

## 组织特异性蛋白质复合体的识别

丁 霞<sup>1</sup>, 张晓飞<sup>2</sup>, 易 鸣<sup>3</sup>

(1. 武汉大学数学与统计学院, 湖北 武汉 430072)

(2. 华中师范大学数学与统计学学院, 湖北 武汉 430079)

(3. 华中农业大学理学院, 湖北 武汉 430070)

**摘 要:** 本文研究了组织特异性蛋白质复合体的识别问题. 利用蛋白质相互作用网络数据以及组织特异性基因表达数据构建组织特异性蛋白网络, 利用多种代表性聚类算法对该网络进行聚类, 并利用非负矩阵分解对聚类结果进行合并聚类, 得到了组织特异性蛋白质复合体. 结果表明, 聚类效果得到明显提升, 并且能识别出组织特异性蛋白质复合体.

**关键词:** 蛋白质相互作用网络; 复合体识别; 组织特异性; 非负矩阵分解

MR(2010) 主题分类号: 92B05

中图分类号: O212.4; O212.5

文献标识码: A

文章编号: 0255-7797(2017)05-1093-08

### 1 引言

在现如今的后基因组时代, 对细胞间模块以及基因的关系进行系统分析和全面了解是一个非常重要的课题. 随着生物信息学的高速发展, 基因组学中大规模的高通量技术, 如基于质谱的串联亲和纯化<sup>[1,2]</sup>、酵母双杂交<sup>[3,4]</sup>以及蛋白芯片技术为我们提供了海量的大规模生物网络, 也为我们对生物网络进行系统的分析创造了可能.

众所周知, 蛋白质很少单独行动, 它们往往结合在一起形成复合体在生命体中进行生物功能<sup>[5]</sup>. 蛋白质复合体的综合研究有助于揭示蛋白质-蛋白质相互作用网络的结构、预测蛋白质的功能, 更有助于阐明各种疾病的细胞机制<sup>[6]</sup>. 经过 10 多年的快速发展, 已经涌现出了许多基于不同聚类机理的蛋白质相互作用网络功能模块检验方法.

尽管在此方面已经有不少研究, 但是这些方法主要关注静态的蛋白质相互作用网络, 而忽略了蛋白质功能作用的动态变化及组织特异机制. 幸运的是, DNA 微阵列技术的出现, 使数以千计的基因的差异表达的各种实验条件被同时且定量监视, 它提供了许多有关于时间以及组织特异的信息<sup>[7]</sup>. 目前也有少许算法研究动态网络, 并探测动态复合体, 但还没有算法涉及到组织特异的复合体探测.

本文通过结合组织特异性基因表达数据以及人类蛋白质相互作用网络构建出一系列组织特异性蛋白网络, 尝试探索组织特异功能模块的研究. 本文的主要方法为对所构建的组织特异性蛋白网络利用多种方法对其进行聚类, 并对结果进行组装, 最后使用非负矩阵分解模型对组装的结果进行有效合并. 实验结果表明, 本文的方法与其他聚类方法相比, 在检测蛋白质复合体上结果更好.

\*收稿日期: 2015-01-06

接收日期: 2015-05-06

基金项目: 国家自然科学基金资助 (11275259); 国家自然科学基金资助 (91330113).

作者简介: 丁霞 (1990-), 女, 湖北鄂州, 硕士, 主要研究方向: 生物信息.

因为组织特异性蛋白复合体对于理解生物学功能以及确定生物标志物和功能靶标十分重要<sup>[8]</sup>, 因此探索组织特异功能模块很有必要.

## 2 方法

在本节中, 本文首先介绍如何构建组织特异性蛋白网络, 随后介绍如何检测组织特异性复合体.

### 2.1 构建组织特异性蛋白网络

组织特异性蛋白网络是结合蛋白质相互作用网络以及组织特异性基因表达数据两者来构建的. 给定一个 PPI 网络, 可以用图  $G = (V, E)$  来表示<sup>[9]</sup>, 其中  $V$  包含  $|V| = N$  个蛋白质, 而  $E$  包含  $|E|$  条边. 图  $G$  可以表示成一个邻接矩阵  $A$ , 其中若有一条边连接蛋白质  $i$  与  $j$ , 则  $A_{ij} = 1$ , 否则  $A_{ij} = 0$ , 在这种情况下, 识别蛋白质复合体这一问题就转化为点的聚类问题. 组织特异性基因表达数据是这  $N$  个蛋白质在  $T$  个组织中的基因水平, 可以用一个  $N \times T$  维矩阵  $F$  表示.

