研究论文 / Research Article www.cjb.org.cn

生物物理学报 2011 年 3 月 第 27 卷 第 3 期: 242-256
ACTA BIOPHYSICA SINICA Vol.27 No.3 Mar. 2011: 242-256

结合基因表达数据和 ChIP-chip 数据的酵母转录调控模块的识别

王晓敏， 王正志， 王广云， 黎刚果

国防科技大学机电工程与自动化学院，长沙 410073

收稿日期：2010-03-22；**接受日期**：2010-05-11
基金项目：国家自然科学基金项目（60835005），国防科大科研计划项目（JC09-03-04）
通讯作者：王晓敏，电话：(0731)84574991，E-mail：wangxiaomin2006@gmail.com

摘要：活细胞依赖其众多的转录调控模块来实现复杂的生物功能，识别转录调控模块对深入理解细胞的功能及其转录机制有着重要的意义。本文结合酵母基因表达数据和 ChIP-chip 数据，提出了一种转录调控模块识别算法。该算法通过采用不同的 *P* 值阈值分别得到了核心集和粗糙集，然后对核心集和粗糙集进行判别，最后对基因进行扩展之后得到基因转录调控模块。将该算法运用到两个酵母基因表达数据中，得到了一些具有显著生物学意义的基因转录调控模块。与其它算法相比，该算法不仅可以识别含有较多基因的转录调控模块，而且可以识别一些其它算法不能识别的基因转录调控模块。识别得到的基因转录调控模块有着不同的生物学功能，并且有助于进一步理解酵母的转录调控机制。

关键词：基因表达数据；ChIP-chip 数据；转录调控模块；调控因子；酵母
中图分类号：Q-332
DOI：10.3724/SP.J.1260.2011.00242

引 言

许多细胞复杂功能和行为的实现都依赖于其复杂的基因调控网络。基因调控网络分析和重建已经成为目前研究的一个热点[1]。研究表明，基因调控网络的结构是层级化、模块化的[2]，正是由于基因调控网络的模块化特性，活细胞能够通过模块中的调控因子同时调控模块中的多个基因，进而实现复杂的功能。这些被一个或者几个调控基因调控并具有相似功能，或参与同一生物过程的基因的集合被称为转录调控模块。通过推断转录调控模块而不是直接推断基因调控网络来理解细胞内的调控机制，可以极大降低推断过程的复杂度[3]。

高通量实验技术的出现使得通过计算重构转录调控模块成为可能。一是基因表达芯片技术[4]，它的出现使得成千上万的基因表达数据可以同时获得。随着大量基因表达数据的出现，全基因组范围的表达分析已成功用于揭示生物转录调控模块对不同细胞过程的控制[5]。在对基因表达数据分析时，通常先通过聚类算法从表达数据中识别共表达基因，然后通过模式识别算法从启动子序列信息中推断出共表达基因中的共调控基因，并且识别调控因子有可能结合的模式[6]。然而这种方式不能直接证实基因间共调控的关系，不能识别出与转录调控模块相关的转录调控因子，也不能识别那些基因间相关性弱的共调控基因。二是染色

体免疫共沉淀芯片技术 (chromatin immunoprecipitation-chip，ChIP-chip)[7]，它的出现使得可以在全基因组范围内掌握转录调控因子的结合情况。通过染色质免疫共沉淀芯片技术可以直接获得转录调控因子与被调控基因间相互结合的数据 (ChIP-chip 数据)。根据 ChIP-chip 数据，Lee 等[8]构建了转录调控因子与被调控基因间的相互作用网络；Harbison 等[9]构建了酵母转录调控因子与被调控基因间结合关系的初始图。但是 ChIP-chip 数据有很多噪声，且依赖于结合数据 *P* 值阈值的选择，不同选择时可能包含很多假阳性或假阴性数据。

既然单独使用两种数据识别基因转录调控模块时都各自存在缺点，结合两种数据来识别转录调控模块就成为更好的选择。一些研究者开发了一些结合两种数据来识别转录调控模块的算法。SAMBA 算法[10]和 ECIM 算法[11]将 ChIP-chip 数据进行转换，然后采用聚类算法来得到转录调控模块。虽然聚类算法所得到的转录调控模块存在一定的生物学意义，但它获得的转录调控模块中调控因子数目过多，给模块生物学意义的深层次分析带来困难。并且 ECIM 算法采用层次聚类的思想不能确定最终所得的转录调控模块数目，还需要研究人员另行指定，存在很大的不确定性，会导致结果不准确。除了聚类算法外，其它一些算法则通过选择结合数据 *P* 值阈值来实现模块的识别。GRAM 算法[12]采用两步策略：首先得到核心模块，然后通过对核心模块进行扩展得到最终的转录调控模块。ReMoDiscovery 算法[13]采用类似于 GRAM 算法的两步策略得到转录调控模块。MOFA 算法[14]则放宽了结合 *P* 值的阈值，从而得到粗糙基因集，进而通过从粗糙基因集中删除基因的策略来获得最终的转录调控模块。这三种算法中，GRAM 算法采用了较小的 *P* 值阈值来进行初始转录调控模块的选择，使得其最终得到的转录调控模块较小，但结果的假阴性率较高；ReMoDiscovery 算法比 GRAM 算法采用了更严格的阈值标准，使得其最终得到的转录调控模块数更少，因此假阴性率也较高；而 MOFA 算法则采用了较大的 *P* 值阈值标准，得到了相对粗糙的基因集，其得到的转录调控模块数目要多一些，且转录调控模块含有更多的基因数，但由于仅采用从粗糙集中去除基因的方式来识别转录调控模块，引入了较多的假阳性数据。

为了得到更为准确的转录调控模块，降低假阳性率和假阴性率，本文提出了一种新的转录调控模块识别算法 (transcription regulatory module identifying algorithm，TRMIA)。该算法首先引入局部相关系数对 ChIP-chip 数据进行了加强，使得部分在较严的 *P* 值阈值标准下不能保留的调控因子与被调控基因间的相互结合关系得以保留，然后在两个水平上分别得到了粗糙基因集和核心基因集，最后在对核心基因集进行判别和基因扩展的同时，对粗糙基因集中除核心基因集外的基因集进行判别和基因删除，进而得到转录调控模块。在酵母数据的实验中，与其它算法相比，TRMIA 算法能得到更为准确的转录调控模块，且能识别一些其它算法没有识别的转录调控模块。

材料与方法

实验数据

基因表达数据

本文选用两个基因表达数据集。第一个是 Spellman 等人[15]使用的时间序列数据，该数据集包含了酵母细胞循环过程中 77 个时间点的基因表达值。第二个数据集是 Gasch 等

人[16]使用的基因表达数据，它包括了不同的环境变化，比如：温度变化、渗压变化、氨基酸状态变化和不同化学试剂的增减等情况下基因的表达数据。本文中基因表达数据用矩阵的形式来表示，每一行表示一个基因在不同条件的表达值，每一列表示在某一条件下，所有基因的表达值。

