PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(8): 914~920
www.pibb.ac.cn

一种从多表达谱数据挖掘基因共表达团的新方法 *

陈 兰[1,3)**] 王世敏[1,3)**] 陈润生[1,2)***]

([1)]中国科学院计算技术研究所，北京 100190；[2)]中国科学院生物物理研究所，北京 100101；[3)]中国科学院研究生院，北京 100049)

摘要 随着近年来高通量基因表达谱数据的涌现，集成多个不同实验条件的表达谱数据，并挖掘在多数据源都保守的基因共表达团，成为预测基因功能或者调控关系的方法之一．但是，常用的方法通常仅简单地集成不同表达谱数据并推导保守基因共表达团，这样可能会导致结果中出现并非真正在多数据源保守的共表达团．提出一种结合最小哈希与局部敏感哈希的新方法，可以高效地寻找在多表达谱数据源中真正保守的基因共表达团．结果分析证明，相比过去的方法，现提出的方法可以获得更加功能相关和调控相关的基因共表达团．

关键词 表达谱，共表达网络，最小哈希，局部敏感哈希
学科分类号 180.1410

随着测量mRNA表达水平的微阵列(microarray)技术的发展，近年来涌现了大量高通量基因表达谱数据．这些数据使得我们可以获得大量基因在同一个实验条件或者时间点的相对表达量，因此成为生物信息领域用于挖掘未知基因功能，或者推测基因间的转录调控关系的重要数据源．基于共享一个生物通路，或者属于相同蛋白质复合物的基因倾向于共调控的假设，许多研究都致力于从大规模的基因表达谱数据中推导共表达的基因团，从而预测基因的调控模式，或者未知基因的生物功能等．从表达谱数据中寻找基因共表达团的方法，可以看做是从该表达谱数据代表的基因共表达网络中寻找比较稠密的子图．基因的共表达网络被定义为，图中的一个点代表表达谱数据中的一个基因，如果两个基因之间共表达，则在图中对应的两点之间连接一条边．判断两个基因是否共表达的方法有很多种，最常用的方法是首先计算基因的表达相关性，再判断该表达相关性的值是否大于一个事先设定的阈值，或者用统计方法计算该表达相关性是否显著的高[1,2]．

然而，在一些研究中指出[3]，相似的基因表达模式并不一定意味着基因的功能相关，反之亦然．例如，即使在相同的实验条件下，也可能有多个代谢通路同时被激活．因此，在同样的实验条件下处于不同代谢通路的基因有可能彼此呈现相似的表达模式，但事实上它们的功能并不相关．此外，芯片实验中的噪音也可能会导致对基因表达相关性估计的误差．因此，简单地通过一个表达谱数据中基因的共表达模式来预测基因功能或者调控的关系，会引入不可预料的误差．考虑到有意义的生物模块倾向于在多个独立的实验条件下被激活，并且，由于数据噪音或者其他偶然因素导致的基因之间的高表达相关，不太可能在多个实验条件下都保守．因此，集成多个不同实验条件的表达谱数据，并挖掘在多数据源都保守的基因共表达团，可以降低对基因调控关系或者功能预测的误差．

集成多表达谱推导保守基因共表达团的方法中，最常见的是首先集成不同表达谱得到一个“综合共表达网络”，然后从中挖掘基因共表达团[1,4]．在这个“综合共表达网络”的图中，每一个点代表一个基因．如果两个基因在多个表达谱数据中都有较高的表达相关性，也就是说这两个基因在多数据源中保守共表达，则在图中对应的两点之间连接一条边．基于该网络寻找的基因共表达团，比从单数

* 国家自然科学基金资助项目(30630040, 30570393, 30600729).
** 共同第一作者.
*** 通讯联系人.
Tel: 010-64888543 , E-mail: crs@sun5.ibp.ac.cn
收稿日期：2008-01-07，接受日期：2008-02-02

据源出发寻找的基因共表达团更加可靠．但是，在一些研究中已经指出[2, 5]，这类方法有一个主要的缺点，那就是从“综合共表达网络”得到的基因共表达团，并不一定真正是在多个共表达网络中都保守的共表达团．因为虽然该网络上的每一条边都代表着对应的两个基因在多数据源中高表达相关，但是不同边代表的基因对可能在不同的数据源高表达相关．因此从“综合共表达网络”得到的共表达团，无法保证该团所有的边都在相同或相似的数据源出现，也就是说，不能确保这个团代表的多个基因之间的共表达关系，是在多数据源中共同出现的(图 1).

