基于多视图融合的蛋白质功能模块检测方法

张 媛[1],贾克斌[1],ZHANG Aidong[2]
(1.北京工业大学电子信息与控制工程学院,北京 100124;
2. Department of Computer Science and Engineering, University at Buffalo, the State University of New York, Buffalo, NY14214)

摘 要: 结合多种生物数据分析蛋白质相互作用网络(Protein-Protein Interaction Network, PPIN)中的功能模块结构,是目前蛋白质功能计算分析领域亟待解决的难题之一.本文提出了一种基于聚合非负矩阵分解(Collective Non-negative Matrix Factorization, CoNMF)的多视图一致性功能模块检测方法,该方法同时逼近多视图数据,寻找统一的最优解达到对原多数据的最优近似.根据该统一解得到功能模块关系,同时该方法能够找到可重叠性的功能模块.实验结果显示本文所提出算法通过融合基因本体、基因表达谱与 PPIN 数据,在模块检测准确度上有一定提高,检测出的蛋白质功能模块具有真实生物意义.

关键词: 蛋白质相互作用网络; 网络模块挖掘; 多数据集成; 可重叠聚类
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112 (2014)12-2337-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.12.001

Consistent Protein Functional Module Detection from Multi-View of Biological Data

ZHANG Yuan[1], JIA Ke-bin[1], ZHANG Ai-dong[2]
1. *Dept. of Electronic Information and Control Engineering, Beijing University of Technology, Beijing* 100124, *China*;
2. *Dept. of Computer Science and Engineering, University at Buffalo, the State University of New York, Buffalo, NY*14214, *USA*)

Abstract: Detecting functional modules from protein-protein interaction networks (PPINs) is an active research area with many practical applications. To date, multiple biological data sources are available such as gene expression data and gene ontology (GO). These data explain the biological roles of proteins from different views and provide additional information to alleviate false information in PPINs. This work focuses on extracting consistent information from diverse data sources. To address this problem, this work proposes a collective non-negative matrix factorization (CoNMF) method which efficiently integrates views of gene ontology, gene expression data and PPINs. In our method, the integration problem is reduced to optimimum approximations of multi-view data by the productions of their common matrix factor with basis matrices. As a result, the common matrix factor provides an intuitive interpretation of soft clustering. Extensive experiments show that CoNMF outperforms most of the baseline methods listed in the paper and is an effective method to extract functional modules in PPINs.

Key words: protein-protein interaction network; functional module detection; multiple data sources Integration; soft clustering

1 引言

基于蛋白质相互作用网络(Protein-Protein Interaction Network, PPIN)分析蛋白质的具体功能是目前生物信息学研究中的一大热点.PPI 描述的是两个蛋白质间的物理直接连接或两者间的功能一致性[1].PPIN 以每个蛋白质为结点,两个蛋白质间的相互关系作为两结点的边,形成一无向图.众所周知,蛋白质大多通过相互作用形成功能意义上紧密联系的集合,也就是我们所说的功能模块,以共同执行相应的一种或多种生命活动,分析 PPI 功能意义是了解和掌握生命活动的分子机制的基础.

至今,学术各界已提出多种聚类方法用来检测 PPIN 中的功能模块,其中基于非监督学习的聚类方法是最常用的 PPI 模块挖掘方法[2~4].然而,由于 PPI 数据的高噪声和不完整性,聚类结果大多不能令人满意.有学者证实酵母双杂交实验(yeast two-hybrid)所得到的 PPI 数据中其误报率(false positive)可达 50%[5].基于如

收稿日期:2013-08-20;修回日期:2013-11-07;责任编辑:覃怀银
基金项目:国家自然科学基金课题(No.30970780);教育部博士点基金项目(No.20091103110005)

此大噪声比例的数据,单纯依据 PPIN 的拓扑特征来寻找紧密联系的功能模块并不能给我们提供可靠的结果.这也是之前一些传统功能模块挖掘方法,如 MCODE[2],CFinder[6],马尔科夫聚类(Markov CLustering,MCL)[7]等方法的一大掣肘.

近年来生物数据收集手段日新月异,通过多种手段对 PPI 数据做补充研究成为可能.首先,基因表达数据被用于寻找共表达的基因和基因产物.这一方法的基本假设是,在一段生命过程中表达模式相似的基因(或基因产物)倾向具有相同的功能,同时在 PPIN 中也更倾向于相互联系形成密集功能模块[8].Segal 等人依据基因共表达的模式特征来提取功能模块[9];也有学者将基因共表达作为 PPIN 权重进行分析[10].另一方面,Cho 等人[11]利用基因本体注释信息(Gene Ontology,GO)[12]计算蛋白质相似度,并据此构建了一个模拟蛋白质功能流在网络中的传输模型,功能流所到之处被划分为同一功能模块.这一方法体现了蛋白质功能在网络中的动态传递概念,并解决了多功能蛋白质的多功能模块从属,即功能模块的重叠问题.

这些方法均得到一定程度的成功.然而不同的数据各有倾向:基因共表达倾向于检测细胞生命过程中表达模式一致的蛋白质群,而 GO 则是静态描述的功能信息.同时,二者各有弊端:首先,基因表达数据在收集过程中,实验本身引入的噪声不可避免;其次,由于人类技术所限,还有丰富的基因功能特质依然未知,GO 注释信息也仍在不断完善[13].将每个角度所获得的数据看做一个视图,如何从多视图中提取一致信息是数据挖掘界近来面临的一个难题.基于以上观察,本文提出一种同时分析基因共表达、GO 注释和 PPIN,提取多视图中聚合特征最为一致的功能模块的方法,简称作 CoNMF(Collective Non-negative Matrix Factorization).本文的基本假设是,具有相同功能的蛋白质一般在相互作用网络中倾向紧密联系,在基因表达谱中具有更为相似的表达模式,同时在基因功能标注系统中也应倾向于有相似的语义信息.CoNMF 的创新之处在于利用基础的矩阵分解概念将一致性检测问题转化为寻找多视图的共同基础向量问题,且依据图邻接相似度矩阵的对称性质提出了更为适用的优化目标函数.本文利用两个不同规模的 PPIN 数据来验证 CoNMF 算法的有效性.通过与多种当前常用算法比较,结果显示 CoNMF 整体优于现今主流的基准比较方法,是一种有效融合多种生物信息源的蛋白质功能模块提取方法.

