PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(9): 996~1005
www.pibb.ac.cn

综合 ChIP-chip 数据、基因敲除数据和表达谱数据重构基因调控网络 *

江丽华 李亦学 刘 琪 **
(上海交通大学生命科学技术学院，上海 200240)

摘要 揭示生物体内在的调控机制是生物信息学的一项重要研究内容．各种高通量生物数据的涌现，为从基因组的尺度上重构基因调控网络提供了可能．由于单数据源仅能提供关于调控关系的片面信息且存在噪声，因此整合多种生物学数据的方法有望得到可靠性较高的调控网络．提出了一种综合 ChIP-chip 数据、knock out (敲除)数据和各种条件下的表达谱数据来推断调控关系的新方法．ChIP-chip 数据和 knock out 数据能分别提供转录因子和目标基因对关系的直接物理结合和功能关系的证据，这两类数据的整合有望获得较高的识别准确率．但这两类数据的重合性通常较低，基于共调控的基因通常具有较高的表达相似性这一假设，在一定程度上降低了这两类数据重合性较低所带来的影响．算法所识别的大部分调控关系都被 YEASTRACT，高质量 ChIP-chip 数据和文献所验证，从而证明了该方法在调控关系的预测上具有较高的准确性．与其他方法的比较，也表明了该方法具有较高的预测性能．

关键词 基因调控网络，ChIP-chip 数据，Knock out(敲除)数据，基因表达谱数据
学科分类号 310.6199
DOI: 10.3724/SP.J.1206.2010.00184

基因调控网络是细胞协同各基因的功能来应对各种内部和外部刺激的重要途径．通过构建基因调控网络，可以解读生物组织内部基因及其产物的生成过程和复杂的调控关系，即在系统尺度上理解生物学进程．

基因网络的研究源于 20 世纪 60 年代，Kauffman [1]通过简单的逻辑规则对基因网络动力学进行了研究．但由于当时实验条件和科技水平有限，研究发展缓慢．近年来，几大领域的发展为大规模研究基因调控网络提供了契机．一是越来越多的生物全基因组序列的测定，以及高通量的实验技术如基因芯片，染色质免疫沉淀(chromatin immunoprecipitation，ChIP)与基因芯片相结合的 ChIP-on-chip 技术的发展，为基因调控的研究提供了大量的数据．二是计算机技术的飞速发展，使得大规模分析这些实验数据、全基因组范围内的扫描和预测成为可能．

目前国内外在基因调控网络的预测方面已取得了一些成果[2-4]．Friedman 等[2]最早提出将贝叶斯网络运用到调控关系的重构中．Segal 等[3]根据表达谱的相似性将基因分成不同的模块，并根据转录因子(transcription factor，TF)与整个模块中基因表达的相关关系得到调控网络.Blanchette 等[4]运用支持向量机基于基因表达谱数据来预测调控关系．基于 ChIP-chip 数据，研究人员对酵母、人和小鼠中某些转录因子可能的目标基因进行了搜索．

这些基于单数据源的方法虽然取得了一定的成功，但由于单数据源通常只能提供关于调控关系的片面信息且实验本身存在的噪声，使得基于单数据源方法的性能受到限制．因此整合多种类型数据的方法有望提高重构网络的可靠性和覆盖率，从而成为发展基因调控网络重构技术的必然趋势．人们在这一方面已经有了一些初步尝试，并取得了一些成果[5-8]．基本思路都是从多方面，而不是单方面来

* 国家重点基础研究发展计划(973)(2009CB918404, 2006CB910700), 国家高技术研究发展计划(863) (2007AA02Z329), 国际合作项目(2007DFA31040)和国家自然科学基金(30700154, 31070746)资助项目.
** 通讯联系人.
Tel: 021-34204348, E-mail: liuqi@sjtu.edu.cn
收稿日期：2010-04-07，接受日期：2010-05-31

寻找潜在调控关系的证据，从而提高重构的可靠性．GRAM 基于表达谱相似性找到那些 ChIP-chip 数据中采用严格阈值而漏掉的靶基因[5]．ReMoDiscovery 结合 ChIP-chip 数据、表达谱数据和转录因子结合模体三种数据来构建调控网络[6]．基于基因表达谱、ChIP-chip 和转录因子序列数据，COGRIM 模型综合了贝叶斯分级模型和吉布斯样本方法[7]．最近，Gilchrist 等[8]用 ChIP-chip 数据和 ChIP-seq 对基因转录调控进行了研究．

在综合多种数据重构调控网络时，最关键的问题是综合何种类型的数据以及如何综合．一般说来，综合那些独立的，能提供互补信息的数据有利于构建准确度较高的调控网络．本文提出了一种综合 ChIP-chip 数据、转录因子基因敲除(knock out)后的表达谱数据和各种条件下的表达谱数据来推断调控关系的新方法．ChIP-chip 数据和 knock out 数据能提供转录因子和目标基因对关系的直接物理结合和功能相关的证据，但这两类互补数据的重合率较低，为了提高重合率且不带来较多的假阳性结果，我们在以前的工作中通过寻找这两类数据最显著的交集来实现这两类数据的最优综合[9]，但这两类数据的交集中也可能包含一些随机出现的基因．为此，我们综合表达谱数据作为揭示调控关系的间接证据，基于共调控的基因通常具有较高相似性这一假设，对这两类数据的交集进行扩充和筛选，此外随机模拟背景噪声的方法，也进一步提高了调控关系构建的可靠性．

1　材料和方法

1.1　数据源

ChIP-chip 数据来源于 Harbison 等的实验结果，它提供了 203 个转录因子在 6 229 个基因启动子区的结合位点信息．Knock out 数据是 Hu 等提供的转录因子敲除后的芯片数据，它记录了 269 个转录因子敲除后的 6 429 个基因的表达谱．第三种数据是近百种实验条件下的基因表达谱数据．所有数据文献来源见网络版附录Ⅰ (http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20100184&flag=1).