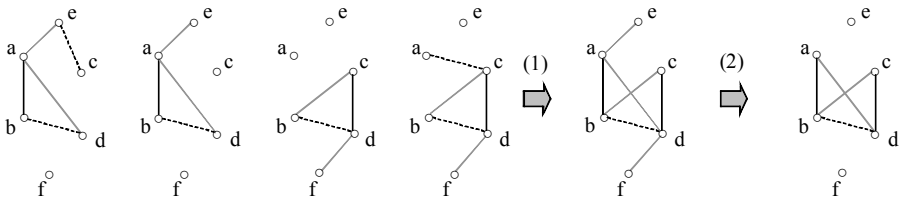


Fig. 1　An example of the“summary co-expression network”method

As shown in the figure, four co-expression networks are derived from four microarrays．It is assumed that an edge is a“conserved co-expressed edge”if it occurs in at least two co-expression networks．In the step (1), a “summary co-expression network”, which involves all the conserved co-expressed edges，is constructed by aggregating all of the four co-expression networks．In the step (2), a dense subgraph of the“summary co-expression network”is obtained by graph clustering method．It is found that the co-expressed cluster represented by the dense subgraph does not occur in any of the original co-expression network.

相对更合理的方法，是要求共表达团的边在相似的共表达网络中出现，也就是寻找在多个表达谱数据对应的共表达网络中频繁出现的子图．由于数据规模的原因，直接在多个共表达网络中搜索多次出现的子图的计算时间是不可接受的．因此，本文提出了一种可扩展的高效方法：首先基于随机算法——最小哈希(min-hashing)与局部敏感哈希(locality-sensitive hashing)结合的方法来降低搜索空间，然后应用普通的图聚类算法寻找最终的结果．在结果分析中发现，应用该方法寻找到的基因共表达团，比“综合共表达网络”方法得到的结果具有更加显著的生物意义.

1　材料与方法

1.1　数据集

我们集成了来源于 Stanford Microarray Database (http://genome-www5.stanford.edu/)以及 NCBI Gene Expression Omnibus(http://www.ncbi.nlm.nih.gov/geo/)的有关酵母(*Saccharomyces cerevisiae*)的 10 套表达谱数据，并根据试验条件将这些数据划分为 20 个表达谱文件，每一个文件包含相似或相关的实验条件(详细数据见本文网络版附录，http://www.pibb.ac.cn/cn/ch/common/create_pdf.aspx?file_no=20080016&flag=1)．所有的表达谱数据经过了标准化，缺失数据超过 50%的基因被丢弃．剩余的所有基因被纳入计算．最终，每一个文件包含 5 397 个基因，以及 7～56 个表达谱数据点．所有的文件一共有 312 个数据点.

1.2　方法

1.2.1　寻找多数据源保守共表达团的方法.

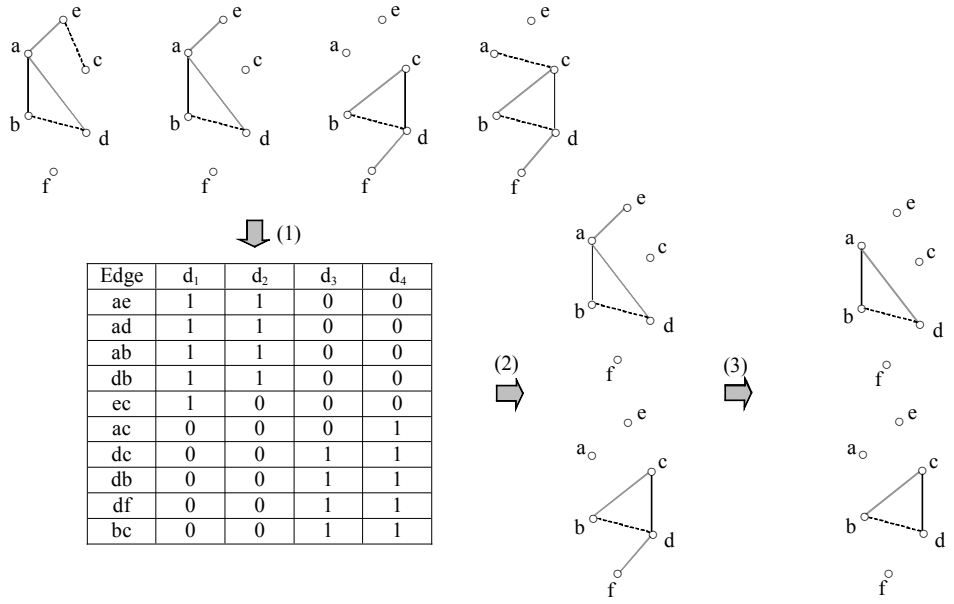
本文提出的方法，目的在于寻找多个表达谱数据对应的共表达网络中频繁出现的子图，也就是在多表达谱数据源中保守的基因共表达团．该方法可分为下列 3 个步骤(图 2)：(1) 每一个表达谱数据推导出相应的共表达网络；(2) 利用最小哈希和局部敏感哈希的方法，将在相似数据源中出现的共表达边聚类，形成候选的保守基因共表达团；(3) 删除候选保守基因共表达团的假阳性边，并采用图聚类方法进行聚类.

下面详细介绍这 3 个步骤.