2 相关工作

在此作者先给出问题的简单描述,并简要介绍已往算法,以说明本文算法的新颖之处.对于给定网络 G 存在多个视图描述 $\boldsymbol{A}^{(i)}\in\mathbb{R}_{+}^{N\times N}$,$i=1,2,\cdots,I$,$\boldsymbol{A}^{(i)}$ 是网络 G 的结点邻接相似度矩阵,M 代表视图个数.

与本文算法相关的已有算法包括直接权重融合算法,基于核函数的融合算法和近来比较受关注的集合聚类方法等等.权重融合算法,将 $\boldsymbol{A}^{(i)}$ 的权重线性组合,即 $\boldsymbol{A}=\sum_{i=1}^{I}\boldsymbol{A}^{(i)}$,作为聚类算法的基础数据,以期获得多数据融合的模块检测结果.较有代表性的聚类算法有谱聚类(spectral clustering)[14],MCL 等.

区别于直接的原始数据叠加,第二种较通用的融合算法对多视图数据进行核函数的叠加.首先,经由核函数 $\boldsymbol{\Phi}$ 将原始数据 $\boldsymbol{A}^{(i)}$ 映射到同一特征空间 $\boldsymbol{F}$ 之后将所得特征叠加,即:$\Phi=\sum_{i=1}^{I}\Phi^{(i)}$.其中核函数的选择有多种,较适用于网络模型的之一是谱核函数(spectral kernel)[15],如下式:

$$\Phi^{(i)}=\sum_{k=1}^{d}\boldsymbol{v}_k^{(i)}(\boldsymbol{v}_k^{(i)})^{\mathrm{T}} \tag{1}$$

式中 $\boldsymbol{v}_k^{(i)}$ 是拉普拉斯矩阵 $\boldsymbol{L}^{(i)}$ 的第 k 小的特征值所对应的特征向量;$\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{D}^{-\frac{1}{2}}$ 是 $\boldsymbol{A}$ 的归一化拉普拉斯矩阵,$\boldsymbol{D}$ 中仅对角元素非零,且 $\boldsymbol{D}_{ii}=\sum_{j}\boldsymbol{A}_{ij}$;$d\ll N$ 代表每个矩阵所用到的最小特征值的个数.

集成聚类:集成聚类是一种对知识重新利用的方法,近年来在机器学习和数据挖掘领域获得了很大关注.Strehl 和 Ghosh[16]提出了两种图聚类的集成方法,基于实例(instances)和基于聚类簇(clusters)的方法.两种方法区别在于完成基础聚类后二次聚类对象及其相似度的选取.在进行集合聚类时,基于目标的集成聚类方法(INstance-based clustering ENSemble,INENS)以原实例即原图中结点为集合聚类对象,以结点同落于一个聚类簇的频率为两结点相似度;而基于聚类簇的方法(CLuster-based clustering ENSemble,CLENS)则以基础聚类模型所得聚类结果为对象,以两个聚类簇共同含有的结点数来计算两聚类簇的相似度.Fern[17]等人提出了一种超二分图集成聚类方法(Hybrid Bipartite Graph Formulation,HBGF),该方法将实例和聚类簇同作为二次聚类的结点,二者之间的从属关系构成一个超二分图.Fern 利用谱聚类算法对该超二分图做聚类划分最终得到集成聚类结果.集成聚类方法能够整合多聚类模型或同一聚类模型下不同参数的聚类结果,利用基础聚类的初步知识得到相对稳定的结果.但是这一类算法的一大弊端是集成结果严重依赖于基础聚类模型的选择.

本文方法与核函数聚类方法有一定共通性,但是贡献在于,本文方法通过基础的矩阵分解概念将一致性检测问题转化为寻找多视图的共同基础矩阵因子的

问题.此外,本方法依据图邻接相似度矩阵的对称性质对目标函数进行改进使其更适用于图的聚类问题.

3　多视图下的 CoNMF 算法

传统的 NMF 问题将非负矩阵 $\boldsymbol{A}$ 分解为两个低阶非负矩阵的乘积形式[18],如式(2)所示.

$$\min_{W,H\geq 0}\frac{1}{2}\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\|_{\mathrm{F}}^{2} \tag{2}$$

其中,$\boldsymbol{A}\in\mathbb{R}_{+}^{m\times n}$, $\boldsymbol{W}\in\mathbb{R}_{+}^{m\times k}$, $\boldsymbol{H}\in\mathbb{R}_{+}^{n\times k}$, $k\ll\min\{m,n\}$, $\|\cdot\|_{\mathrm{F}}$ 是 Frobenius 范数. NMF 将原特征空间中的 $\boldsymbol{A}$ 分解为基向量($\boldsymbol{W}$ 中的列向量)的线性组合形式.根据 NMF 的非负特性,我们非常自然地将其应用在聚类问题中,$\boldsymbol{W}$ 矩阵通常被看做原数据 $\boldsymbol{A}$ 行向结点的聚类标示矩阵,当 $\boldsymbol{W}(i,j)=1$,表示 $\boldsymbol{A}$ 中第 i 行结点属于第 j 个聚类.同样,$\boldsymbol{H}^{\mathrm{T}}$ 则表示原数据行向结点的聚类标识,即 $\boldsymbol{H}^{\mathrm{T}}(i,j)=1$,表示 $\boldsymbol{A}$ 第 i 列结点属于第 j 个聚类. NMF 不能同时保证在两个变量上的凸性质,只能得到局部最优解.由此,单就矩阵近似分解而言,相对于其他矩阵低秩近似方法,如奇异值分解方法(Single Value Decomposition,SVD),NMF 对原矩阵 $\boldsymbol{A}$ 的最有近似结果并没有体现出任何优势,但是 NMF 方法的非负特性使其比 SVD 更加适用于现实问题的聚类方法,因为这些问题所涉及的特征、特征相似度等量度都自然地具有非负性质[19].同时,NMF 的非负性质使其具有良好的可解释性,这是 SVD 方法所不具备的.