1.2　方法

本方法的框图见图 1，主要分为以下 3 个步骤：

a．预处理．预处理中需对表达谱数据进行归一化，并保证这 3 种数据包含相同的基因．对 ChIP-chip 数据和转录因子 knock out 数据，只保留它们共同的转录因子．

b．确定核心模块．(1)对于基因 i，在严格阈值 t_1 下，找出既满足 ChIP-chip 数据结合显著性(即衡量转录因子和基因 i 结合显著程度的 P-value 小于或等于阈值 t_1)又满足转录因子 knock out 数

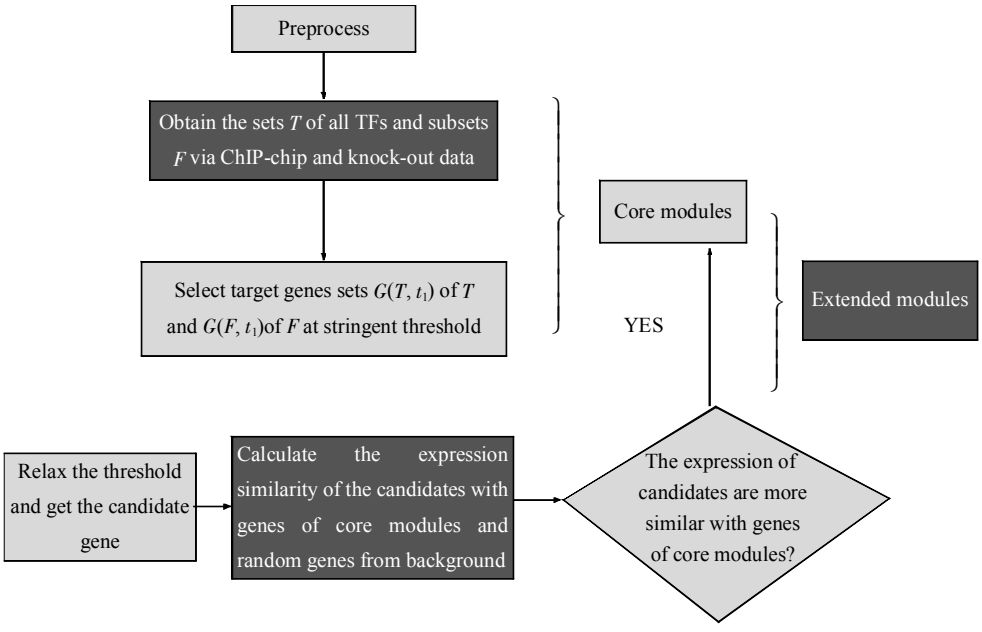

Fig. 1　Schema of our method

据表达差异显著性(即衡量转录因子敲除后，基因 i 差异表达显著性的 P-value 小于或等于阈值 t_1)的所有转录因子组成的全集 $T(i, t_1)$，这里严格阈值 t_1=0.01；(2)找出集合 $T(i, t_1)$的所有子集合 $F(i, t_1) \subseteq T(i, t_1)$，$F(i, t_1)$指在阈值 t_1 下，与基因 i 同时满足结合显著性和差异表达显著性的所有转录因子全集的子集，通过列举全集的子集获得；(3)遍历所有的基因，得到转录因子组成的全集 T 和子集 F，$T=\bigcup_{i=1}^{n} T(i, t_1)$，$F=\bigcup_{i=1}^{n} F(i, t_1)$，$n$ 为总基因数；(4)寻找集合 T，F 的所有目标基因集合，用 $G(T, t_1)$和 $G(F, t_1)$表示；(4)如果目标基因集合的元素个数大于或等于 3，那么转录因子构成的集合 T 或 F 与对应的目标基因构成的集合 $G(T, t_1)$或 $G(F, t_1)$就构成一个核心模块，这样算法就遍历了所有可能结合的基因核心模块．该核心模块包含了可靠性较高的转录调控关系，它为后续的扩展奠定了基础．

c．对核心模块进行扩充和筛选得到扩展模块．确定核心模块后，算法放宽阈值要求即 t_2=0.05，并回访 ChIP-chip 数据和转录因子 knock out 数据．(1)对每个核心模块，算法遍历所有基因，寻找不在核心模块中，但满足放宽的阈值要求的“候补”基因；(2)基于共调控基因具有较高表达相似性的假设，利用表达谱数据求出候补基因与可选核心模块中所有目标基因的皮尔森相关系数(计算皮尔森相关系数的绝对值，因为调控关系可能是正相关也可能是负相关)，然后计算“候补”基因与随机的 3 个基因之间的皮尔森相关系数 100 次，如果“候补”基因与可添加的核心模块中基因的皮尔森相关系数至少有 90 次都比随机相关系数大，则将此“候补”基因加入核心模块得到扩展模块．

总的来说，算法首先从 ChIP-chip 数据和 knock out 数据的交集中通过严格阈值筛选得到核心模块，然后合理地放宽阈值，通过综合近百种实验条件下的微阵列表达谱数据来向核心模块加入候补基因，并通过随机模拟的方法，进一步提高调控关系构建的可靠性．为了减少运算量，算法从包含最多数目的转录因子集合开始，到包含最小数目的子集结束，从而完成模块的初始化与扩展．需要注意的是，如果一个基因已经包含在集合 T 的模块 $M(T)$中，则 T 的子集 F 的模块 $M(F)$不再添加此基因，因为被几个转录因子共同调控的基因，必然被这几个转录因子组成的子集调控．

2　结　　果

算法共识别出 679 对调控关系，涉及到 46 个转录因子和 529 个基因(如图 2，所有的调控关系对见见网络版附录Ⅱ，http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20100184&flag=1)．从图 2 中可以看到一些转录因子，比如

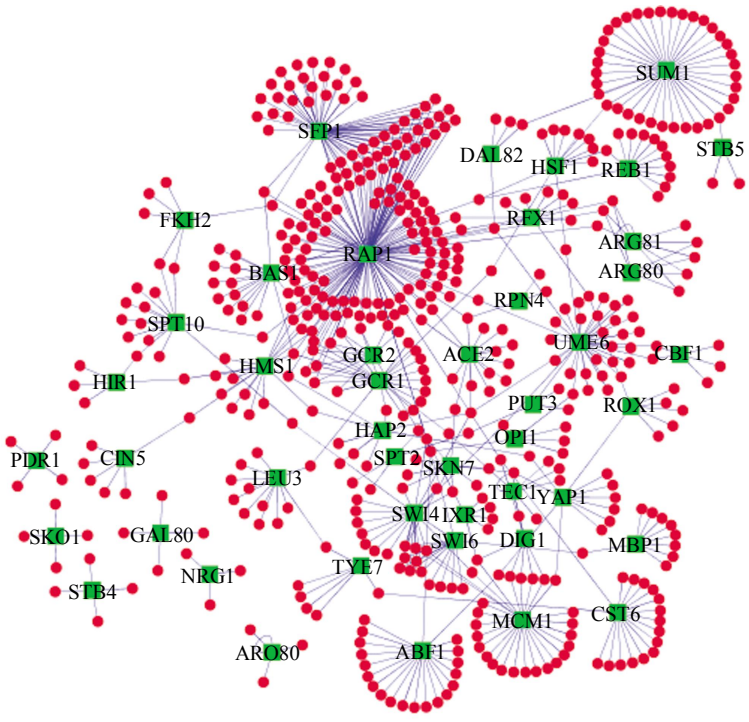

Fig. 2　Visualization of the regulatory networks inferred by our method

TFs are represented by green circles and their target genes by red circles.

GCR1 与 GCR2，ARG80 与 ARG81，SWI4 与 SWI6 等都共同调控一些基因，它们都是已经确定的具有协同调控作用的转录因 子对.