(1)　共表达网络的生成

假设共有 m 个不同的表达谱数据，为了得到相应的共表达网络，需要判断每一个表达谱数据中任意两个基因之间是否显著高表达相关．具体方法如下：

首先采用当前广泛使用的皮尔逊(Pearson)相关系数来计算基因之间的表达相关性．为了判断该表达相关值是否显著的高，我们将该值转换为下列



Edge	d_1	d_2	d_3	d_4
ae	1	1	0	0
ad	1	1	0	0
ab	1	1	0	0
db	1	1	0	0
ec	1	0	0	0
ac	0	0	0	1
dc	0	0	1	1
db	0	0	1	1
df	0	0	1	1
bc	0	0	1	1

Fig. 2　An example of the "conserved subgraph of multiple networks" method

As shown in the figure, four co-expression networks are derived from four microarrays. In the step (1), a table is formed to record the occurrence of each co-expressed edge (d_1, d_2, d_3, d_4 indicating the four co-expression networks). If one edge occurs in any of the co-expression networks, the corresponding item is "1", otherwise is "0". In the step (2), the edges which occur in the similar co-expression networks are clustered by min-hashing and locality-sensitive hashing methods, resulting two candidate clusters. In the step (3), graph clustering method is applied to the candidate clusters to get two final co-expressed clusters, which really are conserved across the co-expression networks.

形式

$$r_1=\sqrt{(n-2)r^2/1-r^2}$$

上述公式中，r 表示两个基因之间的皮尔逊表达相关性，n 表示计算这两个基因的表达相关性所使用的数据点数．转换后的 r_1 值服从自由度为 $n-2$ 的 t 分布．如果两个基因之间的 r_1 值大于预设的 p-value 对应的 t 分布值，则认为这两个基因显著表达相关，在对应该表达谱数据的共表达网络中，连接这两个基因之间的边被赋予权重"1"，表示这条边出现，否则为"0"．

我们建立了一张"共表达表"，来存储共表达网络的边在不同数据源中出现的情况．"共表达表"的每一行代表一条边，每一列代表一个表达谱数据．表中第 i 行第 j 列的元素，$c_{ij}(1\leqslant i\leqslant m, 1\leqslant j\leqslant n^2)$，表示第 i 条边在第 j 个表达谱数据中的权重取值，"1"表示该边出现，也就是该边对应的基因对高表达相关，否则为"0"．通过上述方式，我们将 m 个共表达网所有边的信息用一张二进制表存储．实际情况中，可以仅记录保守共表达边的信息.

(2)　共表达边的聚类

如前所说，直接寻找在多个共表达网络中频繁出现的子图是不可行的．因此，在这一步，首先将相似数据集中出现(权重为"1")的边聚类以降低搜索空间．在每一个结果类中，边都被确保在相似的数据集中出现，因此这些边构成的子图是同时出现在相似数据集的子图．从这些子图中再应用普通的图聚类方法，得到的结果就是我们要寻找的保守基因共表达团.

如果将"共表达表"的每一行看做一个集合，集合的元素是该行对应的边在不同数据集下的权重．我们仅关心两条边是否出现在相似的数据集中，也就是仅关心两个集合之间"1"元素的相似度．因此采用 Jaccard 相似度定义两条边的权重相似性．设 C_i 表示边 i 的权重集合，C_j 表示边 j 的权重集合，则 C_i 和 C_j 的 Jaccard 相似性定义为：

$$S(C,\ C_j)=\frac{|C_i\cap C_j|}{|C_i\cup C_j|}$$

如果直接聚类的话，需要比较"共表达表"里任意两行的相似性，一共需要比较的次数是 $O(n^4m)$，其中，n 是基因的个数，m 是表达谱数据

的个数．由于基因的个数 n 往往在 10^3～10^4 的范围，这样的计算复杂性是无法承受的．因此，我们采用了结合最小哈希和局部敏感哈希的高效近似聚类方法．该方法在大规模数据集的规则挖掘中曾被采用[6]．下面简单介绍该方法的思想．

采用最小哈希方法的目的是将“共表达表”内任意两行之间的 Jaccard 相似性估计值转换成海明距离估计值．首先，随机交换原共表达表的各列，这样得到一张随机表．然后从左到右搜索每一行第一个等于“1”的元素，记录该元素的列位置作为该行数据在这次随机扰动过程中得到的最小哈希值 $h(C_i)$．对原共表达表的任意两行向量，它们在一次随机扰动中分别得到的最小哈希值 $h(C_i)$和 $h(C_j)$与这两行向量之间的 Jaccard 相似性 $S(C_i, C_j)$有下列关系[6].

$$\Pr[h(C_i)=h(C_j)]=\frac{|C_i\cap C_j|}{|C_i\cup C_j|}=S(C_i,\ C_j)$$

重复上述随机扰动过程 k 次，对应原共表达表的每一行可产生 k 个独立的最小哈希值 $h_1(C_i)\cdots h_k(C_j)$，由此可产生一个最小哈希表，每一行包含的是对应原共表达表相应行的 k 个最小哈希值．该方法可以保证最小哈希表中任两行向量之间的海明距离，是原共表达表对应两行向量之间 Jaccard 相似性的一个良好的估计[6].