ChIP-chip 数据

对酵母而言，目前有两个基因组范围内的 ChIP-chip 数据集。一个是 Lee 等人的数据集[8]，另一个是 Harbison 等人的数据集[9]。因为 Harbison 等人的数据集包含比 Lee 等人的数据集更多的转录调控因子结合数据，所以本文中使用 Harbison 等人的数据集。文中 ChIP-chip 结合数据也用矩阵来表示，每一行表示一个基因与所有调控因子之间的结合情况，每一列表示一个调控因子与所有基因间的结合情况。

方法

在具体介绍 TRMIA 算法前，先给出一些参数的定义。

$E=\{e_{ij}\}$是基因表达数据矩阵，行表示一个基因在不同条件下的表达值，e_{ij} 表示基因 i 在条件 j 下的表达值。

$C=\{c_{ij}\}$表示 ChIP-chip 结合数据矩阵，行表示基因，列表示转录调控因子，c_{ij} 表示转录调控因子 j 结合到基因 i 上的结合 P 值。

$B=\{b_{ij}\}$是结合分数矩阵，行表示基因，列表示转录调控因子，b_{ij} 表示转录调控因子 j 结合到基因 i 上的结合分数。文中结合分数矩阵是通过结合基因表达数据和 ChIP-chip 结合数据，并进行一定变换得到的。

F 表示转录调控因子的集合。TF 表示 F 的一个子集，$G_\delta(TF)$表示与 TF 中所有调控因子结合分数都大于 δ 的基因的集合，具体表示为 $G_\delta(tf_a\cdots,tf_c)=(g_x,\cdots,g_y)$，其中，$tf_a\cdots,tf_c$ 为 TF 中的调控因子，$g_x,\cdots,g_y$ 为集合中所包括的基因。$H_\delta\{G_\delta(TF),TF\subset F\}$表示在给定分数阈值 δ 的条件下，F 中所有 TF 对应的 $G_\delta(TF)$基因集的集合。

图 1 给出了 TRMIA 算法的结构框图。由结构框图可以看出，算法主要分为两部分。第一部分为结合分数矩阵的获得；第二部分为最终基因转录调控模块的获得。在结合分数矩阵获得过程中，算法引入了局部相关系数，将其与 ChIP-chip 数据 P 值相结合构建了新的结合分数矩阵，这样使得部分在较严的 P 值阈值标准下不能保留的调控因子和被调控基因间的相互结合关系得以保留；在最终基因转录调控模块获得过程中，采用两种阈值标准得到了转录调控模块的核心集和粗糙集。然后对核心集和粗糙集中的模块进行判别和扩展得到最终的转录调控模块集。

结合分数矩阵的获得

对于结合数据，研究表明，当结合 P 值大于 0.01 时，转录调控因子与基因间的结合关系的置信度很低并且大多数情况下不能被实验验证；当结合 P 值小于 0.001 时，转录调控因子与基因间的结合关系具有很高的置信度，并且通常能被实验验证；当 P 值在 0.01 与 0.001 之间时，转录调控因子与基因间的结合关系很模糊，且这些关系中有一部分能够通过实验验证[8]。因此，根据不同的结合 P 值，采用了不同的方式对结合分数矩阵进行赋值。具

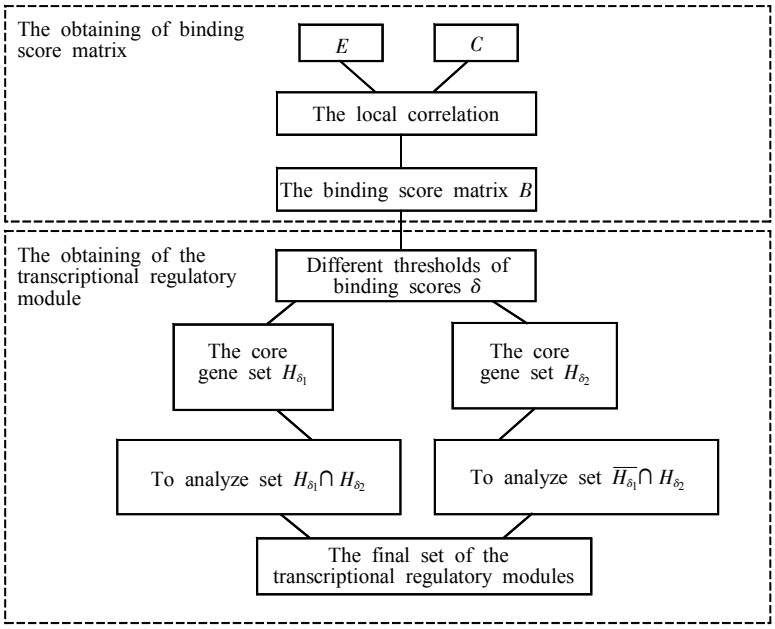

图 1 TRMIA 结构框图 算法主要分为两部分：结合分数矩阵的获得和转录调控模块的获得。在结合分数矩阵获得过程中，引入局部相关系数，将其与 ChIP-chip 数据 *P* 值相结合构建结合分数矩阵。在转录调控模块获得过程中，采用两种阈值标准得到了转录调控模块的核心集和粗糙集，然后对核心集和粗糙集分别进行判别分析构建最终的转录调控模块集

Fig.1 Frame of TRMIA The algorithm consists of two steps: the binding score matrix obtaining procedure and the transcriptional regulatory module obtaining procedure. During the binding score matrix obtaining procedure, the local correlation is introduced and combined with the *P* value of ChIP-chip to construct the binding score matrix. During the binding score matrix obtaining procedure, two different *P* value thresholds are used to obtain the core and loose sets of transcriptional regulatory modules. Then the core and loose sets of transcriptional regulatory modules are analyzed to construct the final set of transcriptional regulatory modules

体的赋值表达式如下：

$$b_{ij}=\begin{cases}0, & \text{如果 } c_{ij}>0.01\\ -\log_{10}c_{ij}\times P, & \text{如果 } 0.01<c_{ij}\leqslant 0.01\\ -\log_{10}c_{ij}, & \text{如果 } c_{ij}\leqslant 0.001\end{cases} \tag{1}$$

其中 *P* 表示基因 *i* 和转录调控因子 *j* 间的局部相关系数，定义如下：

$$P=\max\{abs(P_{t_1,t_2})\} \tag{2}$$

$$P_{t_1,t_2}=\frac{\sum_{t=t_1}^{t_2}(e_{it}-\overline{e_i})(e_{jt}-\overline{e_j})}{\sqrt{\sum_{t=t_1}^{t_2}(e_{it}-\overline{e_i})^2}\sqrt{\sum_{t=t_1}^{t_2}(e_{jt}-\overline{e_j})^2}} \tag{3}$$