为了评估预测结果的准确性，我们将预测结果与 YEASTRACT 数据库、高质量的 ChIP-chip 实验数据，以及其他的实验和计算文献进行比较，结果表明，大部分调控关系都有相关的证据支持，这在一定程度上证明了该方法在预测调控关系上具有较高的准确性.

2.1　与 YEASTRACT 数据库的分析比较

首先我们通过 YEASTRACT 数据库对方法的预测性能进行评估．YEASTRACT 是分析啤酒酵母转录调控关系的重要数据库，收集了转录因子和目标基因的大部分调控关系，每一对关系至少有一个实验证据的支持[10]．46 个转录因子预测到的目标基因数及目标基因中被 YEASTRACT 数据库验证的百分比(图 3).

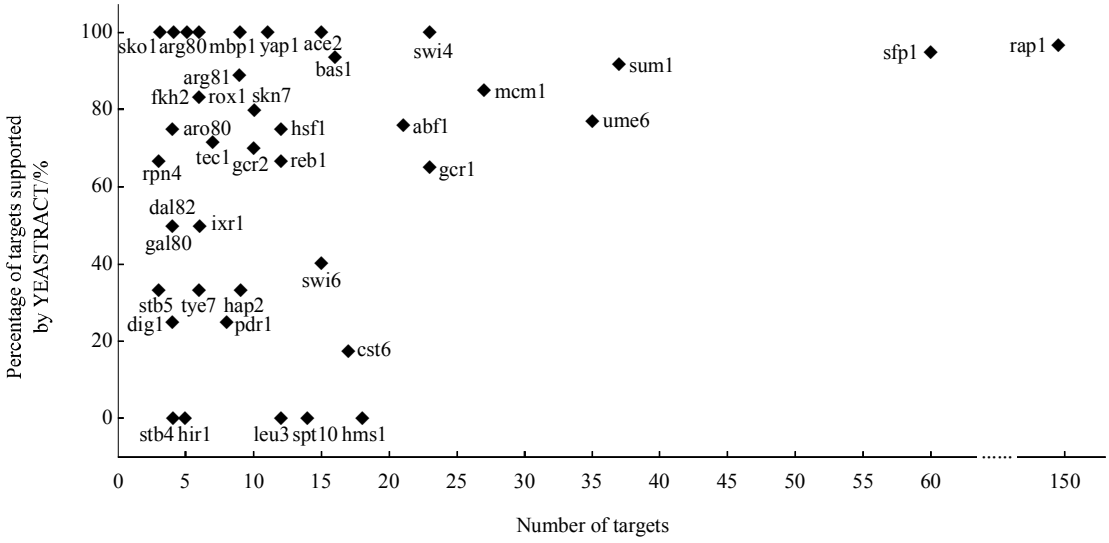

Fig. 3　Comparison with YEASTRACT

We compared all the predicted results with YEASTRACT, the x axis denotes the number of targets predicted by our algorithm, the y axis represents the percentage of targets supported by YEASTRACT.

由图 3 可知，预测的目标基因的支持百分比大多处于 70%以上．特别是 SKO1，SWI4，ACE2，YAP1，ARG80，MBP1 等转录因子，所预测的目标基因全部被 YEASTRACT 所验证．对于目标基因最多的转录因子 SFP1(60 个)，RAP1(149 个)其支持率也在 95%以上；对于那些在 YEASTRACT 支持率较低的转录因子，我们发现，这主要是由于 YEASTRACT 记录不全造成的，这些支持率较低的转录因子，其目标基因通常也能被其他的文献和计算证据所验证．部分验证结果如下.

2.1.1　实验证据.

LEU3 调控参与氨基酸合成所需酶的转录表达．算法预测的 12 个目标基因均未得到 YEASTRACT 的验证(表 1)．而 LEU4 与 ILV5 是已经得到公认的 LEU3 的目标基因，它们属于氨基酸合成通路的基本要素．此外，目标基因 LEU1, ILV2，BAT1，OAC1，ILV3 都被报道受 LEU3 调控．BAP2，ALD5 和 ISU2，也被 Nielsen 等[11]通过综合 ChIP-chip 数据和基因表达谱的计算方法得以证明．因此, 除 2 个功能未知的目标基因 YAL009C 和 YMR046C 外，其他的 10 个基因与 LEU3 的调控关系都有计算和实验文献的支持.

HMS1 共预测出 18 个目标基因(表 1)，在 YEASTRACT 中均没有相关记录．其中 13 个基因 RPL23A，RPL21A，RPS29B，RPP1A，RPL35B，RPL41A，RPS13，RPS25A，RPS28B，RPS22B，RPS19B，RPS19A，RPS12 都编码核糖体蛋白，4 个基因 BUD19，LOC1，SSS1，YSY6 都与核糖体蛋白相关，而研究人员发现 HMS1 是调控核糖体 DNA 转录的重要因子[12]，由此可以推断这 17 个基因很有可能受 HMS1 调控.

SPT10 和 SPT21 是组蛋白转录所必需的转录

因子[13]. 算法预测的 SPT10 的目标基因中 HHF1，HHF2，HHT1，HHT2，HTA1，HTA2，HTB2 都是已确定的核心组蛋白(表 1)，由此可推断它们很有可能是 SPT10 的目标基因.

Table 1 Listed TFs and their targets predicted by our method

TF	ORF
LEU3 (0%)	BAP2 BAT1 OAC1 ILV2 ILV3 ILV5 LEU1 LEU4 ALD5 ISU2 YAL009C YMR046C
HMS1(0%)	BUD19 LOC1 PGA2 RPL21a RPL23a RPL35b RPL41a RPP1a RPS12 RPS13 RPS19a RPS19b RPS22b RPS25a RPS28b RPS29b SSS1 YSY6
SPT10(0%)	CKA1 HHF1 HHF2 HHT1 HHT2 HTB2 HTA1 HTA2 RLI1 RPL34b MNI1 TIF11 YNL010w ARO7
GAL80(50%)	GAL1 GAL7 GAL10 FUR4

* The percentage in bracket denotes the percent supported by YEASTRACT.

转录因子 GAL80 对 GAL 基因 GAL1, GAL10, GAL7 具有转录调控作用，这在文献中早有报道. FUR4 转录水平增加依赖 Gal4, 而 GAL4 和 GAL80 的相互作用通常决定 GAL 基因的转录，由此可以推断 FUR4 可能受 GAL4 和 GAL80 的调控.

2.1.2 其他计算证据. 算法预测的 SWI6 的 15 个目标基因只有 6 个被 YEASTRACT 所验证，但分别有 7 个和 6 个目标基因被 MacIsaac 和 Pham 方法所识别(表 2)[14-15]. MacIsaac 等根据进化保守性的原理开发出基因序列模式搜索算法，通过结合 PhyloCon 和 Converge 的方法反复精炼 ChIP-chip 信息构建酵母基因调控网络. Pham 的研究小组基于规则演化的原理，综合三种不同的微阵列表达谱数据和 ChIP-chip 数据预测调控关系. 这些计算方法综合了不同的数据，使用了不同的模型而得到了相同的结论，这也在一定程度上证明了该调控关系具有较大的可能性. 此外，Cherepinsky 等[16]通过聚类推测 CWP1 是 SWI6 的目标基因，AGA1，AXL2, YOR248W 在其他几种算法得到了验证[17-19]. 这样除 PLB3，YMR144W 外其他的 13 个基因都有其他证据的支持.