不同于传统 NMF 聚类,本文中所涉及的蛋白质相似度矩阵 $\boldsymbol{A}^{(i)}$是对称矩阵,$\boldsymbol{A}^{(i)}\in\mathbb{R}_{+}^{n\times n}$, $\boldsymbol{W},\boldsymbol{H}\in\mathbb{R}_{+}^{n\times k}$且二者理论上应相等.为检测多视图中一致信息,CoNMF 将多视图 $\boldsymbol{A}^{(i)}$变换为基向量的线性组合并对基向量空间及线性组合变换加限制条件,使得其满足两个限定目标:

(1)所有视图通过因式分解后获得的聚类标示矩阵相互一致,即与真正的蛋白质功能聚类标示矩阵 $\boldsymbol{H}$;

(2)由于相似度矩阵 $\boldsymbol{A}^{(i)}$为对称矩阵,提现结点聚类结果的标示矩阵 $\boldsymbol{W}$ 与 $\boldsymbol{H}$ 应一致.因此,必须对传统 NMF 方法做相应修改,以保证获得对称标示矩阵并最优还原原相似度矩阵 $\boldsymbol{A}^{(i)}$.

3.1　CoNMF 的实现

基于上述思想,本文提出的目标函数如式(3),采用三因子乘积算多视图的近似分解结果,同时加入限定惩罚因子以引导目标函数得到满足上述两个限定目标的最优结果.

$$\min\frac{1}{2}\sum_{i=1}^{I}\|\boldsymbol{A}^{(i)}-\boldsymbol{H}^{(i)}\boldsymbol{S}^{(i)}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}+\frac{\alpha}{2}\sum_{i=1}^{I}\|\boldsymbol{H}^{(i)}-\boldsymbol{H}^{*}\|_{\mathrm{F}}^{2}$$

$$\text{s.t.}\quad \boldsymbol{H}^{(i)}\geq 0,\boldsymbol{S}^{(i)}\geq 0 \tag{3}$$

其中,$\alpha\geqslant 0$,本文中此经验阈值根据实验限定为 0.45;$i\in(1,\cdots,I)$代表视图个数.$\boldsymbol{H}^{*}$ 代表同时最终拟合多个视图的最优统一解.式中前半部分旨在得到原矩阵的最优逼近分解矩阵,$\boldsymbol{H}$ 依然作为聚类标示矩阵,而 S 集中提取出聚类相互之间的关系,也即相似度.后半部分作为惩罚因子限定了的聚类结果最终趋于一致性最优解,也即各视图的聚类指示矩阵趋于同一最优解.为解决这一优化问题,文本采用迭代更新算法.具体来讲,该迭代算法包括三个步骤:

(1)给定 $\boldsymbol{H}^{(i)}$,$\boldsymbol{H}^{*}$,求解 $\boldsymbol{S}^{(i)}$的迭代更新规则:

令 Ψ 为限定条件 $\boldsymbol{S}^{(i)}\geqslant 0$ 所对应的拉格朗日乘子,L_1 代表目标函数,则目标函数的拉格朗日对偶为:

$$\Gamma_1=L_1+\mathrm{Tr}(\Psi\boldsymbol{S}^{(i)}) \tag{4}$$

其中,Tr 代表了求矩阵的迹.对上式求极值等价于最小化下式:

$$\begin{aligned}\Gamma_1'=&\mathrm{Tr}(\boldsymbol{A}^{(i)}(\boldsymbol{A}^{(i)})^{\mathrm{T}}-2\boldsymbol{A}^{(i)}\boldsymbol{H}^{(i)}\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\\&+\boldsymbol{H}^{(i)}\boldsymbol{S}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)}\boldsymbol{S}(\boldsymbol{H}^{(i)})^{\mathrm{T}}+\\&\alpha(\boldsymbol{H}^{(i)}(\boldsymbol{H}^{(i)})^{\mathrm{T}}-2\boldsymbol{H}^{*}(\boldsymbol{H}^{(i)})^{\mathrm{T}})+\mathrm{Tr}(\Psi\boldsymbol{S}^{(i)})\end{aligned} \tag{5}$$

Γ_1' 函数对 $\boldsymbol{S}^{(i)}$求导,

$$\begin{aligned}\frac{\partial\Gamma_1'}{\partial\boldsymbol{S}^{(i)}}=&-(\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{A}^{(i)}\boldsymbol{H}^{(i)}\\&+(\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)}\boldsymbol{S}^{(i)}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)}+\Psi\end{aligned} \tag{6}$$

根据 KKT 条件,我们得到:

$$\frac{\partial\Gamma_1'}{\partial\boldsymbol{S}^{(i)}}=0,$$

$$\Psi_{nk}\boldsymbol{S}_{nk}^{(i)}=0,\quad 1<n<N,1<k<K \tag{7}$$

由此我们得到 $\boldsymbol{S}^{(i)}$的更新迭代规则:

$$\begin{aligned}\boldsymbol{S}^{(i)}=&((\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)})^{-1}(H^{(i)})^{\mathrm{T}}\\&\boldsymbol{A}^{(i)}\boldsymbol{H}^{(i)}((\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)})^{-1}\end{aligned} \tag{8}$$

(2)固定 $\boldsymbol{S}^{(i)}$,$\boldsymbol{H}^{*}$,求解 $\boldsymbol{H}^{(i)}$的迭代更新规则:

