酵母基因上游转录因子结合位点分布的统计分析

崔 伟[a],黄 林[a],梁丽静[b,c]

(西北农林科技大学 a.信息工程学院, b.生物信息研究中心,c.生命科学学院,陕西 杨凌 712100)

[摘 要] 【目的】揭示真核生物基因上游转录因子结合位点的分布规律。【方法】挖掘了啤酒酵母基因组数据库(SGD)中基础域结构和β支架结构两超类(superclass)的转录因子结合在基因上游 0～2 000 bp 的位置;将转录因子按照结构和调节基因的不同功能进行聚类,并对其组内和组间的结合位点数进行比较分析。【结果】转录因子结合位点集中分布在转录起始位点上游 100～500 bp(61%);结构差异较大的转录因子的结合位点分布差异显著($P<0.01$);大多转录因子(19/22)在不同功能基因间结合位点分布差异不显著($P>0.05$)。【结论】转录因子结合位点分布与转录因子的结构相关性较大,而与所调控基因的功能相关性较小。推测不同家族转录因子结合位点在基因上游的分布具有特异性,有助于提供新的参数用以改进现有理论预测转录因子结合位点的方法。

[关键词] 啤酒酵母基因;转录因子;转录起始位点;SGD

[中图分类号] Q71 [文献标识码] A [文章编号] 1671-9387(2008)10-0215-06

Statistical analysis of the distribution of transcription factor binding sites in upstream regions of *Saccharomyces cerevisiae* genes

CUI Wei[a], HUANG Lin[a], LIANG Li-jing[b,c]

(*a. College of Information Engineering, b. Bioinformatics Center, c. College of Life Sciences, Northwest A&F University, Yangling, Shaanxi* 712100, *China*)

Abstract: 【Objective】 The research was to study the distribution of these transcription factors binding sites in upstream regions of eukaryote genes. 【Method】 The data of binding sites for transcription factors in the *Saccharomyces cerevisiae* genes database SGD database was mined and analyzed. Based on the structure of the transcription factors and the function of the genes they regulate, two superclass basic domains and beta-scaffold were formed and compared. 【Result】 The results revealed the distribution of transcriptional regulatory sites and most of them lay between 100 and 500 bp upstream of the transcription start site (TSS), showing the relationship between the binding sites, the structure of the transcription factors and the function of the genes they regulate. 【Conclusion】 These results may reveal the existence of class-specific features in the sits distribution bound by each family of TFs. , which can help people make sure which region contains the transcription regulation information, and make a better algorithm for the discovery of transcription factor binding sites.

Key words: *Saccharomyces cerevisiae* gene; transcription factor; transcription start site; SGD

真核生物生长发育取决于特定基因在特定时空 下的表达及调控,而基因调控最基本的环节是转录

* [收稿日期] 2007-11-05
[作者简介] 崔 伟(1983－),男,四川自贡人,在读硕士,主要从事基于网络的生物信息应用研究。E-mail:fsdd_liang@163.com
[通讯作者] 黄 林(1951－),女,陕西杨凌人,教授,硕士生导师,主要从事基于网络的计算机应用研究。E-mail:hl@nwsuaf.edu.cn

调控[1-2]。越来越多的研究表明,转录调控作用是DNA非编码区的重要功能之一[3]。早在20世纪80年代末、90年代初,就有人注意到真核基因上游存在多个转录调控结合位点,转录因子通过识别这些退化了的短序列(5～25 bp),调节基因在复杂的生命过程中的活性和表达[4]。转录因子及对应的结合位点,构成了转录调控网络最基本的结构单位,一对相互作用的反式(trans-)和顺式(cis-)调控元件,与下游的目标基因组成最简单的调控结构,这种结构有方向性,调控信息由转录因子向目标基因传递[5-6]。

分子生物学实验和新近发展的高通量技术,已经积累了许多转录调控资料,其中啤酒酵母(*Saccharomyces cerevisiae*)是基因组结构相对简单的一种真核模式生物,对其基因组上游转录调控机制已有较多研究报道[7-8]。随着基因组序列数据的积累和计算技术的发展,针对转录因子结合位点计算预测的算法和工具也越来越多,已有的预测方法可大致分为基于保守模体的方法和基于比较基因组学的方法两类[9-10]。