Table 2 List of SWI6's targets and computational evidence

TF	ORF	YEASTRACT	MacIsaac *et al*.[14]	Pham *et al*.[15]	Other computational evidence	All evidence
SWI6	CHA1			*		*
SWI6	AGA1				*	*
SWI6	AXL2				*	*
SWI6	PUP3	*	*	*		*
SWI6	SWI4	*	*			*
SWI6	TOS6	*				*
SWI6	PLB3					
SWI6	CWP1				*	*
SWI6	CWP2	*	*			*
SWI6	EXG1		*	*		*
SWI6	YOX1	*	*	*		*
SWI6	YMR144W					
SWI6	SKM1		*	*		*
SWI6	SRL1	*	*	*		*
SWI6	YOR248W				*	*

DIG1 只有 2 个目标基因 GPA1 和 SST2 被 YEASTRACT 验证，但未在 YEASTRACT 中得到验证的目标基因中分别有 5 个和 4 个与 MacIsaac 和 Pham 方法所得到的结果重合(表 3). 仅剩的 FUS1 基因报道受繁殖信息素的调控，而转录因子 DIG1 调控繁殖相关基因的表达[20]，所以我们推测 FUS1 很有可能被 DIG1 调控.

Table 3　List of DIG1's targets and computational evidence

TF	ORF	YEASTRACT	MacIsaac *et al.*[14]	Pham *et al.*[15]	All evidence
DIG1	AGA2		*	*	*
DIG1	BAR1		*	*	*
DIG1	FUS1				
DIG1	GPA1	*			*
DIG1	MFA1		*	*	*
DIG1	SST2	*			*
DIG1	STE2		*	*	*
DIG1	TEC1		*		*

2.2　与高质量 ChIP-chip 数据对比

Hu 等[21]认为低质量 ChIP-chip 数据是造成 ChIP-chip 和 knock out 实验数据重合率低的一个主要原因．他们发现若采用高质量的 ChIP-chip 数据，重合度会有较大的改善．而我们的结果表明，即使采用低质量的数据，方法也能获得较高的识别准确度．我们得到的 RAP1 目标基因与高质量的 ChIP-chip 数据[22]比较结果见图 4.

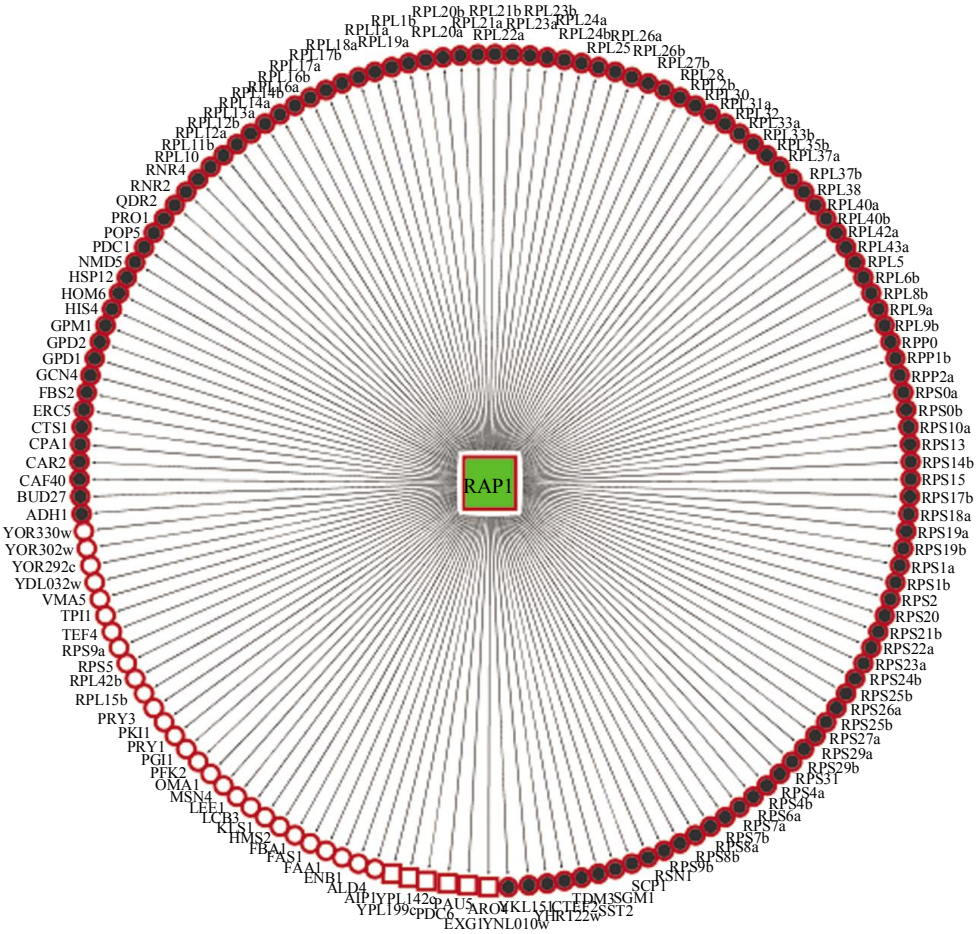

Fig. 4　Comparison of results of RAP1 with high quality ChIP-chip data and YEASTRACT

Circle nodes are for genes with relations supported by YEASTRACT, otherwise rectangular nodes. Solid nodes are for the genes with relations supported by high-quality ChIP-chip data.