局部敏感哈希是一个高效的相似搜索算法，广泛地被应用在各种领域中，包括基因组大规模的序列比对[7]．该方法的思想是，如果两个数据向量足够的相似，那么，随机抽取这两个数据向量中的一部分，这一部分数据会以很高的概率相同．基于这样的思路，我们的实现如下：

从上一步得到的最小哈希表中，任意抽取 r 列，针对每一行 C_i，把这 r 列数据的值连接起来作为该行的一个哈希关键字．独立重复上述过程 l 次，每一次，拥有相同哈希关键字的行被聚类到一起，作为候选集合．每一个候选集合内的行，由于局部值相同，因此有可能彼此的相似度大于我们感兴趣的阈值.

(3)　保守基因共表达团的生成

由于局部敏感哈希方法的思想是用一定的结果不精确性换取时间的节省，因此需要检查每一个候选集合，去除假阳性数据，也就是与其他行的相似度没有达到规定阈值的行向量．针对每一个候选集合，我们迭代的删除与整个集合相似性最低的行，直到剩余行的相似度大于阈值为止(详细方法见本文网络版附录)．经过该步骤以后，每一个候选集合的行向量之间的相似性都满足了要求．由于每一个行向量代表的是连接两个基因之间的共表达边，因此每一个候选集合可以看做一个图，该图由集合内所有行向量代表的边构成．通常认为寻找稠密的基因共表达团能消除更多的噪音．因此，可以采用普通的图聚类方法在每一个候选集合对应的图中寻找稠密子图．这些稠密子图就是在多数据源中保守的基因共表达团．本文采用了 Frey B 和 Dueck D 等提出的“affinity propagation”方法来寻找每一个候选集合对应图的稠密子图[6]．本文采用的子图稠密度(density)定义为: $density=2m/(n(n-1))$，m 是子图边的数目，n 是子图节点的个数.

1.2.2　“综合共表达网络”方法的实现．为了将本文挖掘多数据源保守基因共表达团的方法与传统的“综合共表达网络”方法进行比较，针对相同的数据集，本文也计算了“综合共表达网络”方法得到的结果．计算方法如下：对 m 个表达谱数据采用上述相同的方法生成对应的 m 个共表达网络．连接两个基因的共表达边，如果在多个共表达网络(≥3)中都被赋值为“1”，则认为是保守的共表达边．所有保守的共表达边可以构建成一个“综合共表达网络”．该网络集成了所有共表达网络的综合信息．接着，对“综合共表达网络”直接进行图聚类，挖掘稠密子图作为最终的保守基因共表达团．依然采用“affinity propagation”方法来进行图聚类[6]．以下的讨论中，在不引起混淆的情况下，称本文挖掘多数据源基因共表达团的方法为“多网络频繁团”方法，“综合共表达网络”方法简称为“综合网络”方法．而“保守基因共表达团”也简称为“基因共表达团”.

2　计算结果及讨论

2.1　参数的选择及实现

用表达谱数据文件构建相应的共表达网络时，我们设定判断两个基因是否显著高表达相关的 p-value 为 0.000 1，因为用该参数得到的共表达网络规模适中．在≥3 个数据源里显著高表达相关的边称为保守共表达边．20 个表达谱数据文件共得到 357 546 条保守共表达边，约占整个共表达网络所有可能边数的 2.5%．由于“多网络频繁团”和“综合网络”方法在构建共表达网络这一步是完全相同的，所以两个方法得到相同的保守共表达边.

在“多网络频繁团”方法中使用了最小哈希结

合局部敏感哈希的方法将相似数据集中出现的保守共表达边聚类．由于该方法的随机性，需要设定参数来控制其假阳性和假阴性数据的产生(详细参数设定见本文网络版附录)．对产生的每一个候选集合使用“affinity propagation”方法进行图聚类(该方法的参数设定见本文网络版附录)，寻找稠密子图．本文设定稠密度≥0.25，并且拥有≥3 条边的子图为最终寻找的结果，我们称为一个类，也是一个基因共表达团．由于最小哈希与局部敏感哈希方法的随机性，为了增加结果分析的稳定性，我们共计算了 3 次结果，下列分析都是取这 3 次结果的平均值呈现．“多网络频繁团”方法平均产生了约 4 800 个类，包含约 19 000 条边．

而“综合网络”方法对上述保守共表达边建立的网络直接调用“affinity propagation”方法进行聚类，寻找频繁子图，对频繁子图的要求同上．该方法产生类的数目远远小于“多网络频繁团”方法，仅有 259 个类，包含 14 417 条边．