本文将利用矩阵  $A$  以及矩阵  $F$  来构建组织特异性蛋白网络. 若蛋白质  $i$  与  $j$  有相关关系, 即  $A_{ij} = 1$ , 并且在组织  $t$  中, 蛋白质  $i$  与蛋白质  $j$  均显著表达, 即  $F_{it} > 0$  并且  $F_{jt} > 0$ , 则蛋白质  $i$  与蛋白质  $j$  在组织  $t$  中存在相关关系. 根据上述方法, 对  $T$  个组织进行构建, 则可得到  $T$  个组织特异性蛋白网络.

### 2.2 识别组织特异性蛋白质复合体

在本节中, 本文先对组织特异性蛋白质相关关系网络中的每一个组织分别使用基本聚类方法, 并使用非负矩阵分解模型来合并相似组织特异性蛋白质复合物, 得到新的复合体, 算法的基本流程如图 1 所示.

#### 2.2.1 基本聚类方法

本文首先利用 7 种基本的聚类方法分别对这  $T$  个组织特异蛋白网络进行聚类, 构建蛋白质复合体, 所用的 7 种方法分别为 MCL、MCODE、MINE、ClusterONE、DPCLUS、SPICi、CoAch.

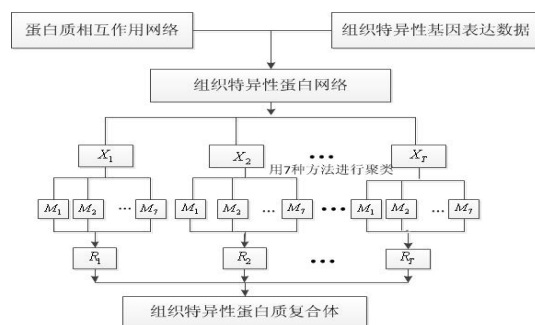


图 1: 识别组织特异性蛋白质复合体的流程图

MCL 是通过模拟在 PPI 网络中流的自由行走来检测蛋白质复合体的经典算法, 它定义了指派节点概率的 Expansion 操作和改变节点游走概率的 Inflation 操作来模拟随机游走的扩展和收缩行为<sup>[10,11]</sup>.

MCODE 是一种基于蛋白质的连接值来检测蛋白质复合体的计算方法, 它首先利用节点的局部邻域密度给 PPI 网络中每个节点进行加权, 然后选择具有最高权值的节点作为初始聚类的种子节点, 并由种子节点向外扩张形成最后的簇 (蛋白质模块)<sup>[11,14]</sup>.

MINE 是一种类似于 MCODE 的凝聚聚类算法, 但它使用了一个改进的顶点加权策略, 并且可以衡量网络模块性, 而这两者都有助于避免使用生长群内包含的临界点来定义模块的边界<sup>[13]</sup>.

DPCLUS 是一种通过簇边界的跟踪进行聚类的算法, 它不仅利用模块密度而且利用新定义的粗特性 CP 完成复合体检验<sup>[11,14]</sup>.

ClusterONE 是一种能识别带重叠的蛋白质复合体的一种算法, 它依赖于重叠领域扩张<sup>[15]</sup>.

CoAch 是一种利用核心依附关系进行复合体检测的算法, 该算法分为两个阶段, 第 1 阶段从邻接图中定义核心顶点, 然后从中检测蛋白质复合体的核心蛋白质, 第 2 阶段为将附属蛋白质逐个连接到核心蛋白质所代表的复合体中<sup>[11,16]</sup>.

SPICi 是一种高效算法, SPICi 种子集群根据其加权度的节点, 如果支撑足够高, 并且集群的密度低于用户定义的阈值, 则此非集群节点将会添加到集群中, 否则, 群集被输出, 这个簇的节点将会从网络中移除<sup>[17]</sup>.