由图 4 可知，我们预测出 RAP1 的 149 个目标基因中有 143 个在 YEASTRACT 数据库中有记录，115 个和高质量 ChIP-chip 数据吻合．若采用低质量的 ChIP-chip 数据共筛选出 169 个目标基因中，其中只有 111 个与高质量 ChIP-chip 数据一致，而若采用低质量的 ChIP-chip 数据和 knock out 数据重合的结果，仅筛选出 70 个基因，其中 62 个与高质量的 ChIP-chip 数据一致．由此可以得出，综合 ChIP-chip 数据和 knock out 数据并采用表达谱数据的相似性进行进一步筛选的方法是可靠的．

2.3 预测结果的 GO 功能富集水平分析

转录因子和其目标基因之间功能的相似性可以为评估调控关系的准确性提供间接的依据．我们采用 SGD 数据库提供的 GO 分析软件 GO Term Finder 分析得到的转录因子与预测目标基因之间是否有显著的生物相关性，部分结果见表 4．GO 功能富集分析表明目标基因集合的功能与相应的转录因子的功能通常是相关的．这从另一方面说明算法符合生物学意义，预测结果在生物学上有利用价值．

Table 4 List of some enriched Gene Ontology(GO) annotations

Regulators	Functional description[1)]	GO annotations[2)]	*P* value
GAL80	Repression of GAL genes	(3/4) Galactose catabolic process	$P=1.17\times10^{-9}$
SFP1	Ribosome biogenesis， regulates G2/M transitions	(55/60) Metabolic process	$P=4.93\times10^{-10}$
RAP1	Transcription	(126/149) Translation	$P=9.77\times10^{-36}$
GCR1	Glycolysis	(13/23) Glucose catabolic process	$P=1.46\times10^{-19}$
GCR2	Glycolysis	(8/10) Glucose catabolic process	$P=9.93\times10^{-16}$
TYE7	May function as a transcriptional activator in Ty1-mediated gene expression	(3/6) Ty element transposition	$P=3.9\times10^{-4}$
DIG1	Inhibits pheromone responsive transcription	(7/8) Response to pheromone	$P=2.02\times10^{-11}$
ARG80	Regulation of arginine-responsive genes	(3/6) Arginine biosynthetic process	$P=2.03\times10^{-7}$
ARG81	Regulation of arginine-responsive genes	(6/9) Arginine metabolic process	$P=1.14\times10^{-13}$
SUM1	Transcriptional repressor required for mitotic repression of middle sporulation-specific genes; general replication initiation factor; telomere maintenance	(16/37) Sporulation resulting in formation of a cellular spore	$P=6.81\times10^{-17}$
OPI1	Phosphorylation transcriptional regulator	(3/5) Cellular lipid metabolic process	$P=7.01\times10^{-3}$
STB5	Regulate multidrug resistance and oxidative stress response forms a heterodimer	(2/3) Pentose-phosphate shunt, oxidative branch	$P=5.14\times10^{-5}$
BAS1	The purine and histidine biosynthesis	(9/16) Heterocycle metabolic process	$P=2.72\times10^{-9}$
YAP1	Regulation of meiotic recombination	(7/11) Response to stress	$P=0.0002$
MBP1	Transcription factor involved in regulation of cell cycle progression from G1 to S phase	(5/9) Cell cycle process	$P=2.87\times10^{-3}$

1) Functional description of regulators is from the Saccharomyces Genome Database

2) Gene Ontology analysis was done using GO Term Finder in SGD in Aug 31, 2008; 5952 genes were included in the background set with *P*-value cut-off < 0.01.

值得一提的是 GO 功能只是目标基因集合的功能类做分析，有的基因虽然没有被归为一类，但是也是具有生物相关性的，比如转录因子 GCR1 预测的 23 个目标基因中虽然只有 13 个目标基因是归为一类的，但 23 个靶基因均被 YEASTRACT 数据库验证．

2.4 与其他算法的比较

此外，我们挑选了几组独立的数据作为评估的正集，将我们的方法与其他计算方法进行了比较，其中包括 ReMoDiscovery[6]，GRAM[5]，SVMs[4]，Pham 等[15]的方法和 COGRIM[7]等．ReMoDiscovery 结合 ChIP-chip 数据、表达谱数据和转录因子结合模体三种数据来构建调控网络[6]．GRAM 算法先在严格阈值下从 ChIP-chip 数据中得到与转录因子结合的核心基因模块，然后再通过表达谱数据来扩充核心模块[5]；SVMs 方法基于基因表达谱数据，运用支持向量机方法来预测调控关系[4]．Pham 等[15]的方法基于规则演化的原理，综合三种不同的微阵列表达谱数据和 ChIP-chip 数据来预测调控关系．GOGRIM 是基于贝叶斯分层和吉布森抽样法，综

合 ChIP-chip 数据、表达谱数据和序列数据来进行转录因子的目标基因预测[7]. 评估的正集采用独立于本方法的实验和文献数据，其中 RAP1 的数据是 Lieb 等[22]提供的高质量的 ChIP-chip 数据，SFP1 的数据是 Marion 等[23]提供的目标基因，GCR1 的数据是 Lopez 等[24]的研究结果，SWI4 的数据是 Iyer 等[25]的 ChIP-chip 实验结果，ACE2 的数据来源于 Workman 等[26]的 ChIP-chip 实验，LEU3 则是直接从文献中提取的目标基因[27]. 基于这些独立的数据源，可以比较它们与各预测结果的 Jaccard 相似性分数 TP/(TP+FP+FN)，其中 TP 代表真阳性，FP 代表假阳性而 FN 则表示假阴性[6]. Jaccard 相似性分数越高，则表明方法的预测性能越好. 各方法的比较结果见表 5. 从表 5 中可以看出，我们的方法在各个数据集上都获得了较高的 Jaccard 相似性分数值，由此也从一定程度上证实了我们的方法在预测调控关系上具有较好的性能，同时也表明了提供调控关系功能证据 knock out 数据的加入有利于提高算法的准确性.

Table 5　Jaccard similarity score (JSS) of different algorithms

TFs	Algorithm					
	ReMoDiscovery[6]	GRAM[5]	SVMs[4]	Pham *et al.*[15]	COGRIM[7]	Our algorithm
RAP1	0.221 782	0.260 204	0.036 53	0.263 291	0.313 131	0.285 36
SFP1	0.229 73	0.126 761	–	0.043 165	–	0.297 297
GCR1	–	0.072 727	0.032 895	0.071 429	0.05	0.134 328
SWI4	0.147 059	0	0.020 08	0.181 818	0.296	0.180 808
ACE2	–	0.142 857	–	0.058 824	0.093 75	0.148 936
LEU3	–	0.222 222	–	0.190 476	0.064 516	0.307 692

"–" Denotes the TF is not included in the algorithm.

3　讨　　论

由于 ChIP-chip 实验数据，knock out 数据或者基因表达谱数据在揭示基因调控关系上都具有不完整性、不确定性和一定的互补性，因此本算法试图通过整合三种实验数据来预测调控关系，从而克服了单数据源的限制，有利于提高重构网络的可靠性和覆盖率.

本算法不像传统方法设定一个严格的阈值进行筛选，导致较高假阴性率的预测结果，也没有不设限制，或随意设定宽松的阈值进行筛选，导致较高假阳性的预测结果，而是通过综合 ChIP-chip 数据、knock out 数据以及表达谱数据来预测调控关系. 在综合的方式上我们首先对两种数据进行相对严格阈值的设定，保证核心转录调控关系的准确性，因为这是后续添加“候补”基因的基础，然后放宽阈值并结合表达谱数据，利用皮尔森相关系数对其做了进一步扩充和精炼，把由阈值带来的误差尽可能地缩小，大大降低了假阴性率. 在最后的扩充中，还加入了背景模块的判断. 从算法最后结果来看，我们总共预测出 46 个转录因子的 679 对调控关系，其中被 YEASTRACT 数据库验证的调控关系为 513 对，支持率达到 75.6%. 在未被 YEASTRACT 数据库验证的目标基因中，有一些经验证是潜在的目标基因. 另外通过查询文献、算法或其他实验结果，对很多调控关系进行了不同程度的验证. 最后通过与其他算法的比较，也表明我们的方法具有较高的可靠性. 算法除了预测调控关系外，还能发现一些协同作用的转录因子对，比如 GCR1 和 GCR2，SWI4 和 SWI6 等.