2.2 两种方法的结果类比较

从上述结果可以看出，“多网络频繁团”方法产生的类，数量远大于由“综合网络”方法产生的类，主要有下列两个原因：

a．“多网络频繁团”方法对结果类的要求比“综合网络”方法更加严格．在“综合网络”方法中，只要是保守共表达边(无论其在哪些数据集下保守)构成一个稠密子图，就被认为是一个基因共表达团．而“多网络频繁团”方法进一步要求每一个基因共表达团的边出现在相似的数据集中，也就是要求这个共表达团本身是多次出现的．因此在“综合网络”中的一个类，可能在“多网络频繁团”方法中被划分为出现在不同数据子集的小类，导致类数目的增加．

b．“多网络频繁团”方法使用最小哈希和局部敏感哈希方法的过程中，由于并没有严格限制一个类内所有的边必须出现在完全相同的数据集里(这样过分严格的限制往往会导致遗漏许多有意义的基因关系)，因此，同一条边可能依据其出现的数据集聚类到不同的候选集合，从而该方法允许同一条边在多个结果类中出现．从这个角度来说，本文提出的方法是一种灵活的允许类之间有交集的聚类方法．因此，也造成了结果类数目的增加．

2.3 两种方法的生物意义比较

首先分析由“综合网络”方法得到的类的生物意义．过去的研究发现，基因的保守共表达可能会暗示着基因之间的功能相关[1, 3]．因此，为了验证结果类的生物功能意义，我们从 SGD 数据库[8]下载了已知酵母基因的 Gene Ontology (GO) biological process 注释．本文采用超几何分布的统计方法(详细方法见本文网络版附录)计算每一个类的基因共享某一个 GO 注释的概率．由于一个类的基因可能共享多个 GO 注释，因此存在统计上的多检测问题(multiple hypothesis testing)，我们应用了 False Discovery Rate 方法[9]对计算得到的概率值做相应的调整．如果一个类显著地共享一个或多个 GO 注释(p-value < 0.01)，我们认为该类的基因是功能相关的．计算中，仅考虑拥有至少 4 个已知功能注释的基因的类．满足该条件的由“综合网络”方法得到的结果类有 157 个，其中 48 (30%)个类显著功能相关．而用“多网络频繁团”方法得到的满足上述条件的约 2 800 个类中，平均 71%的类显著功能相关．这些类显著共享的功能分布在生物的各个过程中，例如蛋白质翻译，RNA 代谢，碳水化合物代谢，端粒维护与压力响应等．

基于被相同转录因子调控的基因倾向于共表达的假设[10, 11]，本文进一步从转录调控的角度探讨上述方法得到的结果类的生物意义．我们从 YEASTRACT 数据库[12]下载了酵母所有已知实验支持的转录调控因子与目标基因的调控关系数据．如果一个类的基因显著地被同一个转录因子调控，我们认为这个类是一个可能的转录调控模块．同样，至少 4 个基因有调控关系数据的类才被计算，方法同上．“多网络频繁团”方法得到的 3 800 个结果类中平均 59%是可能转录调控模块，该比例远高于“综合网络”方法得到的结果，仅 27%的类显著被相同转录因子调控．

从上述结果可以看出，“多网络频繁团”方法得到的保守基因共表达团，不仅大部分功能相关，也有紧密的调控关联．比起“综合网络”方法得到的结果类，该方法得到的具有功能相关或者调控关联的结果类的百分比有明显的提高，也就是说“多网络频繁团”方法倾向于得到更加功能相关和调控相关的基因共表达团．这说明该方法对基因共表达团的边在相似数据集中同时出现的要求是合理而有效的．从结果类中可以发现，一些“综合网络”方法得到的没有明显生物意义的较大的类，在“多网络频繁团”方法中依据边出现的数据子集被划分为多个较小的类，并显著的功能相关或者调控相关．这说明了寻找真正在不同数据子集中保守的基因共

表达团可以过滤数据中更多的噪音，提高有意义的信号的强度，从而使得结果基因团更加具有生物意义.

由于“多网络频繁团”方法得到的大部分结果类都有显著的生物功能或者调控关联，因此本方法也可以应用于预测未知基因功能或者基因间的调控关系. 并且，本文提出的寻找多数据集保守基因共表达团的方法，采用了最小哈希和局部敏感哈希技术降低搜索空间，使得总体方法的复杂度即使在大规模数据集情况下也是可以接受的. 从而该方法可以应用到不同物种日益增长的表达谱数据集中，挖掘出更合理的基因共表达团，进一步可靠地推测未知基因的功能和基因的调控信息.