这里，$1\leqslant t_1\leqslant t_2\leqslant n$，$n$ 为表达数据矩阵的列数，即总的条件数目。$\overline{e_i}$、$\overline{e_j}$ 分别为基因 *i* 和转录调控因子 *j* 在条件[t_1，t_2]间表达水平的平均值。局部相关系数的取值范围为[0,1]。当 t_1 取 1、t_2 取 n 时，局部相关系数即为常用的皮尔逊相关系数。*P* 的值越大表示相关的程度越高，作用越强烈。因为当[t_1，t_2]间的长度太短时，所得局部相关系数不能反映基因表达水平的变化规律，当[t_1，t_2]间的长度太长时，又不利于局部相关模式的识别。因此计算局部相

关系数时，以一定长度 $t_{len}=t_2-t_1$ 为计算单位，部分长度 t_{lap} 重叠地分段计算，得到一系列的相关系数，从中取值最大者作为局部相关系数。对于 t_{len} 和 t_{lap} 的取值，采用了与文献[17]中相同的设置。

基因转录调控模块的获得

在得到结合分数矩阵 B 后，采用两种不同的分数阈值来确定基因转录调控模块的搜索范围，经过搜索之后分别得到核心基因集 H_{δ_1}和粗糙基因集 H_{δ_2}。在搜索核心集时取分数阈值为 3，这与取 0.001 的结合 P 值相对应，此时可以保证所得到的基因集的准确性。在搜索粗糙集时，取分数阈值为 1.5，这使得很多局部相关系数很大的转录调控因子与基因间的相互作用不会因较小的结合 P 值而被舍弃。例如，当调控因子与基因间结合 P 值为 0.01 时，要求调控因子与基因间的相关系数的绝对值要不小于 0.75。同时因为这些局部相关系数很大的转录调控因子与基因间的相互作用往往都实际存在[17]，所以这样可以在降低算法假阴性率的同时，不会显著提高算法的假阳性率，使得算法能同时具有较低的假阴性率和假阳性率。

在得到核心基因集 H_{δ_1}和粗糙基因集 H_{δ_2}后，要分别对 $H_{\delta_1}\cap H_{\delta_2}$和$\overline{H_{\delta_1}}\cap H_{\delta_2}$中的基因集进行判别和基因的扩展与删除。

对 $H_{\delta_1}\cap H_{\delta_2}$中的基因集的判别和扩展

由于粗糙集是在较低分数总阈值情况下搜索得到的基因集，所以核心集中的任意一个基因集 $G^1_{\delta_1}$都会在粗糙集中有一个相应的扩展基因集 $G^2_{\delta_2}$。考虑对 $H_{\delta_1}\cap H_{\delta_2}$中的基因集的判别和扩展，即考虑对 $G^1_{\delta_1}\cap G^2_{\delta_2}$的一个判别和扩展。在核心集中可能含有对应各种调控因子数目的基因集，对它们的操作过程与含有两个转录调控因子基因集类似，所以，这里仅以含有两个转录调控因子的情况为例来说明具体操作过程。现假设基因集 $G^1_{\delta_1}$为 $G^1_{\delta_1}(tf_1,tf_2)=(g_1,\cdots,g_m)$，它在粗糙集中的扩展基因集 $G^2_{\delta_2}$可能存在以下几种情况：

1) $G^2_{\delta_2}(tf_1,tf_2,tf_3)=(g_1,\cdots,g_n)$，且 $m\leqslant n$，这表明由于分数阈值的降低，扩展基因集不仅转录因子数变多，而且基因数也增多。此时 $G^1_{\delta_1}\cap G^2_{\delta_2}=G^1_{\delta_1}$，需要首先判别 tf_3 是否被看作为该基因集的转录调控因子，然后对该基因集进行扩展。对于 tf_3 的判别，首先计算它与基因集 $G^1_{\delta_1}$中所有基因间的相关系数的均值作为其与该基因集的相关性，然后随机从剩余的基因中抽取与基因集 $G^1_{\delta_1}$相同个数的一组基因，并计算 tf_3 与该组基因的相关性。随机抽取 1000 次，得到 tf_3 与基因组相关性的经验分布，进而可以计算出 tf_3 与基因集 $G^1_{\delta_1}$相关性的显著性 P 值。如果该 P 值小于 0.05，则认为 tf_3 是基因集 $G^1_{\delta_1}$的调控因子。如果经过判别 tf_3 是基因集 $G^1_{\delta_1}$的转录调控因子，则需要再对 $G^1_{\delta_1}$的基因进行扩展。否则直接将基因集 $G^1_{\delta_1}$作为一个转录调控模块保留。假设经过判别，tf_3 是基因集 $G^1_{\delta_1}$的调控因子，则需要考虑将那些属于 $G^2_{\delta_2}$但不属于

$G^1_{\delta_1}$的基因对 $G^1_{\delta_1}$进行扩展。按照这些基因与三个转录调控因子的结合分数的总和大小将这些基因由大到小排序，从总和最大的基因开始，采用与判别转录调控因子与基因集相关性相同的方法，判别这些基因与基因集 $G^1_{\delta_1}$的相关性，将那些最终判断为与 $G^1_{\delta_1}$相关的基因扩展到 G_1 中，并将 G_1 扩展后的基因集作为一个转录调控模块保留。

2) $G^2_{\delta_2}(tf_1,tf_2,tf_3)=(g_1,\cdots,g_n)$，且 $n<m$，这表明虽然由于分数阈值的降低，扩展后转录因子变多，基因集中的基因数目变少。此时 $G^1_{\delta_1}\cap G^2_{\delta_2}=G^2_{\delta_2}$，只需要判别基因 $g_{n+1},\cdots,g_m$ 是否能对 $G^2_{\delta_2}$进行扩展，其扩展方法与情况 1) 中相同，最终将 $G^2_{\delta_2}$扩展后的基因集作为一个转录调控模块保留。

3) $G^2_{\delta_2}(tf_1,tf_2)=(g_1,\cdots,g_m,\cdots,g_{m+j})$，且 $j\geqslant 0$，这表明分数阈值的降低使得基因集中的基因数增加。此时 $G^1_{\delta_1}\cap G^2_{\delta_2}=G^1_{\delta_1}$属于 $H_{\delta_1}\cap H_{\delta_2}$。当 $j=0$ 时，两个基因集完全相同，则直接将其作为一个转录调控模块保留。当 $j>0$ 时，只需要考虑基因 $g_{m+1},\cdots,g_{m+j}$ 是否能对 $G^1_{\delta_1}$进行扩展，其扩展方法与情况 1) 中相同，最终将 $G^1_{\delta_1}$扩展后的基因集作为一个转录调控模块保留。

在上面的情况 1) 和情况 2) 中只考虑了扩展基因集 $G^2_{\delta_2}$含有 3 个转录调控因子的情况。当然扩展基因集 $G^2_{\delta_2}$可能含有 4 个或更多转录调控因子，此时对 $G^1_{\delta_1}\cap G^2_{\delta_2}$基因集的判别和扩展过程与情况 1) 和情况 2) 中所述类似，在此不再赘述。