本算法具有普遍适用性，不仅可用于预测酵母的调控关系，还可以拓展到小鼠甚至人等各种生物，不仅可运用到调控关系的重构方面，而且为多种异源数据的整合提供了思路. 若能加入更多的基因表达谱芯片，采用高质量的 ChIP-chip 数据并考虑到转录因子 knock out 后的一些补偿效应，或者综合其他的组学数据，算法的准确性可能得到进一步提高.

参　考　文　献

[1] Kauffman S. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theoretical Biology, 1969, **22**(3): 437-467

[2] Friedman N, Linial M, Nachman I, *et al*. Using Bayesian networks to analyze expression data. J Computational Biology, 2000, **7**(3-4): 601-620

[3] Segal E, Shapira M, Regev A, *et al*. Module networks: Discovering regulatory modules and their condition specific regulators from gene expression data. Nature Genetics, 2003, **34**(2): 166-176

[4] Blanchette M, Bataille A, Chen X, *et al*. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Research, 2006, **16**(5): 656-668

[5] Bar-Joseph Z, Gerber G, Lee T, *et al*. Computational discovery of gene modules and regulatory networks. Nature Biotechnology, 2003, **21**(11): 1337-1342

[6] Lemmens K, Dhollander T, De Bie T, *et al*. Inferring transcriptional modules from ChIP-chip, motif and microarray data. Genome Biology, 2006, **7**(5): R37

[7] Chen G, Jensen S, Stoeckert Jr C. Clustering of genes into regulons using integrated modeling-COGRIM. Genome Biology, 2007, **8**(1): R4

[8] Gilchrist D, Fargo D, Adelman K. Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation. Methods, 2009, **48**(4): 398-408

[9] Cheng H, Jiang L H, Liu Q, *et al*. Inferring transcriptional interactions by the optimal integration of ChIP-chip and Knock-out data. Bioinformatics and Biology Insights, 2009, **3**: 129-140

[10] 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**(Database issue): D446-D451

[11] Nielsen P S, van den Hazel B, Didion T, *et al*. Transcriptional regulation of the *Saccharomyces cerevisiae* amino acid permease gene BAP2. Mol Gen Genet, 2001, **264**(5): 613-622

[12] Hontz R, Niederer R, Johnson J, *et al*. Genetic identification of factors that modulate ribosomal DNA transcription in *Saccharomyces cerevisiae*. Genetics, 2009, **182**(1): 105-119

[13] Eriksson P, Mendiratta G, McLaughlin N, *et al*. Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements. Molecular and Cellular Biology, 2005, **25**(20): 9127-9137

[14] MacIsaac K D, Wang T, Gordon D B, *et al*. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. BMC Bioinformatics, 2006, **7**: 113

[15] Pham T, Clemente J, Satou K, *et al*. Computational discovery of transcriptional regulatory rules. Bioinformatics-Oxford, 2005, **21**(2): 101-107

[16] Cherepinsky V, Feng J, Rejali M, *et al*. Shrinkage-based similarity metric for cluster analysis of microarra y data. Proc Natl Acad Sci USA, 2003, **100**(17): 9668-9673

[17] Tom M, De Smet Riet J, Van de Peer Yves M. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. BMC Systems Biology, 2009, **3**: 49

[18] Zhang M. Promoter analysis of co-regulated genes in the yeast genome. Computers and Chemistry, 1999, **23**(3-4): 233-250

[19] Zou M, Conzen S D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics, 2005, **21**(1): 71-79

[20] Cook J, Bardwell L, Kron S, *et al*. Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast *Saccharomyces cerevisiae*. Genes & Development, 1996, **10** (22): 2831-2848

[21] Hu Z, Killion P, Iyer V. Genetic reconstruction of a functional transcriptional regulatory network. Nature Genetics, 2007, **39** (5): 683-687

[22] Lieb J, Liu X, Botstein D, *et al*. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature Genetics, 2001, **28**(4): 327-334

[23] Marion R, Regev A, Segal E, *et al*. Sfp1 is a stress-and nutrient-sensitive regulator of ribosomal protein gene expression. Proc Natl Acad Sci USA, 2004, **101**(40): 14315-14322

[24] Lopez M, Baker H. Understanding the growth phenotype of the yeast gcr1 mutant in terms of global genomic expression patterns. J Bacteriology, 2000, **182**(17): 4970-4978

[25] Iyer V, Horak C, Scafe C, *et al*. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature, 2001, **409**(6819): 533-538

[26] Workman C, Mak H, McCuine S, *et al*. A systems approach to mapping DNA damage response pathways. Science, 2006, **312**(5776): 1054-1059

[27] Friden P, Schimmel P. LEU3 of *Saccharomyces cerevisiae* activates multiple genes for branched-chain amino acid biosynthesis by binding to a common decanucleotide core sequence. Molecular and Cellular Biology, 1988, **8**(7): 2690-2697

Reconstruction of Gene Regulatory Networks by Integrating ChIP-chip, Knock out and Expression Data*

JIANG Li-Hua, LI Yi-Xue, LIU Qi**

(*School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai* 200240,China)

Abstract Uncovering the underlying regulatory mechanism has become a major research in bioinformatics studies. The availability of various kinds of high-throughput biological data makes the reconstruction of regulatory networks on a genomic scale possible. Since each single data source provides only partial and noisy information of the regulatory relationships, methods combining diverse data sources are expected to get more reliable networks. Here a method was presented to infer the regulatory networks by combining ChIP-chip, TF (transcription factor) knock out and expression data. Since ChIP-chip and TF knock out data provide direct physical binding and functional evidences of relations between TF and target genes, combining these two data is expected to obtain high prediction accuracy. However， the overlap of these two data is low. Based on the assumption that co-regulated genes often have high expression similarity, the method reduced the effect of the low overlap of these two data to some extent. The results show that most inferred regulatory relations are validated by YEASTRACT, high quality ChIP-chip data and literatures, which demonstrate our method is powerful and reliable. Moreover, the comparison between our method and others also shows that it has better performance.

Key words gene regulatory networks，ChIP-chip，TF knock out expression data

DOI: 10.3724/SP.J.1206.2010.00184

---

*This work was supported by grants from National Basic Research Program of China (2009CB918404, 2006CB910700), Hi-Tech Research and Development Program of China (2007AA02Z329), International S&T Cooperation Program of China (2007DFA31040) and The National Natural Science Foundation of China (30700154, 31070746).