参 考 文 献

1 Lee H K, Hsu A K, Sajdak J, *et al*. Coexpression analysis of human genes across many microarray data sets. Genome Res, 2004, **14**(6): 1085～1094

2 Hu H, Yan X, Huang Y, *et al*. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics, 2005, **21**(Suppl 1): i213～221

3 Stuart J M, Segal E, Koller D, *et al*. A gene-coexpression network for global discovery of conserved genetic modules. Science, 2003, **302**(5643): 249～255

4 Aggarwal A, Li Guo D, Hoshida Y, *et al*. Topological and functional discovery in a gene coexpression meta-network of gastric cancer. Cancer Res, 2006, **66**(1): 232～241

5 Yan X, Mehan M R, Huang Y, *et al*. A graph-based approach to systematically reconstruct human transcriptional regulatory modules. Bioinformatics, 2007, **23**(13): i577～586

6 Cohen E, Datar M, Fujiwara S, *et al*. Finding interesting associations without support pruning. IEEE Transactions on Knowledge and Data Engineering, 2001, **13**(1): 64～78

7 Buhler J. Efficient large-scale sequence comparison by locality-sensitive hashing. Bioinformatics, 2001, **17**(5): 419～428

8 Cherry J, Adler C, Ball C, *et al*. SGD: Saccharomyces genome database. Nucl Acids Res, 1998, **26**(1): 73～79

9 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological), 1995, **57**(1): 289～300

10 Allocco D, Kohane I, Butte A. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics, 2004, **5**(1): 18

11 Yu H, Luscombe N M, Qian J, *et al*. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends in Genetics, 2003, **19**(8): 422～427

12 Teixeira M C, Monteiro P, Jain P, *et al*. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucl Acids Res, 2006, **34** (Suppl 1): D446～451

A Method to Detect Gene Co-expression Clusters From Multiple Microarrays*

CHEN Lan[1,3)**], WANG Shi-Min[1,3)**], CHEN Run-Sheng[1,2)***]

([1)] *Bioinformatics Research Group, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100190, China;

[2)] *Bioinformatics Laboratory, Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China;

[3)] *Graduate School of The Chinese Academy of Sciences, Beijing* 100049, China)

Abstract A number of recent studies have focused on discovering genetic functional or transcriptional modules by integrating information from the rapidly accumulating large-scale microarray expression datasets. Such studies commonly model each microarray as a co-expression network, and detect the conserved gene co-expression clusters from these co-expression networks. Currently, the commonly used method is mining conserved co-expression clusters directly from a “summary network”, which is obtained by aggregating all the co-expression networks derived from different microarrays. However, this method may generate false conserved clusters, which never occur in any of the original individual co-expression networks. Here a scalable and efficient method were proposed to detect the truly conserved gene co-expression clusters from multiple microarrays. This problem is formulated as mining frequently occurring subgraphs across multiple co-expression networks, and involves three steps: (1) Translating each microarray into co-expression network; (2) Clustering edges which occur in the similar co-expression networks by min-hashing and locality-sensitive hashing techniques to obtain the candidate clusters; (3) Applying graph clustering method to the candidate clusters to detect the conserved co-expressed clusters. This method was applied to yeast microarrays and the results demonstrate that, compared to the previous study, the

conserved co-expressed clusters detected by the method were more likely to be functionally homogeneous entities or potential transcriptional modules.

Key words microarray, co-expression network, min-hashing, locality-sensitive hashing

---

*This work was supported by grant from The National Natural Sciences Foundation of China(30630040, 30570393, 30600729).

**CHEN Lan and WANG Shi-Min contributed equally to this work.

***Corresponding author.

Tel: 86-10-64888543, E-mail: crs@sun5.ibp.ac.cn

Received: January 7, 2008 Accepted: February 2, 2008

附 录

1 数据集

本文采用的表达谱数据来源于 Stanford Microarray Database http://genome-www5.stanford.edu/) 以及 NCBI Gene Expression Omnibus(http://www.ncbi.nlm.nih.gov/geo/)的有关酵母(*Saccharomyces cerevisiae*)的 10 套表达谱数据．首先根据试验条件将这些数据划分为若干个表达谱文件，每一个文件包含相似或相关的实验条件．在以往的研究中指出，即使表达谱数据经过了“标准化”(normalization)，随机基因对的平均相关性仍不一定为 0[1]．如果一个表达谱的随机基因对的平均相关性远远偏离 0 的话，例如，远大于 0，将会导致更多的基因对拥有显著高的表达相关性．在构建相应的共表达网络时，更多的边将会建立，但其中有些边可能仅仅是因为数据噪音引起的．为了消除这些可能的噪音边，我们删除了平均表达相关性绝对值大于 0.05 的表达谱数据．总共剩下 20 个表达谱数据文件，每一个文件包含了相似或相关的实验条件．在下列的表中罗列了每个文件的实验条件(实验条件)，数据点的个数(数据点)，以及相应的参考文献(参考文献).