对$\overline{H_{\delta_1}}\cap H_{\delta_2}$中的基因集的判别和扩展

粗糙集 H_{δ_2}中除了含有核心集 H_{δ_1}中基因集的相关扩展基因集外，还含有很多核心集没有的基因集。对属于$\overline{H_{\delta_1}}\cap H_{\delta_2}$中的基因集还需要对其进行判别和扩展。在$\overline{H_{\delta_1}}\cap H_{\delta_2}$中可能含有对应各种调控因子数目的基因集，对它们的操作过程与含有两个转录调控因子基因集类似，所以这里仅以含有两个转录调控因子的情况为例说明具体操作过程。现假设 $G^3_{\delta_2}(tf_4,tf_5)=(g_i,\cdots,g_k)$为$\overline{H_{\delta_1}}\cap H_{\delta_2}$中的一个基因集，此时，需要首先判别转录调控因子 tf_4 和 tf_5 分别与基因集 $G^3_{\delta_2}$的相关性，其判别方法与前面所述方法相同。若经过判别两者都与 $G^3_{\delta_2}$显著相关，则保留 $G^3_{\delta_2}$为转录调控模块。若经过判别只有一个（假设为 tf_4）与 $G^3_{\delta_2}$相关，则将 $G^3_{\delta_2}$中的基因按照这些基因与转录调控因子 tf_5 结合分数的大小将这些基因由小到大排序，每次删除结合分数最小的那个基因，然后再判断转录调控因子 tf_5 与剩下的基因所构成的基因集的相关性，直到转录调控因子 tf_5 与剩下的基因所构成的基因集判别为相关或剩下基因数目为 0 时结束。此时，剩下的基因集被保留作为基因转录调控模块。若经过判别两者都不与 $G^3_{\delta_2}$相关，则直接将 $G^3_{\delta_2}$删除。

经过对 $H_{\delta_1}\cap H_{\delta_2}$和$\overline{H_{\delta_1}}\cap H_{\delta_2}$中的所有基因集进行类似上面的操作后，得到最终的转录调控模块集。

模块功能类统计显著性评估

在识别出转录调控模块后，用 KEGG 数据库[18]作为这些模块功能分类的标准。假设实验总共包括 G 个基因，已知其中 m 个基因属于 KEGG 功能类 F，并且在算法识别的一个大小为 D 的模块中，有 h 个基因是属于 KEGG 通路 F。在 m 个属于通路 F 的基因随机分布在各类中的假设下，这 h 个基因应满足超几何分布[19]，且其功能显著性 P 值可以按下式进行计算。

$$P[X\geqslant h]=1-\sum_{i=0}^{h-1}\binom{D}{i}\binom{G-D}{m-i}\Big/\binom{G}{m} \tag{4}$$

如果该模块功能显著性 P 值及其统计显著性值都小于 0.05，则认为该模块具有功能 F。该模块功能显著性 P 值的统计显著性值的计算过程如下：从总体基因集中随机抽取 h 个基因，根据公式(4)可以计算出另一个功能显著性值。重复这一过程 1000 次，就可以得到一个功能显著性值的经验分布，进而可以根据该经验分布计算出该模块功能显著性 P 值的统计显著性值。

结　果

对于 Spellman 数据，TRMIA 算法识别了 297 个基因转录调控模块，一共包含有 1459 个基因，被 103 个调控因子调控。对于 Gasch 数据，TRMIA 算法识别了 338 个基因转录调控模块，一共包含了 1679 个基因，被 111 个调控因子调控。

算法所识别模块的有效性分析

通过对算法所识别模块的分析可知，TRMIA 算法识别了很多具有生物学意义的模块。1) 用 KEGG 数据库在 0.05 显著性水平下对所识别的模块的生物学意义进行检验，Spellman 数据所识别的 297 个模块中，共有 199 个转录调控模块在统计意义上具有显著生物学意义；Gasch 数据所识别的 338 个转录调控模块中，一共有 213 个转录调控模块在统计意义上具有显著生物学意义。2) 大多数调控因子功能与其所对应的基因转录调控模块的功能相一致。通过将调控因子并入转录调控模块基因中，并通过 GO 词素搜索软件 GO-TermFinder[20]来寻找其中相同的词素，Spellman 数据所识别的 103 个调控因子中有 78 个调控因子，在 0.01 的显著性水平上与其所对应的转录调控模块参与相同的生物过程或具有相同的功能；Gasch 数据所识别的 111 个调控因子中，有 83 个在 0.01 的显著性水平上与其所对应的转录调控模块参与相同的生物过程或具有相同的功能。例如：两个数据集中都识别了调控因子为 HIR1、HIR2 和 HIR3 的转录调控模块。该转录调控模块包含了六个组蛋白基因 HHF1、HHT1、HTB1、HTA1、HHF2 和 HHT2。三个调控因子 HIR1、HIR2 和 HIR3 与转录调控模块的六个基因都被 GO-TermFinder 注释到核小体组装这一词素，其 P 值为 7.49×10^{-21}。这说明调控因子 HIR1、HIR2 和 HIR3 与其所对应的转录调控模块功能一致。实际上，HIR1、HIR2、HIR3 在细胞循环过程中对这六个组蛋白基因的调控作用已经在文献中被证实[21]。3) 大多数情况下，转录调控模块对应多个调控因子。Spellman 数据所识别的 297 个转录调控模块中，仅有 53 个转录调控模块的调控因子数为 1，其余模块都对应两个以上的调控因

子；Gasch 数据所识别的 338 个转录调控模块中仅有 62 个转录调控模块的调控因子数为 1，其余模块都对应两个以上的调控因子。并且一个转录调控模块中的多个调控因子间往往存在相互作用。还是以六个组蛋白基因 HHF1、HHT1、HTB1、HTA1、HHF2 和 HHT2 的转录调控模块为例，它们的调控因子 HIR1、HIR2 和 HIR3 属于同一族的基因，而它们之间相互结合形成复合物以实现对组蛋白的调控[21]。

与其它算法的比较

表 1 给出了 GRAM、ReMoDiscovery 和 TRMIA 三种算法在 Spellman 数据中所识别的转录调控模块的基本情况。

表 1 不同算法在 Spellman 数据中所识别转录调控模块基本情况的比较

Table 1 Comparison of transcriptional regulatory modules identified by different algorithms on Spellman dataset

Method	No-SM/ No-TM(Ratio)	Average functional enrichment of SM	Genes			Regulatory factors		
			Mean	Min	Max	Mean	Min	Max
GRAM	138/274(50.3%)	0.0178	6.8	5	33	2.35	1	8
ReMoDiscovery	12/18(66.7%)	0.00538	67.72	6	200	3.5	2	6
TRMIA	199/297(67.1%)	0.0142	14.16	5	99	2.42	1	7

表中给出了具有显著功能的转录调控模块的数目 (No-SM)，转录调控模块的总数 (No-TM)，算法识别的模块中所含的基因以及调控因子的平均值 (Mean)，最小值 (Min) 和最大值 (Max)，以及具有重要功能的转录调控模块的平均功能富集度 (Average functional enrichment of SM)

The number of transcriptional regulatory modules with significant functions (No-SM), the total number of transcriptional regulatory modules (No-TM), and the mean (Mean), minimum (Min) and maximum (Max) number of genes and regulatory factors in the identified modules are displayed, as well as the average functional enrichment of the transcriptional regulatory modules with significant functions