**Corresponding author.

Tel: 86-21-342043348, E-mail: liuqi@sjtu.edu.cn

Received: April 7, 2010 Accepted: May 31, 2010

附　录

Ⅰ算法所用数据的文献来源

ChIP-chip 数据源

[1] Harbison C T, Gordon D B, Lee T I, *et al*. Transcriptional regulatory code of a eukaryotic genome. Nature, 2004, **431**(7004): 99-104

Knock-out 数据源

[2] Hu Z, Killion P J, Iyer V R. Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet, 2007, **39**(5): 683-687

多种条件下表达谱数据源

[3] Casagrande R, Stern P, Diehn M, *et al*. Degradation of proteins from the ER of S. cerevisiae requires an intact unfolded protein response pathway. Mol Cell, 2000, **5**(4): 729-735

[4] Gasch A P, 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

[5] Gross J W, Hegeman A D, Vestling M M, *et al*. Characterization of enzymatic processes by rapid mix-quench mass spectrometry: the case of dTDP-glucose 4,6-dehydratase. Biochemistry, 2000, **39**(45): 13633-13640

[6] Keller G, Ray E, Brown P O, *et al*. Haa1, a protein homologous to the copper-regulated transcription factor Ace1, is a novel transcriptional activator. J Biol Chem, 2001, **276** (42): 38697-38702

[7] Ogawa N, DeRisi J, Brown P O. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. Mol Biol Cell, 2000, **11**(12): 4309-4321

[8] Protchenko O, Ferea T, Rashford J, *et al*. Three cell wall mannoproteins facilitate the uptake of iron in *Saccharomyces cerevisiae*. J Biol Chem, 2001, **276**(52): 49244-49250

[9] Rutherford J C, Jaron S, Ray E, *et al*. A second iron-regulatory system in yeast independent of Aft1p. Proc Natl Acad Sci USA, 2001, **98**(25): 14322-14327

[10] 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

[11] Zhu G, Spellman PT, Volpe T, *et al*. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. Nature, 2000, **406**(6791): 90-94

Ⅱ所有的调控关系对

TF	documented percent	the number of target gene	target genes
HIR1	0%	5	ADK1, HTB2, SLP1,YNR018w,YSY6
GAL80	50%	4	GAL7,GAL10*,GAL1*,FUR4
SFP1	95%	60	CDC60*,CNS1*,CYB2*,DIA1*,ERB1*,HPT1*,NOP58*,PPT1*,RPB5*,RPL12a*, RPL13a*,RPL14a*,RPL16a*,RPL16b*,RPL17a*,RPL19b*,RPL1b*,RPL20a*,RPL20b*, RPL22a*,RPL23b*,RPL24a*,RPL24b*,RPL25,RPL27b*,RPL28*,RPL2b*,RPL30*, RPL31a*,RPL33a*,RPL35a*,RPL40a*,RPL42b*,RPL6b*,RPL8a*,RPL8b*,RPP0*, RPP2a*,RPS11b*,RPS15*,RPS16b*,RPS18a*,RPS1b*,RPS23a*,RPS26a*,RPS27b*, RPS29a*,RPS31*,RPS4a*,RPS5*,RPS6b*,RPS7b*,RPS9a*,SDH4*,UTP6*,YBL028c*, YER079w,YHR020w*,YLR076c,YOR309c*
STB4	0%	4	IMP2,PCS60,YIL166c,YMR034c
RAP1	96%	149	ADH1*,AIP1*,ALD4*,ARO4,BUD27*,CAF40*,CAR2*,CPA1*,CTS1*,ENB1*,ERG5*, EXG1,FAA1*,FAS1*,FBA1*,FRS2*,GCN4*,GPD1*,GPD2*,GPM1*,HIS4*,HMS2*, HOM6*,HSP12*,KES1*,LCB3*,LEE1*,MSN4*,NMD5*,OMA1*,PAU6,PDC1*,PDC6, PFK2*,PGI1*,POP5*,PRO1*,PRY1*,PRY3*,QDR2*,RKI1*,RNR2*,RNR4*,RPL10*, RPL11b*,RPL12a*,RPL12b*,RPL13a*,RPL14a*,RPL14b*,RPL15b*,RPL16a*,RPL16b*, RPL17a*,RPL17b*,RPL18a*,RPL19a*,RPL1a*,RPL1b*,RPL20a*,RPL20b*,RPL21a*, RPL21b*,RPL22a* ,RPL23a*,RPL23b*,RPL24a*,RPL24b*,RPL25*,RPL26a* ,RPL26b*, RPL27b*,RPL28*,RPL2b*,RPL30*,RPL31a*,RPL32*,RPL33a*,RPL33b*,RPL35b*, RPL37a*,RPL37b*,RPL38*,RPL40a*,RPL40b*,RPL42a*,RPL42b*,RPL43a*,RPL5*, RPL6b*,RPL8b*,RPL9a*,RPL9b*,RPP0*,RPP1b*,RPP2a*,RPS0a*,RPS0b*,RPS10a*, RPS13*,RPS14b*,RPS15*,RPS17b*,RPS18a*,RPS19a*,RPS19b*,RPS1a*,RPS1b*, RPS2*,RPS20*,RPS21b*,RPS22a*,RPS23a*,RPS24b*,RPS25b*,RPS26a*,RPS26b*, RPS27a*,RPS29a*,RPS29b*,RPS31*,RPS4a*,RPS4b*,RPS5*,RPS6a*,RPS7a*,RPS7b*, RPS8a*,RPS8b*,RPS9a*,RPS9b*,RSN1*,SCP1*,SGM1*,SST2*,TDH3*,TEF2*,TEF4*, TPI1* ,VMA6*,YDL032w*,YHR122w*,YKL151c*,YNL010w*,YOR292c*,YOR302w*, YOR338w*,YPL142c,YPL199c
SKO1	100%	4	FSH1*,HXT11*,SOL1*,SPI1*
PDR1	25%	4	RPL11a,RPS24a*,YFL065c,YHR138c
SPT10	0%	14	CKA1,HHF2,HHT2,HTB2,HTA2,HHF1,HHT1,RLI1,HTA1,RPL34b,MNI1,TIF11, YNL010w,ARO7
CIN5	83%	6	ACO1*,ALD6*,RPS12,VPS25*,YLL007c*,ZEO1*
GCR1	65%	23	AAT2 ADH1*,BMH1*,CDC19* ,ENO2*,FBA1*,GPM1*,OLA1,PDC1*,PFK1*,PFK2, PGI1*,PGM1,PYC1*,RPS19a*,SAM4,SBP1*,TDH2*,TDH3*,TPI1*,YAL004w, YDR417c,YMR046c
TYE7	33%	6	ADH5*,ERP2*,SDS23,YAR009c,YBR012w-a,YER138c
LEU3	0%	12	ALD5,BAP2,BAT1,ILV2,ILV3,ILV5,ISU2,LEU1,LEU4,OAC1,YAR009c,YMR046c
CST6	18%	17	KAR5*,PCH2,SST2,YBR012w-a,YBR089w,YDL096c*,YFR017c,YJR011c,YKL077w, YLR462w,YLR463c,YLR464w,YLR465c,YRF1-1,YRF1-2*,YRF1-4,YRF1-5
GCR2	70%	10	ADH1*,FBA1*,GPM1*,PGI1*,PHD1,TAL1,TDH2*,TDH3*,TPI1*,YKL153w
TEC1	71%	7	FUS1*,PYC1*,TEC1*,YAR010c*,YBR012w-b*,YPR091c,YRF1-6