虽然已经有很多关于转录因子的研究[9-12],但是对于转录因子结合位点的分布规律还缺乏全面的研究,对于结合位点的理论预测也很少考虑到转录因子的具体结合区域[13-14]。本研究从酵母基因组的数据库SGD(Saccharomyces Genome Database)中提取数据,计算转录因子结合位点到转录起始位点(transcription start site,TSS)的距离,以期为揭示转录因子结合位点的分布规律及真核基因的调控机理提供科学依据。

1 材料与方法

1.1 数据来源

酵母基因组数据库SGD是已经完成基因组全序列测定的啤酒酵母基因组数据库,包括啤酒酵母的分子生物学及遗传学等大量信息[15-16]。从文献[12]所报道的117个转录因子及其所调节的基因中,选取基础域结构转录因子14个(Cad1、Cin5、Sko1、Yap1、Yap7、Gcn4、Swi4、Swi6、Cbf1、Phd1、Sok2、Pho4、Ino2、Ino4)和β支架转录因子8个(Mcm1、Rlm1、Hap2、Hap3、Hap4、Hap5、Rox1、Spt2)进行研究。在SGD数据库中查找相应的转录因子所调控的基因,根据文献[17]等对酵母转录因子结合位点的分析可知,转录因子结合位点一般位于转录起始位点上游800 bp区域内。为了得到更全面的信息,本研究查询啤酒基因组所有基因上游2 kb以内上述22个转录因子结合的位点,并计算其与转录起始位点的距离。

1.2 方 法

1.2.1 *转录因子在转录起始位点上游区域内的结合位点数统计* 将转录起始位点上游2 000 bp划分为20个100 bp的区间,统计每个转录因子在不同区间的结合位点数,再分析两者的相关性。

1.2.2 *转录因子的聚类* 根据1.1中所选转录因子的结构,按照GeneRegulation网站上的Transcription Factor Classification分类标准[18](http://www.gene-regulation.com/pub/databases/transfac/cl.html),对所选转录因子进行聚类分析。

1.2.3 *基因功能聚类* 根据转录因子所调控基因的最主要功能[19](参考酵母资源中心公共数据库YRC[20]和SGD数据库中的注释),对基因及其上游的结合位点进行聚类,选择调控基因功能类型大于等于2类、并且每类中结合位点数不少于10个的因子,进行类间结合位点分布差异的比较。

1.3 数据统计分析方法

用SPSS 13.0软件对所得数据进行差异性检验,采用Mann-Whitney U检验两组间转录因子结合位点的分布差异;采用Kruskal-Wallis H法检验多组间转录因子结合位点分布的差异,并采用秩变换技术结合完全随机设计的方差分析[21]进行秩和检验。

2 结果与分析

2.1 转录因子结合位点的分布

在所选取的22个转录因子中,很多转录因子虽然可同时调控多个基因的表达,但在不同基因上游的结合位点与转录起始位点的距离有差异。表1列出了每个因子在基因转录起始位点上游不同区域的结合位点数。表1显示,大多转录因子(18/24)有85%以上的结合位点位于TSS上游1 000 bp区域,其在1 000 bp以外没有结合位点或者只有个别的结合位点,在2 000 bp以外的区域没有结合位点出现。

统计不同区域所有转录因子的结合数(图1),与表1分析的结果一致,有89.0%的结合位点是在1 000 bp以内。图1显示,转录因子结合位点并不是均匀地分布在这些区域,而是呈偏态分布,只有很少的位点是在TSS上游100 bp以内的,而大部分(61%)转录因子结合位点分布在蛋白编码区上游