由表 1 可以看出，ReMoDiscovery 算法识别的转录调控模块平均所含调控因子数目和基因数目都最多。这是因为 ReMoDiscovery 算法将序列模体数目也作为调控因子数目进行计数，同时，ReMoDiscovery 算法在基因扩展时仅采用基因表达间的相似性作为标准，其所得转录调控模块平均含有更多的调控因子数目和基因数目。虽然 ReMoDiscovery 算法每个转录调控模块平均所含调控因子数目和基因数目最多，但其所得到的转录调控模块最少。这是因为 ReMoDiscovery 算法将序列模体数据作为一个转录调控模块选择的标准，而现阶段序列模体数据还不完整，这使得该算法识别的转录调控模块数较少。TRMIA 算法和 GRAM 算法没有考虑序列模体，因此能够识别一些 ReMoDiscovery 算法不能识别出的转录调控模块。例如：TRMIA 算法和 GRAM 算法中都识别了调控因子为 INO2 和 INO4，并涉及到脂肪酸合成的转录调控模块[22]，而 ReMoDiscovery 算法不能识别这一转录调控模块。同时，由表 1 可以看出，TRMIA 算法、ReMoDiscovery 算法和 GRAM 算法所识别的显著功能模块数与模块总数的比率分别为 67.1%、66.7%和 50.3%。如前所述，正是由于 ReMoDiscovery 算法还考虑序列模体数据作为一个转录调控模块选择的标准，这种算法会使得所识别模块的假阳性率降低，并使得算法所识别模块的假阴性率升高，所以

ReMoDiscovery算法所识别的显著功能模块数与模块总数的比率要远高于 GRAM 算法，同时其所识别的模块总数要远比 GRAM 算法少很多。显然，TRMIA 算法所识别的显著功能模块数与模块总数的比率在三者之中最高，同时 TRMIA 算法所识别的模块总数在三者之中也最多，这说明比之其它两种算法，TRMIA 算法在保证所识别模块的假阳性率不升高的情况下，极大地降低了所识别模块的假阴性率。

由表 1 还可以看到，TRMIA 算法不仅比 GRAM 算法识别了更多的转录调控模块，而且还识别了更多含有生物学意义的转录调控模块。比如：调控因子 STE12、TEC1 与 DIG1 的复合物在交配和成丝过程中起到调控作用；在交配和成丝过程中，交配基因通过 STE12 结合位点 (信息素响应元素) 受 STE12 的调节；成丝基因受 STE12 与 TEC1 结合物的调控[23]。TRMIA 算法和 GRAM 算法两者都未能识别调控因子为 STE12、TEC1 和 DIG1 的转录调控模块，但两种算法都识别了调控因子为 STE12 和 DIG1 的转录调控模块。TRMIA 算法还识别了 GRAM 算法所没有识别的调控因子为 STE12 和 TEC1 的转录调控模块。该转录调控模块中的基因为 FUS3、PEP1、TEC1、PCL2、ACT1、MSB2、YJU2、MFA2、PRM1、ERG24、GPA1、STE12 和 AGA1。这 13 个基因中有 10 个基因与这两个调控因子间的相互作用都已得到证实。其中 STE12 和 TEC1 本身间的相互作用在前面已经提到。在交配过程中，FUS3 使得 TEC1 磷酸化，进而调节信息素反应过程中的交配转录产物[24]。有丝分裂原激活性激酶通路通过信号粘蛋白 MSB2 和有丝分裂原激活性激酶的层叠得到刺激，而有丝分裂原激活性激酶 HOG1，通过干扰有丝分裂原激活性激酶 KSS1 与 TEC1 间的信号转导来保证该通路的特异性[25]。MFA2 转录后的翻译也是受有丝分裂原激活性激酶 HOG1 的调节[26]。PRM1 是一个信息素调节的多扩展膜蛋白，在酵母交配过程中有利于质膜的融合[27]。GPA1 的转录水平会因 STE12 的突变而减少[28]。AGA1 的转录水平在 STE12-CELLS 中也会极大地减少[29]。在啤酒酵母的交配过程中，AGA1 由信息素诱导[30]。PCL2 的表达水平因 FUS3 和 KSS1 的协同作用而增加[31]。ACT1 会因响应交配信息素基因的激活而极化[32]。这是因为，虽然 TRMIA 和 GRAM 算法都采用了大小两种 P 值阈值的策略，但 TRMIA 算法在对核心基因集进行判别和基因扩展的同时，还对粗糙基因集中除与核心基因集交集外的基因集进行判别和基因删除，所以 TRMIA 算法能比 GRAM 算法识别更多的转录调控模块。

此外，与 GRAM 算法相比，TRMIA 算法识别的转录调控模块平均含有更多的调控因子，且含有更多的基因数目。比如：TRMIA 算法和 GRAM 算法都识别了调控因子为 RLM1 和 SWI4 的模块，且其转录调控模块的功能通过 GO-TermFinder 都注释为细胞壁组织，但 GRAM 算法仅识别了 5 个细胞壁组织相关功能基因，TRMIA 算法识别了 7 个细胞壁组织相关功能基因。这是因为 TRMIA 算法引入了局部相关系数对 ChIP-chip 数据进行了部分加强，使得部分在 GRAM 算法中不被考虑但实际存在的转录调控因子和调控基因间的相互作用得以保留，所以 TRMIA 算法能比 GRAM 算法识别出更为准确的转录调控模块。

模块的详细描述

对于 Spellman 数据和 Gasch 数据，Lemmens 等人采用 ReMoDiscovery 算法识别了氨基酸代谢、细胞循环、应激响应、交配和成丝等功能相关的模块，并且对这些功能相关的模块进行了较为详细的描述[13]。同样对于 Spellman 数据和 Gasch 数据，TRMIA 算法不仅识别

了 ReMoDiscovery 算法所识别出的氨基酸代谢、细胞循环、应激响应、交配和成丝等功能相关的模块，还识别了糖类代谢和脂质代谢功能相关的模块。下面仅对只被 TRMIA 算法所识别的糖类代谢和脂质代谢功能相关的模块中部分有文献证实的模块进行详细描述。

糖类代谢过程相关模块

两个数据集中都识别了糖类代谢相关模块。这些模块所包含的部分调控因子在糖类代谢过程中的作用已经得到了证实。在糖酵解过程中，GCR1、GCR2 为正调控因子，直接调控糖酵解基因的表达，并间接调控编码 TCA 循环和呼吸基因的表达[33]。在柠檬酸循环过程中，有氧条件下厌氧基因的转录由于 SSN6/TUP1 复合物受 DNA 结合蛋白 ROX1 的调控而受抑制[34~36]。并且 SUT1 在葡萄糖代谢过程中起到葡萄糖转运的作用[37]。NGR1 作为一个 DNA 结合抑制因子，也通过结合 SSN6/TUP1 的复合物来调控 STA1 基因表达，实现对葡萄糖的抑制[38]。STB5 结合并调控磷酸戊糖通路中的大多数基因在氧化应激响应中的表达[39]。GAL4 是在半乳糖代谢过程中 GAL 族基因的激活因子[40]，并且 NRG1 也调控 GAL 族基因对半乳糖代谢的抑制[41]。HAP2、HAP3 和 HAP4 对一些葡萄糖抑制基因的表达起到调控作用[42,43]。TYE7 和 GCR1 作用于并行的通路中来激活糖酵解基因的表达[44]。此外，这些涉及到糖类代谢的模块还包含调控因子 OAF1、PIP2 和 UME6。这些调控因子诱导过氧化物酶基因参与 β 氧化过程[45]。这说明 β 氧化过程可能与糖类代谢有关。