To be continued

Continued

TF	documented percent	the number of target gene	target genes
DIG1	25%	8	AGA2,BAR1,FUS1,GPA1*,MFA1,SST2*,STE2,TEC1
BAS1	94%	16	ADE12*,ADE13*,ADE16*,ADE3*,ADE4*,ADE5,7*,GCV1*,GCV2*,GCV3*,HIS1*, HIS4*,HSC82*,MTD1*,RPL22a*,RPS19a,SHM2*
SWI6	40%	15	AGA1,AXL2,CHA1,CWP1,CWP2*,EXG1,PLB3,PUP3*,SKM1,SRL1*,SWI4*,TOS6*, YMR144w,YOR248w,YOX1*
ARO80	75%	4	ARO10*,ARO80*,HMLALPHA1*,KRE11
HAP2	33%	9	COX9*,GPX2*,MIM1,RPS22b,RPS29b,SUN4,TRX1,YDR417c,YHB1*
HMS1	0%	18	BUD19,LOC1,PGA2,RPL21a,RPL23a,RPL35b,RPL41a,RPP1a,RPS12,RPS13,RPS19a, RPS19b,RPS22b,RPS25a,RPS28b,RPS29b,SSS1,YSY6
MCM1	85%	27	AGA1*,AGA2*,BAR1*,CHS2*,CIS3*,CLN3*,FDH1*,GPA1*,GPI17,GRE2,GYP8*, HTZ1*,KIN3*,LSM4*,MFA1*,NFU1*,PIR1*,PIR3*,PLB3*,RPA34*,SFG1*,SIM1*, STE2*,YAL037w,YFL034w,YML053c*,ZDS2*
RPN4	67%	3	CTS1,RPN2*,RPN3*
ARG80	100%	6	ARG1*,ARG3*,ARG5,6*,ARG8*,CPA1*,YOR302w*
ARG81	89%	9	ARG1*,ARG3*,ARG5,6*,ARG8*,BUD27*,CPA1*,YDL152w*,YOR302w*,ARG81*
UME6	77%	35	ACS1*,AGP1*,BCY1*,CAR1*,CAR2*,FET3,GSM1*,ICE2*,INO1*,IRC15*,NDJ1*, PAN6*,PRX1*,PST1*,PUT3,RIM4*,SPO13*,SPO16*,SRT1*,SWI1*,TID3*,TSL1, UBI4*,XBP1,YBR184w* ,YCR061w*,YCR062W,YHB1,YIL091c*,YIL127c,YLR297w*, YNL305c,YOL131w*,YOR338w*,ZIP1*
FKH2	83%	6	DBP2,HHF1*,HHT1*,PES4*,RPS1b*,UTP4*
SUM1	92%	37	BNA1*,BNA2*,BNA5*,CDA1*,CDA2,CRR1*,DAL1*,DAL4*,DTR1*,GAS2*,HSP82*, HXT14*,LOH1*,MAM1*,NRT1*,OSW1*,PFS1*,PHM6*,SPO19*,SPO75*,SPO77, SPR3*,SPS18*,SPS22*,SPS4*,SSP1*,TNA1*,YDR042c*,YFL040w*,YFR012w, YFR032c*,YGL138c*,YJL043w*,YKL177w*,YKR015c*,YOR365c*,YSW1*
ACE2	100%	15	AMN1*,BAT2*,BUD9*,CTS1*,DSE1*,DSE2*,DSE3*,EGT2*,GAT1*,HMS2*,NIS1*, PRY3*,RPA14*,SCW11*,YHB1*
CBF1	100%	5	ATP7*,ERG20*,NCE103*,PST1*,YIL127c*
SPT2	0%	4	HPF1,TAL1,YGP1,YIL169c
DAL82	50%	4	CAR2*,DAL4*,NTF2,THI2
HSF1	75%	12	CPR6*,ENO2*,FES1*,GLK1,HSC82*,HSP82*,HYP2,ILS1*,OST1,PGK1*,UBI4*,YLR217w*
OPI1	100%	5	CDS1*,FAS1*,FAS2*,INO1*,ITR1*
REB1	67%	12	CSR1*,HOM2*,HSC82*,MIC17,ORM1,OST1*,SEC20,SEC61*,TKL1*,YDR154c*, YGR235c*,YIP1
SWI4	100%	23	AGA1*,CIS3*,CLN1*,CLN2*,CWP1*,CWP2*,EXG1*,FTR1*,HTA1*,LAP4*,PCL1*, PLB3*,PRY2*,PTR2*,SRL1*,SVS1*,TOS6*,TSL1*,UTH1*,YHB1*,YOR248w*,YOX1*, YPS3*
STB5	33%	3	GND1,SOL3,YKL177w*

To be continued

Continued

TF	documented percent	the number of target gene	target genes
ABF1	76%	21	ADO1,AUR1*,BGL2*,BSD2*,FAS1*,HSP60*,HXK2,KRE1*,LAC1*,LPD1*,MRPL36*, NSE4*,PRC1*,PRE7,RTN1*,SLA1*,SOD1*,SSC1*,STE14*,TAH11,YLR349w
SKN7	80%	10	AHP1*,ASF1*,CWP2*,DDR48*,EXG1*,OCH1*,PRY2,RPA14*,SUR2,UTH1*
YAP1	100%	11	AHP1*,GPX2*,GRE2*,GSH1*,MCH4*,ROX1*,SOD1*,TRX2*,TSA1*,YLR108c*, YNL134c
IXR1	50%	6	EXG1,GPA1,PRM4,SCW4,YAL064w,YJR015w
MBP1	100%	9	ALG14*,CDC45*,OPY2*,RAD51*,RFA2*,RNR1*,SWE1*,YBR071w*,YMR144w*
ROX1	83%	6	BUD20*,CYB5*,FET4*,SMF3*,TIP1,YEL047c*
PUT3	100%	3	HPF1*,MCH5*,PUT2*
NRG1	100%	3	FRE4*,TPO4*,YAL018c*
RFX1	100%	9	CIT2*,GUP2*,RNR2*,RNR3*,RNR4*,SMF3*,YDL129w*,YMR279c*,YOR378w*

*Genes supported by YEASTRACT.