### 2.2.2 非负矩阵分解模型

对每一个组织, 分别使用上述 7 种聚类方法, 可以得到 7 个复合体矩阵  $V_1, V_2, \dots, V_7, V_i$  ( $i = 1 : 7$ ) 为  $N \times P_i$  ( $i = 1 : 7$ ) 矩阵, 其中  $N$  代表蛋白质的个数,  $P_i$  为第  $i$  种聚类方法所识别的蛋白质复合体的个数. 对于矩阵  $V_i$ , 若蛋白质  $N_i, N_j, \dots, N_k$  组成第  $e$  个复合体 ( $1 \leq e \leq P_i$ ), 则在第  $e$  列中, 除了蛋白质  $N_i, N_j, \dots, N_k$  所对应的元素为 1 外, 其余元素为 0. 将这 7 个复合体矩阵  $V_1, V_2, \dots, V_7$  横向排列, 得到矩阵  $V = [V_1, V_2, \dots, V_7]$ ,  $V$  为  $N$  行  $P$  列的矩阵 ( $P = \sum_{i=1}^7 P_i$ ), 依造此方法, 可构建出  $T$  个矩阵.

接着, 我们使用了非负矩阵分解模型来合并相似瞬时蛋白质复合物. 它提供了一种对非负矩阵的低秩逼近, 并且已被广泛地运用到聚类当中<sup>[18,19]</sup>. Lee 和 Seung 的非负矩阵分解方法, 设定模型为

$$\min_{W, H} \sum_{i,j} (V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}).$$

利用更新法则

$$H_{au} \leftarrow H_{au} \frac{\sum_i (W_{ia} V_{iu} / (WH)_{iu})}{\sum_k W_{ka}}, \quad W_{ia} \leftarrow W_{ia} \frac{\sum_u (H_{au} V_{iu} / (WH)_{iu})}{\sum_v H_{av}}.$$

最后得到矩阵  $W$  ( $N \times K$ ) 和  $H$  ( $K \times P$ ), 本文只对矩阵  $W$  进行研究, 将其横向归一, 即令  $U_{ik} = W_{ik} / W_{i.}$ . 得到  $U$  之后, 设定过滤阈值  $\tau$ , 若  $U_{ij} > \tau$ , 则蛋白质  $N_i$  是复合体  $K_j$  的组成部分. 由上可知, 本次算法共有两个参数, 所识别的复合体的个数  $K$  以及过滤阈值  $\tau$ . 由于

复合体大多是由 3 个及 3 个以上的蛋白质组合而成, 因此对所识别出的复合体进行过滤, 将蛋白质个数  $<2$  的复合体舍去.

### 3 结果

#### 3.1 构建组织特异性蛋白质网络

本文从 BIOGPS 项目中的 Affymetrix 数据集中获得了 83 个人体组织和细胞系的转录水平<sup>[20]</sup>, 并从 BioGrid 网站<sup>[21]</sup> 中下载到人体蛋白质 - 蛋白质相互作用关系, 构建了 83 个组织特异性蛋白网络, 具体处理数据以及构造方法详见文献<sup>[20]</sup>, 本文挑选了蛋白质对个数  $>10000$  的 26 个组织进行分析, 这 26 个组织或者细胞分别为: BDCA 4+ 树突状细胞、支气管上皮细胞、CD105+ 内皮、CD19+B 细胞、髓细胞、造血干细胞、CD4+ T 细胞、CD56+ 自然杀伤细胞、CD71+ 早期红细胞前体细胞、CD8+T 细胞、心脏肌细胞、肠和直肠腺癌、慢性粒细胞性白血病 k - 562、早幼粒细胞性白血病淋巴细胞 (MOLT-4)、白血病 HL-60、淋巴瘤 burkitt(Daudi)、淋巴瘤 burkitt(Raji)、日间松果体、夜间松果体、前额叶皮层、视网膜、前列腺、平滑肌、甲状腺、全血.

#### 3.2 黄金标准蛋白质复合体

为了衡量所检测出的复合体的精确性, 本文选择了一个广泛使用的复合体标准作为黄金标准, 该标准是从哺乳动物蛋白质复合体的 CORUM<sup>[22]</sup> 数据库中得到, 最终获得由 2151 个蛋白质组成的 324 个复合体, 本文中只选取其中蛋白质个数大于 3 个的复合体.