令 Ω 为限定条件 $\boldsymbol{H}^{(i)}\geqslant 0$ 所对应的拉格朗日乘子,L_1代表目标函数,则目标函数的拉格朗日对偶为:

$$\Gamma_2=L_1+\mathrm{Tr}(\Omega\boldsymbol{H}^{(i)}) \tag{9}$$

对上式求极值等价于最小化下式:

$$\begin{aligned}\Gamma_2'=&\mathrm{Tr}(\boldsymbol{A}^{(i)}(\boldsymbol{A}^{(i)})^{\mathrm{T}}-2\boldsymbol{A}^{(i)}\boldsymbol{H}^{(i)}\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\\&+\boldsymbol{H}^{(i)}\boldsymbol{S}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)}\boldsymbol{S}(\boldsymbol{H}^{(i)})^{\mathrm{T}}+\alpha(\boldsymbol{H}^{*}(\boldsymbol{H}^{*})^{\mathrm{T}}\\&+\boldsymbol{H}^{(i)}(\boldsymbol{H}^{(i)})^{\mathrm{T}}-2\boldsymbol{H}^{*}(\boldsymbol{H}^{(i)})^{\mathrm{T}})+\mathrm{Tr}(\Omega\boldsymbol{H}^{(i)})\end{aligned} \tag{10}$$

我们通过构造辅助函数的方法,求解 $\boldsymbol{H}^{(i)}$的更新迭代规则如式(11),其具体收敛性证明见下小节.

$$\boldsymbol{H}_{nk}^{(i)}=\boldsymbol{H}_{nk}^{(i)}\left(\frac{2(\boldsymbol{A}^{(i)}\boldsymbol{H}^{(i)}\boldsymbol{S}^{(i)})_{nk}+\alpha\boldsymbol{H}_{nk}^{*}}{2(\boldsymbol{H}^{(i)}\boldsymbol{S}^{(i)}(\boldsymbol{H}^{(i)})^{\mathrm{T}}\boldsymbol{H}^{(i)}\boldsymbol{S}^{(i)})_{nk}+\alpha\boldsymbol{H}_{nk}^{(i)}}\right)^{\frac{1}{4}} \tag{11}$$

(3)给定 $\boldsymbol{H}^{(i)}$,$\boldsymbol{S}^{(i)}$,求解 $\boldsymbol{H}^{*}$.

对 $\boldsymbol{H}^{*}$的迭代更新,采用式(12),其推导过程类似

于式(8)的推导.

$$\boldsymbol{H}^{*} = \frac{1}{l}\sum_{i=1}^{l}\boldsymbol{H}^{(i)} \tag{12}$$

重复迭代以上三个步骤,最终达到目标函数的收敛.值得一提的是,NMF 的非负性使其易于拓展至概率模型,本文的方法利用了这一性质获得可重叠聚类.将迭代结果做归一化至[0,1]之间,并以一定阈值取舍得到最终提现重叠性质的聚类结果.

3.2 收敛性证明

为证明上述关于 $\boldsymbol{H}$ 的更新迭代规则的收敛性,本文构造了 $\Gamma_2(\boldsymbol{H})$的辅助函数并证明其凸性质.证明过程用到以下两个引理:

引理 1 给定任意非负矩阵 $\boldsymbol{A}\in\mathbb{R}_{+}^{n\times n}$,$\boldsymbol{B}\in\mathbb{R}_{+}^{k\times k}$,$\boldsymbol{S}\in\mathbb{R}_{+}^{n\times k}$和 $\boldsymbol{S}'\in\mathbb{R}_{+}^{n\times k}$,$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为对称矩阵,且满足以下不等关系[28]:

$$\sum_{ip}\frac{(\boldsymbol{AS}'\boldsymbol{B})\boldsymbol{S}_{ip}^{2}}{\boldsymbol{S}_{ip}'} \geqslant \mathrm{Tr}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{ASB}) \tag{13}$$

引理 2 给定任意非负对称矩阵 $\boldsymbol{A}\in\mathbb{R}_{+}^{k\times k}$和 $\boldsymbol{B}\in\mathbb{R}_{+}^{k\times k}$,对于 $\boldsymbol{H}\in\mathbb{R}_{+}^{n\times k}$下述不等关系成立[29]:

$$\mathrm{Tr}(\boldsymbol{HAH}^{\mathrm{T}}\boldsymbol{HBH}^{\mathrm{T}}) \leqslant \sum_{ik}\left(\frac{\boldsymbol{H}'\boldsymbol{AH}^{\mathrm{T}}\boldsymbol{H}'\boldsymbol{B}+\boldsymbol{H}'\boldsymbol{BH}'^{\mathrm{T}}\boldsymbol{H}'\boldsymbol{A}}{2}\right)_{ik}\frac{\boldsymbol{H}_{ik}^{4}}{\boldsymbol{H}_{ik}'^{3}} \tag{14}$$

首先,定义 $\Gamma_2(\boldsymbol{H})$的辅助函数 $Z(\boldsymbol{H},\boldsymbol{H}')$:

$$\begin{aligned}Z(\boldsymbol{H},\boldsymbol{H}') = {} & \sum_{ij}\boldsymbol{Y}_{ij} - 2\sum_{ijkl}\boldsymbol{H}_{ji}'\boldsymbol{S}_{jk}\boldsymbol{H}_{kl}'\boldsymbol{A}_{li}\left(1+\log\frac{\boldsymbol{H}_{ji}\boldsymbol{H}_{kl}}{\boldsymbol{H}_{ij}'\boldsymbol{H}_{kl}'}\right)\\ & + \sum_{ij}(\boldsymbol{H}'\boldsymbol{SH}'^{\mathrm{T}}\boldsymbol{H}'\boldsymbol{S})_{ij}\frac{\boldsymbol{H}_{ij}^{4}}{\boldsymbol{H}_{ij}'^{3}}\\ & + \alpha\sum_{ij}\boldsymbol{H}_{ij}'\frac{\boldsymbol{H}_{ij}^{2}}{\boldsymbol{H}_{ij}'} - 2\alpha\sum_{ij}\boldsymbol{H}_{ij}^{*}\boldsymbol{H}_{ij}'\left(1+\log\frac{\boldsymbol{H}_{ij}}{\boldsymbol{H}_{ij}'}\right)\end{aligned} \tag{15}$$