脂质代谢相关模块

两个数据集中都识别了脂质代谢相关模块。这些模块所包含的部分调控因子在脂质代谢过程中的作用已经得到了证实。INO2 与 INO4 的复合物共同调控脂肪酸合成的生物过程[22,46]。UPC2 是类固醇生物合成过程中的调控因子，它与 ECM22 一起调控类固醇生物合成过程基因 ERG2 和 EGR3[22,47]。PDR3 除了调控 ATP 结合盒转运蛋白编码基因的表达[48]，还在膜脂和细胞壁生物合成过程中起到调控作用[49]。OAF1 和 PIP2 的复合物的活性依赖于油酸的诱导[50]。而 ADR1 涉及到 OAF1 和 PIP2 间的结合，而反过来 OAF1 和 PIP2 的复合物对于 ADR1 结合到一些油酸响应相关基因非常重要[51]。

讨　论

对于调控网络的研究一般分为三个层次：motif 层次、模块层次和网络层次。模块层次作为一个中间的层次，起着承上启下的作用，对其进行分析不仅使得调控网络的构建更为简单、合理，而且有利于对网络中 motif 结构的分析和理解。图 2 给出了部分涉及到细胞循

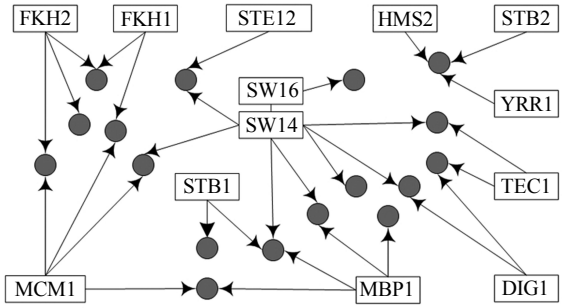

图 2　基于细胞循环功能模块的局部调控网络
图形表示模块，方形表示调控因子

Fig.2　Module-based local regulatory network with function of cell cycle　Modules are represented by small circles. Regulators are represented by rectangles

环的模块所构成的局部网络图。如果将其它各种功能模块所构成的局部网络图拼合起来，便很容易构建出全基因组范围的调控网络，并且所构建的网络具有很好的功能解释性。

由图 2 可知，大部分模块都具有两个以上的调控因子。在 2.1 节中已经提到，当一个转录调控模块对应多个调控因子，这些调控因子间往往存在相互作用，这些相互作用可能是直接的相互结合，也可能是通过其它介质的间接相互作用。具体而言，这些调控因子间及其与模块间的相互关系可以分为以下几种类型：1) 一个调控因子与 DNA 相结合，其它调控因子通过与它相结合来对模块中的基因进行调控。比如：SWI6 与 MBP1 和 SWI4 间及其与模块间的关系就属于这一种类型。具体而言就是 SWI6 通过分别与 DNA 结合蛋白 MBP1 和 SWI4 结合形成复合物 MBF 和 SBF 来转录调控模块中的基因[52]。2) 两个调控因子都与 DNA 结合，且它们具有相同的结合序列。此时这两个调控因子都直接转录调控模块中的基因，但它们通过相同的结合序列而间接相关。比如：调控因子 STE12、SWI4 和 SWI6 及其对应的转录调控模块。虽然 STE12 和 SBF (SWI4 和 SWI6 的复合物) 分别特异调控交配与成丝和细胞循环相关模块，但它们通过相同的结合序列而间接相关[53]，进而使得这三个调控因子能同时成为同一个模块的调控因子。这说明它们所共同调控的模块既是细胞循环相关模块，也是交配与成丝相关模块。3) 两个调控因子都与 DNA 相结合，但不具有相同的结合序列。它们之间的相互作用通过共同募集相同的调控因子来实现，进而形成复合物实现对模块基因的调控。比如：调控因子 ROX1 和 MOT 都是 DNA 结合蛋白，且结合位点不同，但都通过募集 SSN6/TUP1 形成复合物来抑制缺氧基因[54]。实际上，这两个调控因子都可以在对方不表达的情况下单独与 SSN6/TUP1 形成复合物实现抑制作用[55]。

正是通过上面所述的对调控因子间及其与模块间相互关系的分析，可以更容易地分析出模块的调控因子与其调控基因间的 motif 类型，进而可以通过对 motif 的分析更清晰地了解模块的细节特征。到目前为止，在酵母转录调控网络中已经识别出 6 种网络调控 motif：自调控 (autoregulation)、多组件回路 (multi-component loop)、前馈回路、单输入 motif (single input motif)、多输入 motif (multi-input motif) 和调控链 (regulator chain)，从下面几点来说明基于模块层次的研究与基于 motif 层次的研究间的相互关系。1) STE12 是一个序列特异的 DNA 结合蛋白[8]，而且其结合位点就在 STE12 基因编码区的上游序列中，因此它能形成自调控的 motif。TRMIA 算法识别了很多对应调控因子 STE12 的转录调控模块，在对这些模块的调控机制进行深入研究时，就需要考虑由 STE12 形成的自调控。2) ROX1 和 YAP6 都是序列特异的 DNA 结合蛋白[56]，且它们的结合位点分别在对方的上游序列中相互调控，进而形成了多组件回路。TRMIA 算法中识别了对应调控因子 ROX1、YAP6 和 SKN7 的模块，说明该模块中便含有 ROX1 和 YAP6 构成的多组件回路。ROX1 和 SKN7 都是氧化应激响应过程中的调控因子[57]，因此推测 YAP6 通过与 ROX1 构成多组件回路，参与对氧化应激响应过程的调控。3) MCM1 和 SWI4 也都是序列特异的 DNA 结合蛋白[56,58]，MCM1 的结合位点在 SWI4 的上游序列中，同时其结合位点也存在于 CLB2 的上游序列中，这样便形成了前馈回路。因此，MCM1 除了可以通过结合到 SWI4 来调控 CLB2，也可以直接调控 CLB2。实际上，在这个回路中，与 MCM1 和 CLB2 相互作用的蛋白除了 SWI4 外，还应该包括结合到 SWI4 上的调控因子 SWI6。TRMIA 算法识别了对应调控因子 MCM1、