#### 3.3 评价标准

我们将判断预测的复合体是否能很好地对应到已知的复合体作为评判标准. ACC<sup>[23]</sup> 是用来测量几何精度的, 在这项研究中, 它被用来评估预测的复合体与参考的复合体之间的相似性. MMR (the Maximum Matching Ratio) 由 Paccanaro 提出的用来评估相对于参考蛋白质复合体来说预测的蛋白质复合体是否符合期望的一个评价标准.

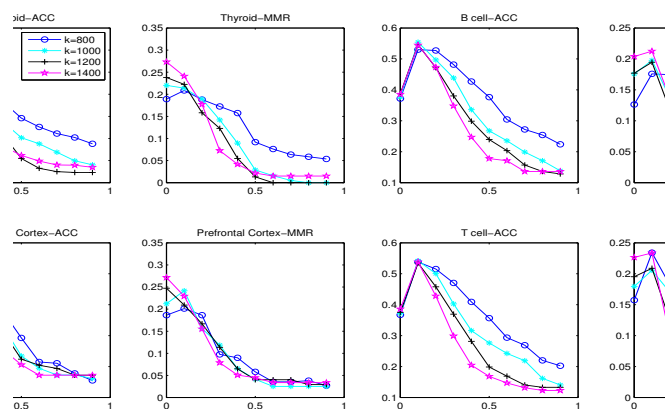


图 2: 参数  $\tau$  在不同组织识别出不同复合体的曲线图

### 3.4 基本聚类方法的参数选择

MCL 有一个用来调整聚类的间隔尺寸的参数, 俗称膨胀率, 本文设定其取值范围从 3.0 到 5.0, 步长为 0.2; MCODE 设定蛋白质个数为 3, 其余参数默认; MINE 设定蛋白质个数为 3, 其余参数默认; DPCLUS 有两个参数, 最小密度  $d$  以及最小聚类性质参数  $cp$ , 本文设定其值分别为 0.7 以及 0.5; ClusterOne 参数设为默认; CoAch 有一个参数  $\omega$ , 用来过滤冗余的核心蛋白质, 本文设定取值范围为 0.225 到 0.925, 步长为 0.05; SPICi 有两个参数, 其中我们设定密度阈值这一参数的取值范围为 0.1 到 1, 步长为 0.1.

对于以上 7 种算法, 挑选出使得每种算法的 ACC 和 MMR 的调和平均数最大的参数值作为最后选定的参数值.

### 3.5 参数选择

本文的算法中, 共有两个参数  $K$  以及  $\tau$ ,  $K$  为所识别的蛋白质复合体的个数, 根据过往者的经验, 设定其取值范围从 600 到 1600, 步长为 200,  $\tau$  为过滤阈值, 设置其取值范围为 0 到 0.9, 步长为 0.1. 在对 26 个组织分别进行上述算法后, 得到表 1.

