THE CONSERVATION OF TRANSCRIPTION FACTOR-BINDINS SITES IN *SACCHAROMYCES* GENOMES.

Kovaleva G.Y.[1,2]***, Mironov A.A.***[1]***, Gelfand M.S.***[2]
[1] Moscow State University, Department of Bioengeneering and Bioinformatics, Moscow, Russia.
[2] Institute for Information Transmission Problems, RAS, Moscow, Russia.
Corresponding author: kovaleva@iitp.ru

Key words: comparative genomics, yeast, *Saccharomyces*, transcription, regulatory site, Gcn4p, Met31/Met32, Cbf1/Met4/Met28, Leu3.

Resume

Motivation: two independent studies on prediction of regulatory sites in genomes of the genus *Saccharomyces* by comparative analysis produced contradictory results. In this study we studied the conservation rate of known and predicted binding sites for regulatory proteins of several metabolic pathways in *Saccharomyces* genomes.
Results: comparative analysis of seven *Saccharomyces* genomes shows that in most cases the binding sites are not perfectly conserved, or conserved but in a different position. This observation contradicts previously postulated statements that functionally significant regions are absolutely conserved and occupy the same in related genomes.

Introduction: extracting the complete functional information encoded in a genome — including genic, regulatory and structural elements — is a central challenge in biological research. Prediction of non-protein-coding functional regions, such as regulatory elements, is especially difficult because of are usually short (6-15 bp for *S.cerevisiae* and many other eukaryotic genomes), often degenerate, and can reside on either strand of DNA at variable distances from the genes they control. Since functional sequences tend to be conserved through evolution, they can appear as 'phylogenetic footprints' in alignments of genome sequences of different species (Hardison et al., 1997). Recently, two groups sequenced several *Saccharomyces* genomes. The main goal of these studies was to predict the regulatory sites in *Saccharomyces* spp. using multiple whole-genome alignment in one case (Kellis et al., 2003) and multiple alignments of the gene upstream regions in another (Cliften et al., 2003). Results were represented as two lists of predicted binding motifs. Our comparison of these lists shows a very moderate intersection even accounting to fact that in study by Kellis et al. the predicted motifs were constructed using IUB codes. This prompted us to analyze the conservation rate for known and predicted binding sites in *Saccharomyces* genomes in more detail.

Methods

Complete genome sequences of *Saccharomyces cerevisiae* and *Candida albicans* were extracted from GeneBank. The fragments covering 750 nucleotides of upstream regions and 150 nucleotides of protein-coding regions were considered. Search for orthologs was done using fungiBLAST (http://www.ncbi.nlm.nih.gov/BLAST/Genome/FungiBlast.html). In some cases orthologous regions for the protein-coding part of a used fragment were not found; and such fragments were ignored. On the contrary, regions upstream of orthologous genes were used even if they did not produce a strong alignment.
For identification of orthologous genes and site patterns, Genome Explorer program (Mironov et al., 1999) was used. SignalX (Mironov et al., 2000) was used to construct nucleotides weight matrices. Multiple sequence alignments were done using ClustalX (Thompson et al., 1997).

Results and discussion

Site conservation. Unlike many eukaryotic organisms yeasts are able to synthesize amino acids. Therefore, they have corresponding metabolic pathways and regulators. The global regulator of amino acid biosynthesis is Gcn4p. It translation is activated condition of starvation for some amino acids. The regulator binds to Gcn4p-responsive element (GRE), with the consensus TGACTC in the upstream regions of regulated genes (Natarajan et al., 2001; Hinnebusch and Natarajan, 2002). It is also known that in upstream regions of genes regulated by Gcn4p usually contain at least two GRE. Based on published experimental data, we seleted nine genes that are certainly regulated by Gcn4p: *HIS3*, *ARG8*, *ARG1*, *ADE4*, *ILV1*, *TRP4*, *HIS4*, *HIS7*, and *ILV2* (Natarajan et al., 2001). Using the TRANSFAC database we constucted a position weight matrice to identify significant binding sites in the upstream regions of these genes. The conservation rate for all these sites are shown in Table 1.
From these data it follows that the conservation rate of known and strong predicted binding sites is similar. Still, even strongest sites are not necessary conserved in all examined genomes, as it has been thought previously.

Genome/conserved sites	known	predicted	weak
S.cerevisiae	**11**	**6**	**32**
S.paradoxus	**10**	**4**	**13**
Non-conserved	1(0)	2(0)	19(0)
S.mikatae	**6**	**4**	**10**
Non-conserved	4(1)	2(0)	19(3)
S.kudriavzevii	**3**	**3**	**5**
Non-conserved	1(7)	0(3)	7(20)
S.bayanus	**0**	**2**	**2**
Non-conserved	0(11)	0(4)	7(23)

Table 1. Conservation rates of known, strong and weak predicted binding sites of Gcn4p in 5 of 7 analyzed *Saccharomyces* genomes. The number of conserved sites in each genome is set in bold. The numbers in parentheses are the number of sites that could not be conserved since orthologous region is absent or not sequenced in corresponding genome.