其中 $\boldsymbol{Y}=\boldsymbol{A}^{(i)}(\boldsymbol{A}^{(i)})^{\mathrm{T}}+\alpha(\boldsymbol{H}^{*}(\boldsymbol{H}^{*})^{\mathrm{T}}$.得到上式的中间推导依据有:

$$\begin{aligned}& u\leqslant 1+\log u\\ & \mathrm{Tr}(\boldsymbol{AHSH}^{\mathrm{T}}) \geqslant \sum_{ijkl}\boldsymbol{H}_{ji}'\boldsymbol{S}_{jk}\boldsymbol{H}_{kl}'\boldsymbol{A}_{li}\left(1+\log\frac{\boldsymbol{H}_{ji}\boldsymbol{H}_{kl}}{\boldsymbol{H}_{ij}'\boldsymbol{H}_{kl}'}\right)\\ & \mathrm{Tr}(\boldsymbol{HH}^{\mathrm{T}}) \leqslant \sum_{ij}\boldsymbol{H}_{ij}'\frac{\boldsymbol{H}_{ij}^{2}}{\boldsymbol{H}_{ij}'}\\ & \mathrm{Tr}(\boldsymbol{H}^{*}\boldsymbol{H}^{\mathrm{T}}) \geqslant \sum_{ij}\boldsymbol{H}_{ij}^{*}\boldsymbol{H}_{ij}'\left(1+\log\frac{\boldsymbol{H}_{ij}}{\boldsymbol{H}_{ij}'}\right)\end{aligned} \tag{16}$$

明显有 $Z(\boldsymbol{H},\boldsymbol{H}')\geqslant\Gamma_2(\boldsymbol{H})$且 $Z(\boldsymbol{H},\boldsymbol{H})=\Gamma_2(\boldsymbol{H})$,因此 $Z(\boldsymbol{H},\boldsymbol{H}')$是 $\Gamma_2(\boldsymbol{H})$的辅助构造函数.明显可知 $Z(\boldsymbol{H},\boldsymbol{H}')$的 Hessian 矩阵是半正定的也即 $Z(\boldsymbol{H},\boldsymbol{H}')$为凸函数.通过求解$\partial Z(\boldsymbol{H},\boldsymbol{H}')/\partial\boldsymbol{H}=0$得到 $\boldsymbol{H}$ 的更新迭代式(11)也即使 $\Gamma_2(\boldsymbol{H})$收敛于全局最小值的迭代规则.

4 相似度计算

本节将介绍上文中蛋白质功能相似度 $\boldsymbol{A}^{(i)}$的计算方法.

4.1 基因共表达相关系数

共表达的蛋白质倾向参与同种生命功能活动,属于同一功能模块,所以基因表达数据常被用来计算蛋白质功能相似度,弥补 PPIN 的不可靠性[3,4].本文采用皮尔森相关系数计算基因表达模式的相似程度(归一化至 0~1 范围).基因共表达相关系数用 $\boldsymbol{C}$ 表示,并与 PPIN 相结合:

$$\boldsymbol{w}_1(p_i,p_j)=\boldsymbol{C}(p_i,p_j)\times\boldsymbol{G}(p_i,p_j) \tag{17}$$

其中,p_i,p_i代表任两蛋白质,$\boldsymbol{G}$ 是描述 PPIN 的邻接矩阵,$\boldsymbol{w}_1$ 是将基因共表达相关系数与原蛋白质网络相结合后所得相似性矩阵.共表达分析是降低 PPIN 噪声的一个有力工具,但是基因表达谱数据自身的噪声,是其应用中的一大问题.结合 GO 功能注释信息能够帮助弥补基因表达谱数据中的不可靠因素.

4.2 GO 功能相似度系数

GO 是一个知识体系,概括了基因所参与的生物过程(Biological Process,BP),发挥的分子功能(Molecular Function,MF)和所处在的细胞位置(Cellular Component,CC)[13].GO 将这些知识根据从属关系组织成有向无环图(directed acyclic graph).GO 功能相似测量主要有两种方法:基于路径长度的方法和信息量方法.计算两个注释条目在 GO 结构中的最短路径长度,是最简单的比较两条目间相似度的方法.但是两注释条目间在 GO 系统结构中的路径长度并不能真正功能上的相似度.而从另一方面,基于信息量的方法以计算注释条目的信息量(Information Content,IC)来测量比较两条目间共通的信息.给定 GO 子集中任一个条目 c,θ 为从属于 c 的所有子条目的集合,$p(c)$表示一个基因被 θ 中任一条目所注释的概率,并以负对数形式表示该条目所包含的信息量,即 $-\log(p(c))$.因此,如果条目 c 是所在子集结构的根条目,则 θ 包含属于该子集的所有条目,因此 $p(c)$必然为 1,而该条目 c 的信息量为 0.计算两个条目的功能相似度,就是计算二者最低层的共有父条目结点的信息量大小[20],加入归一化因子后其计算方法如下[21]:

$$\boldsymbol{O}(c_i,c_j)=\max_{c\in P_a(c_i,c_j)}\frac{2\times[\log(p(c))]}{\log(p(c_i))+\log(p(c_j))} \tag{18}$$

式中,$P_a(c_i,c_j)$是 c_i,c_j 共有父条目的集合.