表 1: ACC - 结果比较

|                    | coach       | Dpclus      | mcl  | mcode | mine        | cluster one | spici       | NMF         |
|--------------------|-------------|-------------|------|-------|-------------|-------------|-------------|-------------|
| BDCA 4+++ 树突状细胞    | 0.34        | 0.44        | 0.42 | 0.27  | 0.42        | <b>0.50</b> | 0.46        | <b>0.51</b> |
| 支气管上皮细胞            | 0.67        | 0.64        | 0.54 | 0.49  | 0.61        | 0.67        | <b>0.68</b> | <b>0.68</b> |
| CD105+++ 内皮        | 0.35        | 0.39        | 0.38 | 0.28  | 0.44        | <b>0.45</b> | <b>0.45</b> | <b>0.51</b> |
| 单核细胞               | <b>0.60</b> | 0.45        | 0.43 | 0.32  | 0.49        | 0.52        | 0.50        | <b>0.60</b> |
| CD19+++B 细胞        | 0.39        | <b>0.50</b> | 0.44 | 0.29  | 0.46        | <b>0.50</b> | 0.43        | <b>0.55</b> |
| 髓细胞                | 0.38        | 0.46        | 0.43 | 0.29  | 0.43        | 0.52        | 0.49        | <b>0.52</b> |
| 造血干细胞              | 0.31        | 0.39        | 0.39 | 0.25  | 0.37        | <b>0.47</b> | 0.38        | <b>0.47</b> |
| CD4+++ T 细胞        | 0.41        | 0.51        | 0.46 | 0.33  | 0.53        | <b>0.56</b> | 0.52        | <b>0.58</b> |
| CD56+ 自然杀伤细胞       | 0.35        | 0.47        | 0.44 | 0.29  | 0.44        | <b>0.51</b> | 0.44        | <b>0.52</b> |
| CD71+++ 早期红细胞前体细胞  | 0.46        | 0.56        | 0.48 | 0.43  | 0.51        | <b>0.61</b> | 0.59        | <b>0.62</b> |
| CD8+++T 细胞         | 0.37        | 0.47        | 0.47 | 0.30  | 0.46        | <b>0.52</b> | 0.50        | <b>0.54</b> |
| 心脏肌细胞              | <b>0.70</b> | 0.64        | 0.54 | 0.52  | 0.64        | 0.68        | 0.68        | <b>0.71</b> |
| 肠和直肠腺癌             | 0.62        | 0.62        | 0.56 | 0.41  | 0.61        | 0.63        | 0.62        | <b>0.63</b> |
| 慢性粒细胞性白血病 k - 562  | 0.69        | 0.69        | 0.56 | 0.51  | 0.66        | <b>0.71</b> | 0.65        | <b>0.71</b> |
| 早幼粒细胞性白血病淋巴细胞      | <b>0.59</b> | 0.52        | 0.53 | 0.42  | 0.52        | 0.55        | 0.53        | <b>0.59</b> |
| 白血病 HL-60          | 0.63        | 0.60        | 0.59 | 0.47  | 0.56        | <b>0.64</b> | 0.62        | <b>0.64</b> |
| 淋巴瘤 burkitt(Daudi) | 0.34        | 0.45        | 0.42 | 0.28  | 0.43        | <b>0.51</b> | 0.45        | <b>0.50</b> |
| 淋巴瘤 burkitt(Raji)  | 0.63        | 0.57        | 0.52 | 0.44  | 0.59        | <b>0.65</b> | 0.57        | <b>0.63</b> |
| 日间松果体              | 0.41        | 0.52        | 0.44 | 0.33  | 0.47        | 0.55        | <b>0.56</b> | <b>0.58</b> |
| 夜间松果体              | 0.39        | 0.51        | 0.46 | 0.35  | 0.49        | <b>0.55</b> | 0.52        | <b>0.58</b> |
| 前额叶皮层              | 0.52        | 0.56        | 0.52 | 0.32  | <b>0.62</b> | 0.56        | 0.54        | <b>0.64</b> |
| 视网膜                | <b>0.71</b> | 0.67        | 0.53 | 0.55  | 0.67        | 0.66        | 0.62        | <b>0.73</b> |
| 前列腺                | 0.47        | 0.50        | 0.49 | 0.27  | 0.51        | 0.48        | <b>0.52</b> | <b>0.60</b> |
| 平滑肌                | <b>0.73</b> | 0.66        | 0.59 | 0.49  | 0.64        | 0.68        | 0.70        | <b>0.74</b> |
| 甲状腺                | <b>0.64</b> | 0.63        | 0.59 | 0.43  | 0.59        | 0.63        | 0.62        | <b>0.65</b> |
| 全血                 | <b>0.65</b> | 0.64        | 0.62 | 0.51  | 0.63        | 0.63        | <b>0.65</b> | <b>0.67</b> |

在对所有组织计算中发现, 一般复合体个数在 600–2000 并且阈值在 0 或者 0.1 的情况

下表现良好, 由于篇幅有限, 仅挑选出 4 个组织进行参数分析, 分别为: 甲状腺、B 细胞、前额叶皮层、T 细胞, 如图 2.

### 3.6 与其他聚类方法进行比较分析

在这一章中, 我们将本文的算法与其他 7 种算法对这 26 个组织或者细胞的蛋白质网络进行聚类之后的结果进行比较. 对于其他 7 种基本聚类方法, 我们取其 ACC 和 MMR 的调和平均数为这 26 个组织最后的结果, 从表 1 中可以看出, 本文的算法最后得到的 ACC 值在 24 个组织中处于最大值, 两个组织中居于第二.