100～500 bp,与文献[6]的结果基本一致(74%);而在较远的距离则很少有结合位点,尤其在 1 000 bp 以外,只有个别的转录因子在上面有结合位点。

表 1　转录因子在转录起始上游的结合位点数目

Table 1　Number of the binding sites for every factor in the upstream of TSS

结构类型 Superclass	转录因子 Transcri-ption factor	短距离区域/bp Short distance from TSS									
		0～100	100～200	200～300	300～400	400～500	500～600	600～700	700～800	800～900	900～1 000
基础域结构因子 Basic domain factors	Cad1	2	3	1	2	2	1	1	0	1	0
	Cin5	8	17	12	11	13	4	10	1	2	1
	Sko1	0	0	0	3	0	1	2	2	0	0
	Yap1	0	8	4	2	2	1	0	0	0	0
	Yap7	8	17	17	12	6	5	2	2	3	0
	Gcn4	8	46	38	22	9	6	6	6	4	3
	Swi4	0	12	32	26	16	14	5	9	5	2
	Swi6	0	9	16	16	9	12	2	4	3	2
	Cbf1	5	38	35	37	28	20	7	3	3	5
	Phd1	2	2	4	5	9	5	14	12	8	3
	Sok2	2	0	8	14	6	12	10	10	8	8
	Pho4	2	3	4	8	4	3	1	0	0	0
	Ino2	1	9	7	7	1	2	2	1	1	0
	Ino4	0	8	2	6	2	0	1	2	1	0
β-支架因子 beta-Scaffold	Mcm1	2	11	17	13	6	9	3	2	1	2
	Rlm1	0	0	3	2	2	0	1	1	1	0
	Hap2	2	3	8	4	1	1	2	0	1	0
	Hap3	1	3	6	3	1	1	1	0	0	0
	Hap4	2	14	8	3	3	3	0	0	1	0
	Hap5	2	3	7	4	3	3	2	0	1	0
	Rox1	0	1	1	0	1	1	0	0	0	0
	Spt2	1	4	3	2	4	1	1	0	0	0

结构类型 Superclass	转录因子 Transcri-ption factor	长距离区域/bp Long distance from TSS										总计 No. of all
		1 000～1 100	1 100～1 200	1 200～1 300	1 300～1 400	1 400～1 500	1 500～1 600	1 600～1 700	1 700～1 800	1 800～1 900	1 900～2 000	
基础域结构因子 Basic domain factors	Cad1	0	0	1	0	0	0	0	0	0	0	14
	Cin5	5	7	3	1	3	3	1	0	1	5	108
	Sko1	0	2	0	0	0	1	0	0	0	0	11
	Yap1	0	0	0	0	0	0	0	0	0	0	17
	Yap7	0	0	1	0	0	0	0	0	0	0	73
	Gcn4	2	0	0	0	0	0	0	0	0	0	150
	Swi4	2	6	0	1	2	0	0	0	1	0	133
	Swi6	1	2	1	0	4	0	0	0	0	0	81
	Cbf1	0	5	0	1	0	0	0	1	0	0	188
	Phd1	3	5	4	2	5	2	2	1	0	0	88
	Sok2	5	8	5	3	3	2	1	0	0	0	105
	Pho4	0	1	0	0	0	1	1	0	0	0	28
	Ino2	1	0	0	1	0	0	0	0	0	0	33
	Ino4	1	0	0	0	0	0	0	0	0	0	23
β-支架因子 beta-Scaffold	Mcm1	2	1	1	1	0	0	2	0	0	0	73
	Rlm1	0	0	0	0	1	0	0	0	0	0	11
	Hap2	0	0	1	0	0	0	0	0	0	0	23
	Hap3	0	0	0	0	0	0	0	0	0	0	16
	Hap4	0	0	0	0	0	0	0	0	1	0	35
	Hap5	0	0	2	0	0	1	0	1	0	0	29
	Rox1	0	1	0	0	0	0	1	0	0	1	7
	Spt2	1	0	1	0	0	0	0	0	0	0	18

对其中结合位点数量最多的 3 个转录因子的分布情况进行比较,发现其各自也符合总体规律,都呈偏态分布,而且多集中在 TSS 上游 100～500 bp。但三者之间仍有部分差异,因此又进一步检验了不