基因通常包含多条 GO 注释,因此计算两基因或基因产物的功能相似度,要考虑所有注释条目.我们用式(14)挑选其中一基因所包含的任一注释与另一基因所对应的注释集 θ_j 中具有最大相似度的注释,累计最大

注释相似度并取均值.

$$\boldsymbol{O}'(p_i,p_j)=\frac{1}{U\times W}\Big(\sum_{u\in\theta_i}\max_{w\in\theta_j}(\boldsymbol{O}(c_u,c_w))+\sum_{w\in\theta_j}\max_{u\in\theta_i}(\boldsymbol{O}(c_u,c_w))\Big)\tag{19}$$

其中 θ_j,θ_i 是两个蛋白质所对应的注释集,$U=|\theta_i|$,$W=|\theta_j|$.基因功能相似系数,作为 PPIN 的另一权重,与其邻接矩阵结合获得第二个相似性矩阵:

$$\boldsymbol{w}_2(p_i,p_j)=\boldsymbol{O}'(p_i,p_j)\times\boldsymbol{G}(p_i,p_j)\tag{20}$$

5　实验与讨论

5.1　实验数据与实验说明

本文分别在两组 PPIN 数据上对所提出进行验证.其一来自 Gavin 等人的工作[22],该 PPI 信息来源于串联亲和纯化(Tandem Affinity Purification,TAP),包含 2,551 个蛋白质和 21,413 个相互作用.TAP 数据相对准确,近年来更新较不频繁,本文对其功能模块的检测与另一组来自 Biogrid 数据库做对比.Biogrid 数据库更新及时,PPI 数据较全面.在本文实验中,我们剔除了 Biogrid 中不存在对应基因表达的蛋白质,因此得到了一个包含 4,531 结点和其间 143,226 个相互作用的 PPIN.本文所用的 GO 数据下载自 http://www.geneontology.org/.GO 注释信息中,存在一些尚没有交叉验证的信息,为了取得更准确的相似度信息,本文单独提取并使用仅提取经实验或专家验证后的注释条目,即 IDA,IEP,IGI,IMP,IPI,RCA 和 TAS 证据码下的 GO 注释,而排除 IC,IEA,ISS,NAS 及 ND 条目[23].基因表达数据来自 GEO(Gene Expression Omnibus)数据库,其检索号为 GSE12055[24].

对于实验结果的验证,本文根据 CYC2008[25] 数据集做校验标准.该数据集包含了 408 个酵母蛋白质功能模块,所有模块均经由小规模试验验证或有相关文献支持.本文用查全率(Recall rate,Rec)、查准率(Precision rate,Prec)和二者融合的 F-measure 值[11,26] 验证算法检测到的功能模块与 CYC2008 标准数据集的一致性.上述衡量指标的计算方法如下:

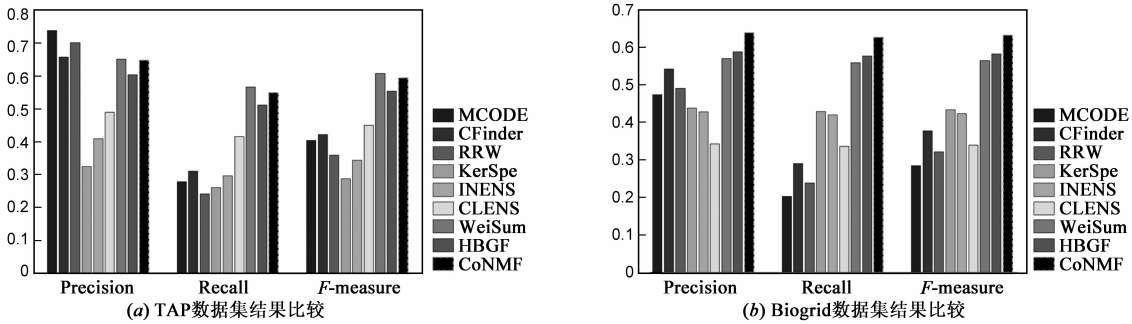
$$\text{Prec}=\text{TP}/(\text{TP}+\text{FN}),\quad \text{Rec}=\text{TP}/(\text{TP}+\text{FP}),\quad F\text{-measure}=\frac{2\times\text{Prec}\times\text{Rec}}{\text{Prec}+\text{Rec}}\tag{21}$$

其中,TP 代表 True Positive,真正性结果,即算法检测到的正确的模块;FP 代表 False Positive,假正性结果,即算法未检测到但在标准数据中存在的正确模块;FN 为 False Negative,假负性结果,算法检测到但标准数据中不存在的模块.

5.2　结果分析与讨论

本文从多种角度选取了比较方法.其中 MCODE,CFinder,RRW[27] 均仅利用 PPIN 的拓扑结构.MCODE 采用一种贪婪算法将权重 PPIN 划分为紧密联系的子区域.CFinder 基于全连通团(clique)的概念扩张网络中团成员形成功能模块.RRW 是通过可重启的随机行走方法寻找起始节点可到达的结点组成功能模块.

如图 1 所示,本文所提出算法 CoNMF 在 Biogrid 数据集上达到最高的 F-measure 值,综合指标明显优于其他比较算法.基于 PPIN 拓扑结构的三种比较算法在 TAP 中精确度较高,但在 Biogrid 数据集上其结果则相对不及其他算法.一个合理的解释是,TAP 数据相对较精确且数据集较小,基于 PPIN 拓扑结构的算法较易受数据集大小和噪声的影响,检测结果不稳定.同时,本文发现直接以权重叠加不同视图的 WeiSum 算法同样在 TAP 数据上表现较为理想.WeiSum 对 Biogrid 数据明显优于 MCODE 等三种算法,这说明了多视图融合的必要性.比较本文算法与其他数据融合算法及聚类集成算法,CoNMF 在 Biogrid 数据的检测结果要好于其他算法,其提高程度明显优于在 TAP 数据上的表现.由此我们能够得出结论,本文所提出算法对噪声较大且数据集也较大的情况下提高较为显著.



(a) TAP数据集结果比较　　(b) Biogrid数据集结果比较

图1　比较结果

表 1 具体展示了比较方法在两数据集上的结果,除查准率、查全率和 F-measure 值外,以此统计了准确查找到的功能模块数、模块平均大小、准确查找到的模块覆盖蛋白质数目和属于多个模块的蛋白质数目.本文算法所检测出的功能模块平均大小更大,覆盖更多的蛋白质,同时所得结果允许某些蛋白质存在于多个功

能模块,解决了模块的重叠问题.