本文将 26 个组织所用的 7 种方法得到的最高值与本文所用的方法进行比较, 提升最高的组织是前列腺, 提高值为 13%. 在与其他 7 种方法分别单独比较时, 提高最高的百分比分别为: 51.61%、33.33%、39.53%、122.22%、27.03%、25.00%、27.91%, 具体提升情况可参见图 3, 从图中我们可以看出, MCODE 算法所得到的结果最差, 在 26 个组织中, 使用非负矩阵分解得到的结果均比其提高 30% 以上; 其次是 MCL, 提高了 8% 到 40%; 而 CClusterONE 表现最好, 有两个组织比本文的算法分别高出 1.96%、3.08%.

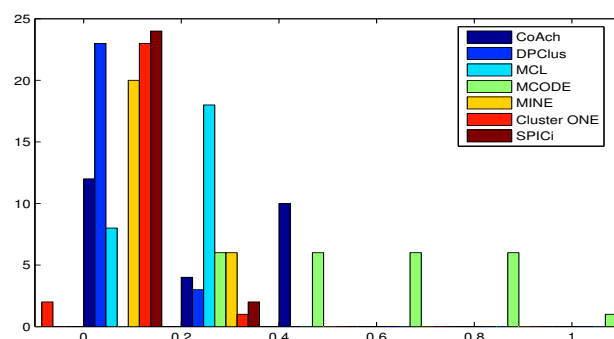


图 3: 与其他组织比较, NMF 提升百分比的频数图

从上述结果中可以看出, 本文所提出的算法与其他 7 种方法相比是具有优越性的.

## 4 讨论与分析

组织特异性蛋白质复合物对于理解生物学功能以及确定生物标志物和功能靶标十分重要, 这也是本文的研究动机. 同一个蛋白质在不同的组织中会与不同的蛋白质相结合, 举例来说, 转运蛋白 1(TNPO1) 在树突状细胞中与蛋白质 CD4、PPP3CA、TNPO3 结合, 在髓细胞中与 SRP19、TNPO3 相结合, 而在平滑肌中则与蛋白质 IPO5、IPO7、NUTF2、RAN、SRP19 结合形成复合物, 由此可以看出在不同的组织中其会与不同的蛋白质相结合, 而 TNPO1 与 TNPO3 则同时出现在不同组织的同一个复合物中, 这与生命活动也是相符合的.

在真正的生命活动中, 蛋白质会在不同的组织中与不同的蛋白质相结合, 而许多现有的检测蛋白质复合物模型都是在静态 PPI 网络模型中直接检测, 而忽略了蛋白质复合物的空间特异性. 本文利用多种方法对不同的组织构建组织特异性蛋白质相互作用网络, 并使用非负矩阵分解模型对其他聚类结果进行合并聚类, 并在获取组织特异性蛋白质复合物时得到了良好的结果. 同时, 本文也有一些不足, 虽然本文的结果在 ACC 标准中表现良好, 但在 MMR 这一标准中仍需改进, 同时, 本文仅选取一个黄金标准复合物, 在接下来的工作中, 我们可

参考多组黄金标准复合体进行方法之间的比较.