SWI4和 SWI6 的模块，同时该模块中包含基因 CLB2，所以该模块中含有这样的前馈回路。4) LEU3、FHL1、RAP1 和 YAP5 也都是 DNA 结合蛋白[56,59]，且它们的结合位点存在于多个基因的上游序列中。LEU3 分别结合到 LEU1、BAT1 和 ILV2 的上游序列中时，形成了单输入 motif。而 FHL1、RAP1 和 YAP5，都可以分别结合到 RPL2B、RPL16A、RPS21B 和 RPS22A 的上游序列，这样形成了多输入 motif。TRMIA 算法识别了对应调控因子 LEU3 的模块，该模块含有它所结合的三个基因 LEU1、BAT1 和 ILV2。TRMIA 算法也识别了对应调控因子 FHL1、RAP1 和 YAP5 的模块，但它们所结合的四个基因中只有两个基因 RPL2B 和 RPS22A 属于该模块。另外两个基因 RPL16A 和 RPS21B 都出现在对应调控因子 RAP1、YAP5 和 GAT3 的模块中，同时，RPL16A 出现在对应 FHL1、PDR1 和 RAP1 的模块中，RPS21B 出现在对应 FHL1 的模块中。这说明单输入 motif 往往存在于一个模块内部，而多输入 motif 则往往存在于模块与模块之间，反映不同模块之间的相互联系。此外，调控链类似由多个单输入 motif 层次链接而成。TRMIA 算法没有识别对应调控因子 MCM1 和 SWI5 的模块，但 SWI5 出现在调控因子为 FKH2 和 MCM1 的模块中，ASH1 出现在调控因子为 SMP1 和 SWI5 的模块中，这说明调控链可以反映出多个模块间的一个先后链接的关系。

总之，通过模块层次的研究，可以构建以模块为基础的全基因组范围内的调控网络，可以知道所识别的模块中及模块间存在怎样的调控关系，应该形成何种类型的网络 motif，进而可以构建更为清晰的转录调控网络，有利于进一步深入理解转录调控机制。

致谢 感谢第三军医大学的倪青山博士对文稿的编辑整理

参考文献：

1. Cavalieri D, De Filippo C. Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug Discov Today*, 2005, 10(10): 727~734
2. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA*, 2004, 101(9): 2981~2986
3. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 2003, 34(2): 166~176
4. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997, 278(5338): 680~686
5. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA. Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell*, 2001, 12 (2): 323~337
6. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet*, 1999, 22(3): 281~285
7. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 2001, 409(6819): 533~538
8. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 2002, 298 (5594): 799~804
9. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 2004, 431(7004): 99~104
10. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA*, 2004, 101(9): 2981~2986
11. Liu X, Jessen WJ, Sivaganesan S, Aronow BJ, Medvedovic M. Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and

ChIP-chip data. *BMC Bioinformatics*, 2007, 8: 283
12. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 2003, 21 (11): 1337~1342
13. Lemmens K, Dhollander T, de Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, de Moor B, Marchal K. Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol*, 2006, 7(5): R37
14. Wu WS, Li WH, Chen BS. Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics*, 2006, 7: 421
15. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, 1998, 9(12): 3273~3297
16. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 2000, 11(12): 4241~4257
17. 王广云. 肿瘤基因芯片表达数据分析相关问题研究. 湖南长沙: 国防科学技术大学自动化所, 学位论文. 2009: 40~45
Wang GY. Research on relevant problems of tumor DNA microarray expression data analysis. Institute of automation, National university of defense technology, Changsha, Hunan. 2009: 40~45
18. Topisirovic I, Culjkovic B, Cohen N, Perez JM, Skrabanek L, Borden KL. The proline-rich homeodomain protein, PRH, is a tissue-specific inhibitor of eIF4E-dependent cyclin D1 mRNA transport and growth. *EMBO J*, 2003, 22 (3): 689~703
19. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 2006, 22 (19): 2405~2412
20. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO-TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 2004, 20(18): 3710~3715
21. Prochasson P, Florens L, Swanson SK, Washburn MP, Workman JL. The HIR corepressor complex binds to nucleosomes generating a distinct protein/DNA complex resistant to remodeling by SWI/SNF. *Genes Dev*, 2005, 19(21): 2534~2539
22. Nohturfft A, Zhang SC. Coordination of lipid metabolism in membrane biogenesis. *Annu Rev Cell Dev Biol*, 2009, 25: 539~566
23. Chou S, Lane S, Liu H. Regulation of mating and filamentation genes by two distinct Ste12 complexes in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 2006, 26 (13): 4794~4805
24. Chou S, Zhao S, Song Y, Liu H, Nie Q. Fus3-triggered Tec1 degradation modulates mating transcriptional output during the pheromone response. *Mol Syst Biol*, 2008, 4:212
25. Birkaya B, Maddi A, Joshi J, Free SJ, Cullen PJ. Role of the cell wall integrity and filamentous growth mitogen-activated protein kinase pathways in cell wall remodeling during filamentous growth. *Eukaryot Cell*, 2009, 8(8): 1118~1133
26. Vasudevan S, Garneau N, Tu Khounh D, Peltz SW. p38 mitogen-activated protein kinase/Hog1p regulates translation of the AU-rich-element-bearing MFA2 transcript. *Mol Cell Biol*, 2005, 25(22): 9753~9763
27. Aguilar PS, Engel A, Walter P. The plasma membrane proteins Prm1 and Fig1 ascertain fidelity of membrane fusion during yeast mating. *Mol Biol Cell*, 2007, 18 (2): 547~556
28. Nakayama N, Kaziro Y, Arai K, Matsumoto K. Role of STE genes in the mating factor signaling pathway mediated by GPA1 in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 1988, 8(9): 3777~3783
29. Oehlen LJ, McKinney JD, Cross FR. Ste12 and Mcm1 regulate cell cycle-dependent transcription of FAR1. *Mol Cell Biol*, 1996, 16(6): 2830~2837
30. Huang G, Dougherty SD, Erdman SE. Conserved WCPL and CX4C domains mediate several mating adhesin interactions in *Saccharomyces cerevisiae*. *Genetics*, 2009, 182(1): 173~189
31. Cherkasova V, Lyons DM, Elion EA. Fus3p and Kss1p control G1 arrest in *Saccharomyces cerevisiae* through a balance of distinct arrest and proliferative functions that operate in parallel with Far1p. *Genetics*, 1999, 151 (3): 989~1004
32. Butty AC, Perrinjaquet N, Petit A, Jaquenoud M, Segall JE, Hofmann K, Zwahlen C, Peter M. A positive feedback loop stabilizes the guanine-nucleotide exchange factor Cdc24 at sites of polarization. *EMBO J*, 2002, 21(7):1565~1576
33. Sasaki H, Uemura H. Influence of low glycolytic activities in gcr1 and gcr2 mutants on the expression of other metabolic pathway genes in *Saccharomyces cerevisiae*. *Yeast*, 2005, 22(2): 111~127
34. Ter Linde JJ, Steensma HY. A microarray-assisted screen for potential Hap1 and Rox1 target genes in *Saccharomyces cerevisiae*. *Yeast*, 2002, 19(10): 825~840
35. McCammon MT, Epstein CB, Przybyla-Zawislak B, McAlister-Henn L, Butow RA. Global transcription analysis of Krebs tricarboxylic acid cycle mutants reveals an alternating pattern of gene expression and effects on hypoxic and oxidative genes. *Mol Biol Cell*, 2003, 14 (3): 958~972
36. Klinkenberg LG, Mennella TA, Luetkenhaus K, Zitomer RS. Combinatorial repression of the hypoxic genes of *Saccharomyces cerevisiae* by DNA binding proteins Rox1 and Mot3. *Eukaryot Cell*, 2005, 4(4): 649~660
37. Weierstall T, Hollenberg CP, Boles E. Cloning and characterization of three genes (SUT1-3) encoding glucose