The biosynthesis of methionine is regulated by three more or less independent regulators/regulatory complexes: Gcn4p, Met31/Met32, and Cbf1/Met4/Met28. Prediction of binding sites for Gcn4p was done using the same matrice as above. For the other two regulatory complexes, only the binding site consensus is known: TCACGTG for the Met31/Met32 complex and AAACTGTGG for the Cbf1/Met4/Met28 complex (Thomas and Surdin-Kerjan, 1997). Nevertheless, we constructed the corresponding matrices and applied them to known regulatory genes (Thomas and Surdin-Kerjan, 1997). Thus subsets of known and predicted binding sites were identified. As it can be seen in Table 2, the conservation rates of these groups of sites are similar to those for Gcn4p (see above).
The leucine biosynthesis pathway is regulated by two regulators Gcn4p, a master regulator of amino acid biosynthesis, and specific regulator Leu3. Leu3 binding sites are not given in TRANFAC database, and we constructed a position weight matrice on the basis of the consensus (Kohlhaw, 2003) and few experimentally known binding sites (most genes of this regulon are known to be regulated from expression array experiments).
The conservation rates for these regulators are given in Table 3 and are more or less similar to the rates for other studied regulators. However, the binding sites for Leu3, both known and predicted, seem to be more conserved compared to the first two cases. The difference can be due to the lengh of the signal: for Leu3 it seems to be about 10 nucleotides, as internal positions also contribute to the strength of interaction.

Genome/conserved sites	Met31+ Met32	known	predicted	weak	Cbf1-complex	known	predicted	weak
S.cerevisiae		**10**	**11**	**80**		**8**	**15**	**51**
S.paradoxus		**8**	**9**	**22**		**2**	**6**	**16**
Non-conserved		0(2)	1(1)	43(5)		5(1)	6(3)	35(0)
S.mikatae		**4**	**3**	**7**		**2**	**2**	**8**
Non-conserved		2(4)	2(6)	22(51)		2(4)	2(11)	20(23)
S.kudriavzevii		**2**	**1**	**2**		**2**	**0**	**4**
Non-conserved		1(7)	2(8)	10(68)		2(4)	2(13)	11(36)
S.bayanus		**2**	**2**	**5**		**1**	**1**	**0**
Non-conserved		0(8)	2(7)	22(53)		3(4)	1(13)	18(33)

Table 2. Conservation rates of known, strong and weak predicted binding sites of Met31/Met32 and Cbf1/Met4/Met28 complexes in five of seven analyzed *Saccharomyces* genomes. Notations as in Table 1.

Genome/conserved sites	known	predicted	weak
S.cerevisiae	**5**	**6**	**12**
S.paradoxus	**4**	**5.5**	**5**
Non-conserved	0(1)	0.5(0)	7(0)
S.mikatae	**4**	**4**	**1**
Non-conserved	0(1)	1(1)	6(5)
S.kudriavzevii	**1.5**	**3**	**1**
Non-conserved	0.5(3)	0(3)	5(7)
S.bayanus	**2.5**	**1**	**1**
Non-conserved	1.5(1)	0(5)	5(7)

Table 3. Conservation rates of known, strong and weak predicted binding sites of Leu3 in five of seven analyzed *Saccharomyces* genomes. Notations as Tables 1 and 2. Fractional numbers reflect that strong (known or predicted) sites may become weaker, but still above the threshold, due to partially inactivating mutations.

Thus, the standard technique of comparative genomics can not be used to yeast genomes without corrections. There is no universal conservation of binding sites that is often observed in prokatyotic organisms.

Site clusterization. It is known that most eukatyotic regulatory signals are short and the selectivity of regulator binding is obtained by clustering of sites them in the upstream regulatory regions. We considered clusterization of binding sites of Gcn4p in the upstream regions of nine genes listed above of *Saccharomyces cerevisiae* and *Candida albicans* genomes. Clusterization of binding sites is not as strict criterion as absolute conservation of the exact motif at a certain position. However, our data shows that clusterization of binding sites is evolutionary significant because of it is observed in so divergent genomes. Still, the statistical parameters are unique for every cluster in each genome. We observed that genes known to be regulated by Gcn4p in most cases have clusters of candidate Gcn4p binding sites in the upstream regions genes from *S.cerevisiae* and from *C.albicans*.

Thus, experimentally known and strong predicted binding sites for transcriptional regulators of several yeast metabolic pathways have a similar conservation rate in related genomes, whereas weak predicted sites are less conserved. However, even in closely related genomes conservation of experimentally defined sites is not guaranteed. A weaker form of comparative analysis,

requiring conservation of site clusters irrespective of their parameters, can be used at larger evolutionary distances, but the recognition rules turn out to be not very specific.

Acknowlegments

We are grateful to Lena Stavrovskaya for kindly provided software. This study was supported by grants from HHMI, LICR, RFBR, the Fund for Support of Russian Sciences, and the Program in Molecular and Cellular Biologyof RAS.

References

Cliften P., Sudarsanam P., Desikan A., Fulton L., Fulton B., Majors J., Waterston R., Cohen B.A., Johnston M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting**.** *Science*, 301, 71-76.

Hardison R.C., Oeltjen J., Miller W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res*., 7, 959-966.

Hinnebusch A.G., Natarajan K. (2002) Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryotic Cell*, 1, 22-32.

Kellis M., Patterson N., Endrizzi M., Birren B., Lander E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423, 241-254.

Kohlhaw G.B. (2003) Leucine biosynthesis in fungi: entering metabolism through the back door. *Microbiol Mol Biol Rev.*, 67, 1-15.

Mironov A.A., Koonin E.V., Roytberg M.A. and Gelfand M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, 27, 2981-9.

Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000) GenomeExplorer: software for analysis of complete bacterial genomes. *Mol. Biol*., 34., 222-231.

Natarajan K., Meyer M.R., Jackson B.M., Slade D., Roberts C., Hinnebusch A.G., Marton M.J. (2001) Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast**.** *Mol Cell Biol.*, 21, 4347-4368.

Thomas D., Surdin-Kerjan Y. (1997) Metabolism of sulfur amino acids in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev.*, 61, 503-32.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25. 4876-4882.