## 参 考 文 献

- [1] Aebersold R, Mann M. Mass spectrometry-based proteomics[J]. *Nature*, 2003, 422(6928): 198–207.
- [2] Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry[J]. *Nature*, 2002, 415(6868): 180–183.
- [3] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome[J]. *Proceed. National Acad. Sci. United States America*, 2001, 98(8): 4569–4574.
- [4] Uetz P, Giot L, Cagney G, Mansfield T A, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*[J]. *Nature*, 2000, 403(6770): 623–627.
- [5] Gavin A C, Bische M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes[J]. *Nature*, 2002, 415(6868): 141–147.
- [6] Lage K, Karlberg E O, Størting Z M, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders[J]. *Nature Biotechnology*, 2007, 25(3): 309–316.
- [7] Lo K, Raftery A E, Dombek K M, et al. Integrating external biological knowledge in the construction of regulatory networks from time-series expression data[J]. *BMC Sys. Bio.*, 2012, 6(2): 101.
- [8] Vasmatzis G, Klee E W, Kube D M, Therneau T M, Kosari F. Quantitating tissue specificity of human genes to facilitate biomarker discovery[J]. *Bioinformatics*, 2007, 23(11): 1348–1355.
- [9] Li D, Li J, Ouyang S, Wang J, Wu S, Wan P, Zhu Y, Xu X, He F. Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: large-scale organization and robustness[J]. *Proteomics*, 2006, 6(2): 456–461.
- [10] Enright A J, Dongen S V, Ouzounis C A. An efficient algorithm for largescale detection of protein families[J]. *Nucleic Acids Res*, 2012, 30(7): 1575–1584.
- [11] 冀俊忠, 刘志军, 刘红欣, 刘椿年. 蛋白质相互作用网络功能模块检测的研究综述 [J]. *自动化学报*, 2014, 40(4): 577–593.
- [12] Bader G D, Hogue C W V. An automated method for finding molecular complexes in large protein interaction networks[J]. *BMC Bioinformatics*, 2003, 4(1): 2.
- [13] Rhrissorrakrai K, Gunsalus K C. MINE: module identification in networks[J]. *BMC Bioinformatics*, 2011, 12(1): 192.
- [14] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks[J]. *BMC Bioinformatics*, 2006, 7(1): 207.
- [15] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks[J]. *Nature Methods*, 2012, 9(5): 471–472.
- [16] Wu M, Li X L, Kwok C K, Ng C K. A core-attachment based method to detect protein complexes in PPI networks[J]. *BMC Bioinformatics*, 2009, 10(1): 169.
- [17] Jiang P, Singh M. SPICi: a fast clustering algorithm for large biological networks[J]. *Bioinformatics*, 2010, 26(8): 1105–1111.
- [18] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. *Nature*, 1999, 401(6755): 788–791.
- [19] Ding C, He X F, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering[J]. *Siam Intern. Confer. Data Min.*, 2005, 5: 606–610.

- [20] Lopes T J, Schaefer M, Shoemaker J, Matsuoka Y, Fontaine J F, Neumann G, Andrade-Navarro M A, Kawaoka Y, Kitano H. Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases[J]. *Bioinformatics*, 2011, 27(17): 2414–2421.
- [21] Chatr-aryamontri A, Breitkreutz B J, Heinicke S, et al. The Biogrid interaction database: 2013 update[J]. *Nucleic Acids Research*, 2013, 41(2): 816–823.
- [22] Havugimana P C, Hart G T, Nepusz T, et al. A census of human soluble protein complexes[J]. *Cell*, 2012, 150(5): 1068–1081.
- [23] Li X, Wu M, Kwoh C K, et al. Computational approaches for detecting protein complexes from protein interaction networks: a survey[J]. *BMC Genomics*, 2010, 11(4): S3.
- [24] Ou-Yang L, Dai D Q, Zhang X F. Protein complex detection via weighted ensemble clustering based on bayesian nonnegative matrix factorization[J]. *Plos One*, 2013, 8(5): 639–642.
- [25] Ou-Yang L, Dai D Q, Li X L, Wu M, Zhang X F, Yang P. Detecting temporal protein complexes from dynamic protein-protein interaction networks[J]. *BMC Bioinformatics*, 2014, 15(1): 16001–16005.
- [26] Zhang X F, Dai D Q, Ou-Yang L, Yan H. Detecting overlapping protein complexes based on a generative model with functional and topological properties[J]. *BMC Bioinformatics*, 2014, 15(2): 836–842.
- [27] Zhang W, Zou X F. A new method for detecting protein complexes based on the three node cliques[J]. *IEEE/ACM Trans Comput. Biol. Bioinform*, 2015, 12(4): 879–886.
- [28] 涂俐兰. 两两序列比对的一种新方法 [J]. *数学杂志*, 2006, 26(1): 67–70.

## IDENTIFICATION OF THE TISSUE SPECIFIC PROTEIN COMPLEXES

DING Xia<sup>1</sup>, ZHANG Xiao-fei<sup>2</sup>, YI Ming<sup>3</sup>

(1.School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)

(2.School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China)

(3.School of Science, Huazhong Agricultural University, Wuhan 430070, China)

**Abstract:** In this paper, we study the identification problem of tissue-specific protein complexes. By using a variety of typical clustering algorithm to cluster the network, we construct a tissue-specific protein-protein interaction network based on the protein-protein interaction networks as well as the tissue-specific gene expression data, then merge the results with non-negative matrix factorization model to obtain tissue-specific protein complexes. The results show that clustering effect has been significantly improved, and can identify tissue-specific protein complexes.

**Keywords:** protein-protein interaction networks; complexes identification; tissue-specific; non-negative matrix factorization

**2010 MR Subject Classification:** 92B05