表 1 算法比较

	Clustering models	Precision	Recall	*F*-measure	Matched modules	Average size	Cover proteins	Overlapping proteins
TAP	MCODE	0.738	0.278	0.404	98	4.73	465	—
	Cfinder	0.657	0.310	0.421	110	8.38	920	—
	RRW	0.701	0.241	0.359	85	7.59	648	—
	KerSpe	0.324	0.260	0.288	92	7.2	699	—
	INENS	0.409	0.297	0.344	105	8.16	849	—
	CLENS	0.490	0.415	0.450	147	8.34	1226	—
	WeiSum	0.651	0.567	0.608	201	5.97	1199	—
	HBGF	0.603	0.511	0.554	181	9.14	1654	—
	CoNMF	0.647	0.548	0.593	194	13.37	1947	524
Biogrid	MCODE	0.474	0.203	0.284	83	6.59	545	—
	Cfinder	0.541	0.289	0.377	118	7.63	901	—
	RRW	0.490	0.239	0.321	97	7.85	764	—
	KerSpe	0.437	0.428	0.433	175	7.39	1292	—
	INENS	0.428	0.419	0.423	171	9.24	1580	—
	CLENS	0.342	0.335	0.338	137	8.79	1202	—
	WeiSum	0.570	0.559	0.564	228	6.44	1468	—
	HBGF	0.588	0.576	0.582	235	11.68	2745	—
	CoNMF	0.618	0.605	0.611	247	14.21	2803	603

以 *P*-value 为基准分析所得模块的 GO 富集度,*P*-value 值越小说明该功能模块越可能具有该 GO 条目所注释的功能.一般文献中以 *P*-value < 0.05 为准.图 2 为随机选取的模块实例,其中模块 1 ~ 6 的 GO 富集情况见表 2,可以发现本文所提取的蛋白质功能模块在 GO 富集度上表现出极高的一致性.由此我们有理由认为,CoNMF 能够发现具有实际生物学意义的蛋白质功能模块.

表 2 部分结果的 GO 注释富集度分析(TAP)

Module (size)	GO (Biological Process)	GO-ID	*P*-value	Protein members
01(16)	proteasomal ubiquitin-independent protein catabolic process	10499	2.43E-36	YKL168C YFL007W YML092C YPR103W YGL011C YJL001W YOR362C YOL038W YIL009C-A YFR050C YMR314W YGR135W YER094C YER012W YBL041W YGR253C
	proteasomal protein catabolic process	10498	3.74E-22	
	proteasomal ubiquitin-dependent protein catabolic process	43161	3.74E-22	
	modification-dependent protein catabolic process	19941	4.98E-18	
	ubiquitin-dependent protein catabolic process	6511	4.98E-18	
	proteolysis involved in cellular protein catabolic process	51603	1.00E-17	
	modification-dependent macromolecule catabolic process	43632	1.72E-17	
	protein catabolic process	30163	5.93E-17	
02(26)	mitochondrial reanslation	32543	2.53E-33	YDR036C YBR251W YDR347W YHL004W YNL186W YDR041W YPL013C YJR113C YGR084C YPL118W YMR188C YNL306W YDR337W YBR146W YKL155CYGL129C YIL093C YKL003C YBL090W YNR037C YJR101W YGR215W YHR059W YDL045W-A YGR170W YGR150C
	mitochondrion organization	7005	6.18E-26	
	translation	6412	2.31E-13	
	organelle organization	6996	2.09E-12	
03(24)	nuclear mRNA splicing, via spliceosome	398	2.53E-33	YER029C YDL087C YGR013W YDR240C YIL061C YKL012W YDR235W YHR086W YGR074W YPR182W YFL017W-A YBR119W YMR125W YLR298C YPL178W YER165W YJR084W YML046W YPR094W YLR275W YLR147C YDR432W YFL003C YKL214C
	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	377	6.18E-26	
	RNA splicing, via transesterification reactions	375	2.31E-13	
	RNA splicing	8380	2.01E-32	

续表

Module (size)	GO (Biological Process)	GO-ID	*P*-value	Protein members
04(12)	double-strand break repair via break-induced replication	727	2.64E-12	YLR274W YMR192W YGL201C YDR161W YBR202W YLR433C YGL001C YBR109C YEL032W YOL146W YPL153C YDR489W
	double-strand break repair via homologous recombination	724	5.93E-11	
	DNA-dependent DNA replication	6261	8.59E-11	
	recombinational repair	725	2.02E-10	
05(16)	DNA repair	6281	2.59E-15	YCR092C YML032C YDL156W YPR065W YDR097C YJL173C YAR007C YHR164C YNL312W YBR136W YDR499W YJR144W YIR002C YMR190C YLR234W YER104W
	DNA replication	6260	2.61E-15	
	response to DNA damage stimulus	6974	4.80E-15	
	anatomical structure homeostasis	60249	7.94E-15	
06(23)	tRNA transcription from RNA polymerase III promoter	42797	8.32E-40	YDL150W YNL248C YJR063W YKL144C YOR116C YDR005C YPR110C YOR207C
	tRNA transcription	9304	8.32E-40	YOR340C YNR003C YPR190C YOR341W
	transcription from RNA polymerase III promoter	6383	2.79E-36	YPR187W YOR224C YJL011C YNL113W
	transcription, DNA-dependent	6351	6.95E-29	YNL151C YKR025W YBR154C YOR210W YDR045C YPR032W YGL156W

6　结论

本文提出了一种针对多生物数据源融合的一致功能模块挖掘算法,该算法的贡献在于针对多视图的对称矩阵,通过非负矩阵分解同时分析多视图数据,以基础特征空间的线性变换来近似原数据,从而得到了一致的聚类结果.实验结果显示,本文方法在噪声较多数量较大的数据集上对结果准确度有较大提高,且所获得功能模块具有实际生物意义,为蛋白质功能分析提供了辅助信息.CoNMF 虽是针对蛋白质相互作用数据提出的多视图融合算法,但该算法思想可拓展至社交网络、Web 数据等多种关系网络,是一种较为有效的融合多视图信息的网络模块挖掘算法.