Table S1 The information of microarray files

Experimental conditions	Experimental data points	Resources
Alpha factor release	18	[2]
Elutriation	14	[2]
Gamma radiation	16	[3]
DNA damage(MMS) response	18	[3]
Mock irradiation	8	[3]
Nitrogen depletion	9	[4]
Sorbitol effects	7	[4]
Heat Shock	22	[4]
Steady State	8	[4]
Diamide	8	[4]
Nutrition limitation	11	[4]
Cell cycle alpha factor	13	[5]
Fkh1_2_alpha_factor	13	[5]
Nutrition	12	[6]
Sporulation	7	[7]
Diauxic_shift_timecourse	7	[8]
Signaling and circuitry of multiple MARK pathways	56	[9]
Glucose pulse on galactose chemostat	26	[10]
Calcineurin/Crzlp Signaling Pathway for Ca	25	[11]
Calcineurin/Crzlp Signaling Pathway For Na	16	[11]

2 去除候选集合的假阳性

局部敏感哈希方法过程中，哈希关键字碰撞的行聚集为候选的相似行集合．但是由于局部敏感哈希的思想是用一定的不精确性换取时间的节省，因此每一个候选集合里的行数据并不一定符合要求．在这一步中，我们一一检查每一个候选集合，去除假阳性数据，也就是与其他行的相似度没有达到事先规定阈值的行向量.

以前的局部敏感哈希方法通常用于一对数据的比对问题，因此最后只需要从候选集合里取出每一对数据，检测它们的相似性是否大于阈值，即可判断是否是最终结果．但在我们的应用中，要保证一个集合里所有行向量的相似度大于阈值．因此我们去除假阳性的方法有所不同.

首先，定义多个行向量的相似度为：

$$S_{\text{multiple}}(C_1,C_2,\cdots,C_i)=\frac{|C_1\cap C_2\cdots\cap C_i|}{|C_1\cup C_2\cdots\cup C_i|}$$

假设一个候选集合里有 n 个行向量，每一个行向量有 m 个元素．我们定义集合中的一个行向量 j 和该集合所有行向量的相似度为：

$S_j^*=\sum_{i=1}^{m}(c_{ij}*(\frac{\sum_{i=1}^{n}c_{ij}}{n}))$，$c_{ij}\in(0,1)(1\leqslant i\leqslant m,1\leqslant j\leqslant n)$；$c_{ij}$ 是第 j 个行向量第 i 列的元素值，取“1”或者“0”

也就是说，该集合所有行向量在每一列的元素“1”的比例($\sum_{j=1}^{n}c_{ij}/n$)作为权重，如果向量 j 在这一列的元素也为“1”，则与整个集合的相似度就增加 $c_{ij}*(\sum_{j=1}^{n}c_{ij}/n)$.

对每一个候选集合，判断集合的相似度 $S_{\text{multiple}}(C_1,C_2\cdots,C_i)$，如果已经大于事先设定的阈值 S^*_{multiple}，则该集合作为最终的结果输出．否则，迭代的删除与整个集合相似性 S^* 最低的行，直到剩余行的相似度大于阈值为止．这个过程的详细描述如下：

(1) 判断候选集合的相似度 S_{multiple} 是否大于阈值 S^*_{multiple}，如果大于，停止，并输出该集合．否则，跳到下一步.

(2) 计算候选集合中每一个行向量与集合的相似度：$S^*_1\cdots S^*_n$

(3) 寻找 $S^*_1\cdots S^*_n$ 中相似度最低的行向量，从集合中删除．跳到步骤(1).

3 参数的设定

3.1 最小哈希和局部敏感哈希方法参数的设定

最小哈希方法得到的最小哈希表，任意两行向量之间的海明距离 $S^m(C_i,C_j)$，与原共表达表中对应两行向量的 Jaccard 相似性 $S(C_i,C_j)$有下列关系[12]：

设 $0<\delta<1,>0,k\geqslant 2\delta^{-2}S_0^{-1}\lg\varepsilon^{-1}$. 对原表的任意两行向量 C_i,C_j

如果 $S(C_i,C_j)\geqslant S_0$，则以至少$(1-\varepsilon)$的概率 $S^m(C_i,C_j)\geqslant(1-\delta)S_0$

如果 $S(C_i,C_j)\leqslant S_0$，则以至少$(1-\varepsilon)$的概率 $S^m(C_i,C_j)\leqslant(1+\delta)S_0$

本文中，设定最小哈希方法的参数，S_0= 0.8， δ = 0.1，ε= 0.1，因此得到 k=575．也就是说对原共表达表每一行向量产生了 575 个最小哈希值．这样导致最终产生的最小哈希表比原共表达表占用更多的空间．但是 k 的增加可以带来很低的假阳性和假阴性．

在局部敏感哈希方法里，假设我们的目标是寻找彼此相似度大于 S_0 的行向量，假阳性数据是指出现在候选集合里，但彼此相似度≤S_0 的行向量；也就是说，彼此相似度≤S_0 的行向量的哈希关键字至少碰撞了一次，因而被罗列到候选集合里．假阳性率为[12]：

$\sum_{S_k \leqslant S_0} P(S_k)P_C(S_k) = \sum_{S_k \leqslant S_0} P(S_k)(1-(1-S_k^r)^l)$，$P(S_k)$代表着行向量之间的相似度为 S_k 的概率，是原数据行向量相似度的分布函数．

