

结合词汇情感倾向计算分式进行情感倾向计算，经对比分析，在式(3)中当参数值"$A \geqslant 4$"时得出的相似度更趋合理，在此将 A 值定为 4，α 和 β 在[0.5,1]之间取值时结果趋于准确，经对比分析分别定为 0.6 和 0.8，γ 和 δ 是对综合倾向值与情感倾向最大值之和进行调节的参数，经比较将参数 γ 和 δ 值分别定为 0.5。最终得出 good 的倾向值为 0.39，大于阈值 0，为褒义形容词。同理对其它测试形容词与特征形容词作以上距离计算，并进行倾向值计算，实验结果如表 3 所列。

表 3 不同特征词集褒贬倾向性准确率(%)

5 对特征形容词			10 对特征形容词			20 对特征形容词		
褒义词准确	贬义词准确率	平均准确率	褒义词准确	贬义词准确率	平均准确率	褒义词准确	贬义词准确率	平均准确率
80	87	84.5	83	92	87.5	90.5	98	94.25

因为所选特征词数量以及测试词数量较少，所以整体效果一般，得出的准确率偏低。从实验结果来看，褒义形容词准确率比贬义形容词准确率低，这与褒贬形容词的特征有关。提高准确度的方法有两点，首先要提高特征形容词的数量以及覆盖面，其次是在算法上进行优化，尤其是在可调节参数的设定上要进行多次测试并将结果进行分析比较，使参数值更趋合理，从而提高词汇情感倾向性的准确率。

结束语　本文以英语形容词语义配价和情感信息描述为主要研究对象，为英语形容词语义信息和情感信息的分析处理探索新途径，在一定程度上填补了形容词配价研究的空白，拓宽了配价语法的研究范围。此外，基于本词典提出了词汇情感倾向性分析模型，在词汇情感倾向分析中引进情感相似最大值作为调节参数，提出了将语义距离和情感距离相结合进行词汇情感倾分析的方法。面向情感倾向分析的应用是构建本词典的重要目标，在不断丰富和完善配价词典的同时，还需积极探索词典在其它具体项目中的应用，如服务于政治、军事领域的语义检索、信息抽取和机器翻译等。

参 考 文 献

[1] Huang Jin-zhu, Yi Mian-zhu. The Design and Implementation of VDEA[C]// 4th International Universal Communication Symposium Proceedings. Beijing, 2010: 324-330

[2] 詹卫东. 一个汉语语义知识表达框架：广义配价模式[OL]. [2014-10-21]. http://ccl.pku.edu.cn/doubtfire

[3] 毕玉德. 面向韩文信息处理的谓词句法语义信息词典的构建[J]. 解放军外语学院学报, 2004, 25(4): 53-57

[4] Herbst T, et al. A Valency Dictionary of English [J]. International Journal of Lexicography, 2009, 22(1): 55-85

[5] 傅玉芳. 常用形容词分类词典[D]. 上海：上海大学出版社, 2004

[6] Bkyt T, et al. Polarity classification of celebrity coverage in the Chinese press[EB/OL]. [2014-08-12]. http://analysis.mitre.org/Proceedings/index.html

[7] 王付君. 性质形容词的价[J]. 长春教育学院学报, 2012, 28(1): 42-44

[8] 邱天. 现代汉语双价形容词研究[D]. 长春：东北师范大学, 2006

[9] 周咏梅，杨佳能. 面向文本情感分析的中文情感词典构建方法[J]. 山东大学学报, 2013, 23(6): 27-33

[10] 潘文彬. 基于情感词典的中文句子情感倾向分析[D]. 北京：北京邮电大学, 2011

[11] 黄硕. 中文情感知识库的构建与应用[D]. 北京：北京邮电大学, 2013

[12] 张成功，刘培玉，等. 一种基于极性词典的情感分析方法[J]. 山东大学学报, 2012, 47(3): 47-50

[13] 刘群，李素建. 基于《知网》的词汇语义相似度的计算[OL]. [2015-03-28]. http://www.wenku.baidu.com/view/6213af995/e79b8968022660.html

[14] 朱嫣岚，闵锦，等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2007, 20(1): 14-20

[15] 杨昱昺，吴贤伟. 改进的基于知网词汇语义褒贬倾向性计算[J]. 计算机工程与应用, 2009, 45(21): 91-93

(上接第 417 页)

[9] Chen J, Hsu W, Lee M L, et al. Increasing Confidence of Protein Interactomes Using Network Topological metrics[J]. Bioinformaties, 2006, 22: 1998-2004

[10] 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 Biol, 2003, 5(1): 1-13

[11] Chua H N, Sung W K, Wong L. Exploiting Indirect Neighbors and Topological Weight to Predict Protein Function from Protein-Protein interactions[J]. Bioinformatics, 2006, 22: 1623-1630

[12] Chou K C. Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition [J]. Proteins: Structure, Function, and Bioinformatics, 2001, 43(3): 246-255

[13] http://www.csie.ntu.edu.tw/~cjlin

[14] Zhu S, Yu K, Chi Y, et al. Combining Content and Link for Classification Using Matrix Factorization[C]// Proceedings of SIGIR'07. Amsterdam, The Netherlands, 2007: 487-494

[15] Cho R J, Campbell M J, et al. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle[J]. Molecular Cell, 1998, 2(1): 65-73

[16] Mewes H W, Frishman D, et al. MIPS: a database for genomes and protein sequences[J]. Nucleic Acids Res, 2002, 30(1): 31-34

[17] Backstrom L, Leskovec J. Supervised Random Walks: Predicting and Recommending Links in Social Networks[C]// Proceedings of the fourth ACM International Conference on Web Search and Data Mining, 2011. Hong Kong, 2011: 635-644

[18] Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic(ROC) curve[J]. Radiology, 1982, 143: 29-36