同转录因子间结合位点的分布差异。

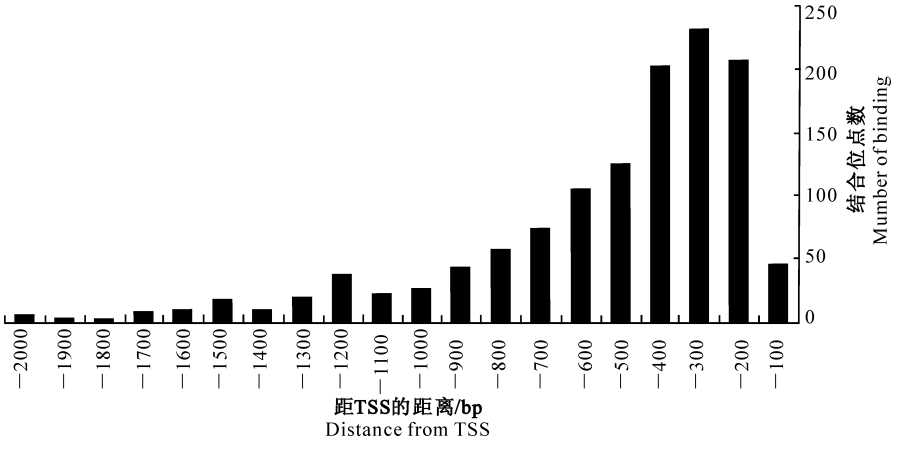

图1 转录因子结合位点的分布

Fig. 1 Distribution of the binding sites of transcriptional factor

2.2 转录因子结构与结合位点的关系

2.2.1 转录因子聚类结果 对以上22个转录因子按照其结构聚类的结果如下:

1超类:基础域结构转录因子(Basic Domains)

1.1类:亮氨酸拉链因子(Leucine zipper factors,bZIP)

包括Cin5、Sko1、Yap1、Yap6、Yap7

1.1.1族:衔接元件(AP-1(-like) components)

包括Gcn4

1.1.2族:只含有亮氨酸拉链(ZIP only)

包括Swi4、Swi6

1.2类:螺旋-环-螺旋因子(Helix-loop-helix factors,bHLH)

包括Cbf1,Phd1, Sok2

1.2.1族:Hairy家族

包括Pho4

1.2.2族:INO家族

包括Ino2、Ino4

2超类:β支架转录因子(beta-Scaffold Factors with Minor Groove Contacts)

2.1类:MADS box家族

包括Mcm1、Rlm1

2.2类:CCAAT家族

包括Hap2、Hap3、Hap4、Hap5

2.3类:HMG家族

包括Rox1、Spt2

2.2.2 不同功能间结合位点的分布差异 对两个超类转录因子结合位点的分布规律进行秩和检验。结果显示,这两组结构有很大差异的转录因子,其结合位点的分布也有显著差异($P=0.004$)。

对每个超类下的每一类中的转录因子分组进行检验可知,基础域结构下的亮氨酸拉链因子、螺旋-环-螺旋因子两类转录因子的结合位点分布存在极显著差异($P<0.001$),β-支架因子中的MADS box家族、CCAAT家族、HMG家族3类之间也有显著性差异($P=0.006$)。

再进一步比较具有相似结构的转录因子之间结合位点的分布,即比较ZIP only家族中的Swi4与Swi6,INO家族中的Ino2与Ino4,MADS box家族中的Mcm1与Rlm1,CCAAT家族中的Hap2、Hap3、Hap4与Hap5,HMG家族中的Rox1与Spt2,检验结果显示,以上同一家族结构相似的转录因子间均无显著性差异($P>0.05$)。表明在结构相似的同一家族的转录因子间,结合位点的分布没有明显差异。

2.3 转录因子结合位点在不同基因间的分布

根据每个基因的主要功能,对上述22个转录因子所调控的1 264个基因进行分类,包括:生物合成(1)、胁迫应答(2)、细胞壁(3)、转录和调控(4)、RNA(rRNA、tRNA)(5)、传送器(6) 6类功能,结果见图2。对转录因子结合位点在6组基因上游的分布进行比较,结果6组间差异显著($P=0.001$);再作秩和检验,结果显示6组间差异也显著($P=0.001$)。多重极差检验结果为:(5)、(3)、(4)组间无显著差异($P=0.063$),(3)、(4)、(2)、(1)组间无显著差异($P=0.592$),(2)、(1)、(6)组间无显著差异