假阴性指的是行向量彼此的相似度≥S_0，却没有出现在候选集合里．也就是说，彼此相似度≥S_0 的行向量，但哈希关键字却一次也没碰撞．假阴性率为[13]：

$$\sum_{S_k \geqslant S_0} P(S_k)P_N(S_k) = \sum_{S_k \geqslant S_0} P(S_k)(1-S_k^r)^l$$

但由于当 $S_k \geqslant S_0$ 时，$(1-S_k^r)^l \leqslant (1-S_0^r)^l$，所以上述假阴性计算公式可化简为：

$$\sum_{S_k \geqslant S_0} P(S_k)P_N(S_k) \leqslant (1-S_0^r)^l \text{ [13]}$$

从上述假阳性和假阴性的计算公式可以看出，我们可以通过调整每次随机抽取的列数 r，以及重复的次数 l，来达到降低假阳性和假阴性的目的．本文中，我们设定 S_0= 0.7, r= 10，假阴性率 = 0.05，计算得到 l=105．我们试着调整 S_0 的设置，使其在合理的范围内变化(0.6～0.9)，但发现该参数的变化对结果的影响不大．因此，在文中仅讨论 S_0= 0.7 得到的结果．

由于局部敏感哈希方法产生的假阳性数据在后续的步骤可以去除，因此我们只控制了假阴性率．在最后一步去除候选集合的假阳性数据时，我们设定集合的行向量相似度阈值 S^*_{multiple} = 0.7(参见本附录的第 2 部分：去除候选集合的假阳性)．也就是说，最终产生的每一个保守基因共表达团，边的数据集相似性至少为 0.7.

3.2 Affinity propagation 聚类方法的参数设定

该方法的实现软件可以从 http://www.psi.toronto.edu/affinitypropagation/ 下载．该方法需要两个文件来描述输入的图结构与聚类的参数．第一个文件描述图的结构．按照该方法的建议，图中如果两个节点之间有边相连，该边的权重可以设定为一个很小的负数，在我们的应用中设定为−1.0e−10．第二个文件描述选择节点作为类的 exemplars 的倾向性．我们认为在图中度越大的节点越倾向于是一个类的 exemplar．因此设定每一个节点成为类的 exemplar 的倾向性与该节点的度的相关．

4 超几何方法计算类

文中计算一个基因类是否显著的共享 GO 注释或者转录调控因子，使用的都是超几何分布的统计方法．下面以计算基因类共享转录调控因子为例，解释该方法：

$$P(n, n_1, m, m_1) = \sum_{x=m_1}^{m} \frac{C_{n1}^{x} C^{m-x}_{n-n_1}}{C_n^m}$$

设当前数据共有 n 个基因有转录因子的信息．其中 n_1 个基因注释被转录因子 A 调控．上述公式计算的是从这 n 个基因中随机抽取 m 个基因，恰好有≥m_1 个基因被转录因子 A 调控的概率．该概率可以用来衡量一个基因类是否显著的共享某种属性．概率越小，代表该基因类越显著的共享此　属性．

参 考 文 献

1 Ploner A, Miller L, Hall P, *et al.* Correlation test to assess low-level processing of high-density oligonucleotide microarray data. BMC Bioinformatics, 2005, **6**(1): 80

2 Spellman P T, Sherlock G, Zhang M Q, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 1998, **9**(12): 3273～3297

3 Gasch A, Huang M, Metzner S, *et al.* Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog mec1p. Mol Biol Cell, 2001, **12**(10): 2987～3003

4 Gasch A, Spellman P, Kao C, *et al.* Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell, 2000, **11**(12): 4241～4257

5 Zhu G, Spellman P T, Volpe T, *et al.* Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. 2000, **406**(6791): 90～94

6 Sudarsanam P, Iyer V R, Brown P O, *et al.* Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae. Proc Nat Acad Sci, 2000, **97**(7): 3364～3369

7 Chu S, DeRisi J, Eisen M, *et al.* The rranscriptional program of sporulation in budding yeast. Science, 1998, **282**(5389): 699～705

8 DeRisi J L, Iyer V R, Brown P O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 1997, **278**(5338): 680～686

9 Roberts C J, Nelson B, Marton M J, *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science, 2000, **287**(5454): 873～880

10 Ronen M, Botstein D. Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. Proc Nat Acad Sci, 2006, **103**(2): 389～394

11 Yoshimoto H, Saltsman K, Gasch A P, *et al.* Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces cerevisiae*. J Biol Chem, 2002, **277**(34): 31079～31088

12 Cohen E, Datar M, Fujiwara S, *et al.* Finding interesting associations without support pruning. IEEE Transactions on Knowledge and Data Engineering, 2001, **13**(1): 64～78

13 Buhler J. Efficient large-scale sequence comparison by locality-sensitive hashing. Bioinformatics, 2001, **17**(5): 419～428