transporters of the yeast *Pichia stipitis*. *Mol Microbiol*, 1999, 31(3): 871~883
38. Park SH, Koh SS, Chun JH, Hwang HJ, Kang HS. Nrg1 is a transcriptional repressor for glucose repression of STA1 gene expression in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 1999, 19(3): 2044~2050
39. Larochelle M, Drouin S, Robert F, Turcotte B. Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. *Mol Cell Biol*, 2006, 26(17): 6690~6701
40. Bhat PJ, Hopper JE. Overproduction of the GAL1 or GAL3 protein causes galactose-independent activation of the GAL4 protein: Evidence for a new model of induction for the yeast GAL/MEL regulon. *Mol Cell Biol*, 1992, 12(6): 2701~2707
41. Zhou H, Winston F. NRG1 is required for glucose repression of the SUC2 and GAL genes of *Saccharomyces cerevisiae*. *BMC Genet*, 2001, 2: 5
42. Bowman SB, Zaman Z, Collinson LP, Brown AJ, Dawes IW. Positive regulation of the LPD1 gene of *Saccharomyces cerevisiae* by the HAP2/HAP3/HAP4 activation system. *Mol Gen Genet*, 1992, 231(2): 296~303
43. Ulery TL, Jang SH, Jaehning JA. Glucose repression of yeast mitochondrial transcription: Kinetics of derepression and role of nuclear genes. *Mol Cell Biol*, 1994, 14(2): 1160~1170
44. Nishi K, Park CS, Pepper AE, Eichinger G, Innis MA, Holland MJ. The GCR1 requirement for yeast glycolytic gene expression is suppressed by dominant mutations in the SGC1 gene, which encodes a novel basic-helix-loop-helix protein. *Mol Cell Biol*, 1995, 15(5):2646~2653
45. Schuller HJ. Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. *Curr Genet*, 2003, 43(3): 139~160
46. Schwank S, Ebbert R, Rautenstrauss K, Schweizer E, Schuller HJ. Yeast transcriptional activator INO2 interacts as an Ino2p/Ino4p basic helix-loop-helix heteromeric complex with the inositol/choline-responsive element necessary for expression of phospholipid biosynthetic genes in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 1995, 23(2): 230~237
47. Vik A, Rine J. Upc2p and Ecm22p, dual regulators of sterol biosynthesis in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 2001, 21(19): 6395~6405
48. Katzmann DJ, Hallstrom TC, Mahe Y, Moye-Rowley WS. Multiple Pdr1p/Pdr3p binding sites are essential for normal expression of the ATP binding cassette transporter protein-encoding gene PDR5. *J Biol Chem*, 1996, 271(38): 23049~23054
49. DeRisi J, van den Hazel B, Marc P, Balzi E, Brown P, Jacq C, Goffeau A. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Lett*, 2000, 470(2): 156~160
50. Phelps C, Gburcik V, Suslova E, Dudek P, Forafonov F, Bot N, MacLean M, Fagan RJ, Picard D. Fungi and animals may share a common ancestor to nuclear receptors. *Proc Natl Acad Sci USA*, 2006, 103(18): 7077~7081
51. Karpichev IV, Durand-Heredia JM, Luo Y, Small GM. Binding characteristics and regulatory mechanisms of the transcription factors controlling oleate-responsive genes in *Saccharomyces cerevisiae*. *J Biol Chem*, 2008, 283(16): 10264~10275.
52. Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, 1993, 261(5128): 1551~1557
53. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*, 2004, 5(8): R56
54. Klinkenberg LG, Webb T, Zitomer RS. Synergy among differentially regulated repressors of the ribonucleotide diphosphate reductase genes of *Saccharomyces cerevisiae*. *Eukaryot Cell*, 2006, 5(7): 1007~1017
55. Sertil O, Kapoor R, Cohen BD, Abramova N, Lowry CV. Synergistic repression of anaerobic genes by Mot3 and Rox1 in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 2003, 31(20):5831~5837
56. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*, 2009, 19(4): 556~566
57. Wong CM, Ching YP, Zhou Y, Eide DJ. Transcriptional regulation of yeast peroxiredoxin gene TSA2 through Hap1p, Rox1p, and Hap2/3/5p. *Free Radic Biol Med*, 2003, 34(5): 585~597
58. Sidorova J, Breeden L. Analysis of the SWI4/SWI6 protein complex, which directs G1/S-specific transcription in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 1993, 13(2): 1069~1077
59. Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, Gebbia M, Talukder S, Yang A, Mnaimneh S, Terterov D, Coburn D, Li Yeo A, Yeo ZX, Clarke ND, Lieb JD, Ansari AZ, Nislow C, Hughes TR. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell*, 2008, 32(6): 878~887

Yeast Transcriptional Regulatory Module Identification by Integrating Gene Expression Data and ChIP-chip Data

WANG Xiaomin, WANG Zhengzhi, WANG Guangyun, LI Gangguo

Institute of Automation, National University of Defense Technology, Changsha 410073, China

This work was supported by grants from The National Nature Science Foundation of China (60835005) and The Scientific Research Program of National University of Defense Technology (JC09-03-04)

Received: Mar 22, 2010 **Accepted:** May 11, 2010

Corresponding author: WANG Xiaomin, Tel: +86(731)84574991, E-mail: wangxiaomin2006@gmail.com

Abstract: A living cell can carry out complex biological functions depending on its transcriptional regulatory modules. Therefore, identifying transcriptional regulatory modules is very important for understanding cell function and its transcription mechanism. Integrating gene expression profiles and ChIP-chip data, a transcriptional regulatory module identifying algorithm is developed. The algorithm introduces local correlation into ChIP-chip data and obtains the core and loose sets of transcriptional regulatory modules by using two different *P* value thresholds. Then the core and loose sets are distinguished and the genes in them are expanded. Finally the transcriptional regulatory modules are obtained. Some transcriptional regulatory modules with significant biological meanings are obtained from two different yeast gene expression profiles by using this algorithm. The comparison with other algorithms shows that this algorithm can identify not only modules with more genes, but also modules that other algorithms can not identify. These identified transcriptional modules are helpful for further understanding yeast transcription mechanism.

Key Words: Gene expresion data; ChIP-chip data; Transcriptional regulatory module; Regulatory factor; Yeast

DOI: 10.3724/SP.J.1260.2011.00242