($P=0.051$);而(3)、(4)、(5)均与(6) 组间有显著差异($P>0.05$)。

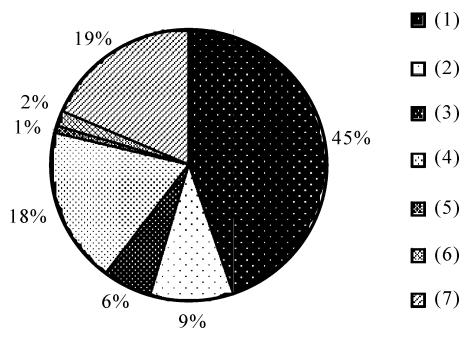

图 2　结合位点按基因功能分类的结果

(1)生物合成;(2)胁迫应答;(3)细胞壁;(4)转录和调控;(5)RNA ;(6)传送器;(7)其他

Fig. 2　The binding sites distribution classified by gene function

(1)Biosynthesis;(2)Stress responsive;(3)Cell wall;(4)Transcription and regulation;(5)RNA;(6)Transporter;(7)Others

2.4　基因功能对结合位点的影响

将每个转录因子按其所调节的基因功能分成若干组(分类标准与 2.3 相同),对同一个转录因子的结合位点在不同组间的分布进行检验。结果表明,除 Cin5 ($P=0.02$)、Phd1 ($P=0.034$)、Mcm1 ($P=0.018$)3 个转录因子在所调控的不同功能基因间结合位点的分布有显著差异外,其他转录因子在其所调控的不同功能的基因间,结合位点分布的差异不显著。表明转录因子所调控基因的功能对其结合在基因上游的位置无显著影响。

3　讨　论

基因转录调控是一个复杂的问题,转录调控位点有可能存在于基因序列中的任何区域。真核基因的转录起始位点上游存在转录因子结合位点的现象,已被越来越多的试验所证实[2,5-6]。有试验结果表明,许多转录因子间有协同调控的作用,但对这种协同作用的机制还缺乏详细认识[19]。本研究中,转录调节因子结合位点在 TSS 上游 100 bp 以内很少有结合位点,这是因为这一区域一般都包括转录起始位点和转录起始所需的各种转录起始元件。在线性的 DNA 上,酵母转录因子主要作用在短距离内(100～500 bp),这个距离可能可以使这些转录因子专注地作用于其所调控的基因上,并能同时减少对周围其他基因不恰当的激活作用;这个区域可能更有利于转录因子间的协同调控;转录因子的结构又对其结合位点的序列有特异要求,所以可以推测,正是酵母基因上游这些位点功能上的束缚,决定了这些转录因子结合位点的分布。这有助于了解它们在转录调控中的作用以及如何起作用,对于理解真核基因的调控机理是一项很重要的工作。

一个转录因子在染色体上的结合位置,可能会与其所调控的基因或其本身的性质有关。从本研究结果可以看出,对于大多数转录因子而言,同一个转录因子在不同功能基因上的结合位点的分布差异并不显著,但是结构有差异的转录因子间结合位点的分布有显著差异。结构差异小的转录因子(同一家族内)结合位点的分布差异不显著,这可能与同一家族转录因子多存在协同作用、使结合更稳定有关。而结构差异较大非同一家族的转录因子结合位点的分布差异显著,这说明转录因子在基因上游所结合的位置与转录因子的结构相关性较大,不同结构的转录因子的结合位点分布有特异性。

现在的转录因子结合位点预测方法,是以结合位点信号的保守性为出发点,基于保守模体或基于比较基因组学[1,9-10]。本研究结果表明,结构不同的转录因子的结合位点分布有差异,对于已知结构的转录因子,可以从这种转录因子结合位点位置的特异性出发,来预测结合位点,以提高预测的特异性。综合考虑位置特异性与结合位点信号的保守性来提高预测性能,是作者下一步将要进行的工作。

致谢:本研究得到了袁志发教授的热情帮助和指导,在此深表感谢!