参考文献

[1] Bonetta L. Protein-protein interactions: Interactome under construction[J]. Nature, 2010, 468(7325): 851 – 854.

[2] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks[J/OL]. BMC Bioinformatics, 2003, 4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC149346/pdf/1471-2105-4-2.pdf, 2003-01-13/2013-11-19.

[3] Young-Rae C, Woochang H, Aidong Z. Efficient modularization of weighted protein interaction networks using k-hop graph reduction[A]. Bioinformatics and Bioengineering, 2006 BIBE 2006 Sixth IEEE Symposium on[C]. Virginia: IEEE, 2006. 289 – 98.

[4] Ulitsky I, Shamir R. Identifying functional modules using expression profiles and confidence-scored protein interactions[J]. Bioinformatics, 2009, 25(9): 1158 – 1164.

[5] Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data[J]. Journal of Molecular Biology, 2003, 327(5): 919 – 923.

[6] Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks [J]. Bioinformatics, 2006, 22(8): 1021 – 1023.

[7] Oti M, Brunner HG. The modular nature of genetic diseases [J]. Clincal Genetics, 2007, 71(1): 1 – 11.

[8] Chen J, Yuan B. Detecting functional modules in the yeast protein protein interaction network[J]. Bioinformatics, 2006, 22(18): 2283 – 2290.

[9] Segal E, Wang H, Koller D. Discovering molecular pathways from protein interaction and gene expression data[J]. Bioinformatics, 2003, 19(suppl 1): i264 – i272.

[10] Li M, Wu XH, Wang JX, Pan Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data[J/OL]. BMC Bioinformatics, 2012, 13. http://www.biomedcentral.com/content/pdf/1471-2105-13-109.pdf, 2012-05-23/2013-11-19.

[11] Young-Rae C, Lei S, Aidong Z. FlowNet: flow-based approach for efficient analysis of complex biological networks [A]. Data Mining, 2009 ICDM '09 Ninth IEEE International Conference on[C]. Florida: IEEE, 2009. 91 – 100.

[12] Consortium GO. The Gene Ontology (GO) database and informatics resource[J]. Nucleic Acids Research, 2004, 32(suppl 1): D258 – D261.

[13] du Plessis L, Skunca N, Dessimoz C. The what, where, how and why of gene ontology-a primer for bioinformaticians[J]. Briefings in Bioinformatics, 2011, 12(6): 723 – 735.

[14] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395 – 416.

[15] Smola AJ, Kondor R. Kernels and regularization on graphs [J]. Learning Theory and Kernel Machines, 2003, 2777: 144 – 158.

[16] Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse

framework for combining multiple partitions[J]. J Mach Learn Res,2003,3(1):583 – 617.
[17] Fern XZ, Brodley CE. Solving cluster ensemble problems by bipartite graph partitioning[A]. Proceedings of the Twenty-first International Conference on Machine Learning[C]. New York: ACM,2004. 36 – 45.
[18] Kim H, Park H. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method[J]. SIAM Journal on Matrix Analysis and Applications,2008,30(2):713 – 730.
[19] Zhang T, Shen R, Lu H. Using non-negative matrix factorization to cluster learners and construct learning communities[J]. Chinese Journal of Electronics,2011,20(2):207 – 211.
[20] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization[A]. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval[C]. New York: ACM, 2003. 267 – 273.
[21] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language[J]. J Artif Intell Res,1999,11(1):95 – 130.
[22] Lin D. An information-theoretic definition of similarity[A]. Proceedings of the Fifteenth International Conference on Machine Learning[C]. Scotland: Springer,1998. 296 – 304.
[23] Gavin A-C. Proteome survey reveals modularity of the yeast cell machinery[J]. Nature,2006,440(30):631 – 636.
[24] Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations[J]. Nat Rev Genet, 2008, 9(7):509 – 515.
[25] Anderson JB, Sirjusingh C, Syed N, Lafayette S. Gene expression and evolution of antifungal drug resistance[J]. Antimicrob Agents Chemother,2009,53(5):1931 – 1937.
[26] Pu SY, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes[J]. Nucleic Acids Research,2009,37(3):825 – 831.
[27] Jianxin W, Min L, Jianer C, Yi P. A fast hierarchical clustering algorithm for functional modules dscovery in protein interaction networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2011,8(3):607 – 620.
[28] Macropol K, Can T, Singh AK. RRW: Repeated random walks on genome-scale protein networks for local cluster discovery [J/OL]. BMC Bioinformatics, 2009, 10. http://www.biomedcentral.com/content/pdf/1471-2105-10-283.pdf, 2009-09-09/2013-11-19.
[29] Ding C, Li T, Peng W, Park H. Orthogonal nonnegative matrix t-factorizations for clustering[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Philadelphia: ACM,2006. 126 – 135.
[30] Wang H, Huang H, Ding C. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization[A]. Proceedings of the 20th ACM International Conference on Information and Knowledge Management[C]. Glasgow: ACM,2011. 279 – 284.

作者简介

张 媛 女,1985 年 10 月出生于河北省深州市. 现为北京工业大学博士生. 研究方向为数据挖掘、生物信息处理等.
E-mail: zhangyuan@emails.bjut.edu.cn

贾克斌 男,1962 年 8 月出生于河南省安阳市. 现为北京工业大学电子信息与控制工程学院教授. 从事多媒体智能处理、生物信息学等方面的研究.
E-mail: kebinj@bjut.edu.cn

Aidong Zhang 女,1994 年获得普渡大学博士学位. 现为美国纽约州立大学布法罗分校计算机科学与工程系教授和系主任. 主要研究方向为数据挖掘和生物信息学.
E-mail: azhang@buffalo.edu