[参考文献]

[1]　钟　东,张振书,刘宇虎,等. 初步建立真核基因调控元件模块的搜索方法 [J]. 第一军医大学学报,2004(2):57-61.
Zhong D, Zhang Z S, Liu Y H, et al. Establishment of the methods for searching eukaryotic gene cis-regulatory modules [J]. Journal of First Military Medical University, 2004(2):57-61. (in Chinese)

[2]　雷耀山,史定华,王翼飞. 基因调控网络的生物信息学研究 [J]. 自然杂志,2004(1):7-12.
Lei Y S, Shi D H, Wang Y F. Reviewing the study of gene regulatory networks from ioinformatics [J]. Chinese Journal of Nature, 2004(1):7-12. (in Chinese)

[3]　Sun Q, Chen G, Streb J W, et al. Defining the mammalian CArGome [J]. Genome Res, 2006, 16:197-207.

[4]　Folter S D, Angenent G C. Trans meets cis in MADS science [J]. Trends Plant Sci, 2006, 11(5):224-231.

[5]　张昆林,张　静,罗静初. 酵母基因上游与内含子可能存在的转录协同作用 [J]. 生物化学与生物物理进展,2005,32(1):46-52.
Zhang K L, Zhang J, Luo J C. Potential transcriptional synergy

between upstream regions and introns of yeast genes [J]. Progress in Biochemistry and Biophysics, 2005, 32(1): 46-52. (in Chinese)

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

[7] 胡 俊,杨建红,李 琰,等. 酵母基因内含子中转录正调控位点的统计分析 [J]. 生物物理学报, 2004(1): 43-49.
Hu J, Yang J H, Li Y, et al. Statistical analysis of positive triscriptional regulatory sites in the introns of yeast genes [J]. Acta Biophysica Sinica, 2004(1): 43-49. (in Chinese)

[8] 苗 福,蒋湘宁. 基因转录调控相关数据库集成系统及其应用 [J]. 生物信息学, 2004(4): 37-41.
Miao F, Jiang X N. Gene transcriptional regulation databases and database integrated systems: their constitution and applications [J]. China Journal of Bioinformatics, 2004(4): 37-41. (in Chinese)

[9] Wyrick J J, Young R A. Deciphering gene expression regulatory networks [J]. Curr Opin Genet Dev, 2002(12): 130-136.

[10] Lenhard B, Sandelin A, Mendoza L, et al. Identification of conserved regulatory elements by comparative genome analysis [J]. Biol, 2003, 2: 13.

[11] Stormo G. DNA binding sites: representation and discovery [J]. Bioinformatics, 2000, 16(1): 16-23.

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

[13] Kellis M, Patterson N, Endrizzi M, et al. Sequencing and comparison of yeast species to identify genes and regulatory elements [J]. Nature, 2003, 423: 241-254.

[14] Cliften P F, Hillier L W, Fulton L, et al. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis [J]. Genome Res, 2001, 11(7): 1175-1186.

[15] Cherry J M, Adler C, Ball C, et al. SGD: Saccharomyces Genome Database [J]. Nucleic Acids Res, 1998, 26(1): 73-79.

[16] Tompa M, Li N, Bailey T L, et al. Assessing computational tools for the discovery of transcription factor binding sites [J]. Nat Biotechnol, 2005, 23: 137-144.

[17] Van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies [J]. J Mol Biol, 1998, 281(5): 827-842.

[18] Wingender E, Chen X, Fricke E, et al. The TRANSFAC system on gene expression regulation [J]. Nucleic Acids Research, 2001, 29(1): 281-283.

[19] Lee T I, Rinaldi N J, Robert F, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae* [J]. Science, 2002, 298: 799-804.

[20] Riffle M, Malmstrom L, Davis T N. The yeast resource center public data repository [J]. Nucleic Acids Res, 2005, 33: D378-D382.

[21] 刘万里,薛 茜,曹明芹,等. 用 SPSS 实现完全随机设计多组比较秩和检验的多重比较 [J]. 地方病通报, 2007(2): 27-29.
Liu W L, Xue X, Cao M Q, et al. Nonparametic test of completely randomized design and multiple comparisons with SPSS [J]. Endemic Diseases Bulletin, 2007(2): 27-29. (in Chinese)