Biophysics, Vol. 48, Suppl. 1, 2003, pp. S26–S29.

Computer Analysis of the *GCN4* Regulon of Yeast *Saccharomyces cerevisiae*

G. Yu. Kovaleva[1], V. Yu. Makeev[2], and M. S. Gelfand[1,2,3]

[1]Department of Bioengineering and Bioinformatics, Moscow State University, 1-73 Vorobievy Gory, Moscow, 119992 Russia
[2]State Scientific Center GosNIIGenetika, 1 1st Dorozhny proezd, Moscow, 117545 Russia
[3]Institute for Problems of Information Transmission, Russian Academy of Sciences, 19 Bolshoi Karetny per., Moscow, 127994 Russia

Received November 10, 2003; in final form, Junuary 26, 2004

Abstract—Binding sites of eukaryotic transcriptional regulators are often very short, and the specificity of recognition is attained by cooperative binding of regulators to clusters of sites. We analyzed clustering of binding sites of the global regulator of amino acid metabolism Gcn4p in regulatory regions of nine genes of yeast *Saccharomyces cerevisiae* and of orthologous genes of a phylogenetically quite distant yeast *Candida albicans*. Despite differences in the parameters characterizing the clusters, the clustering of candidate Gcn4p is retained in most regulatory regions, which confirms the functional significance of this phenomenon. Analysis of a control set of genes that are not regulated by Gcn4p demonstrates that this clustering is not random.

Key words: yeast, transcriptional regulators, Gcn4p binding sites, clustering

INTRODUCTION

The transcriptional apparatus of eukaryotic cells differs from that of prokaryotes. In particular, the regulatory regions of most genes contain binding sites for many regulators. The length of these sites is 6–8 nucleotides, and the specificity of recognition is sustained owing to the existence in the regulatory region of clustered binding sites of one type (so-called isotypic clusters). There are two possible explanations of this phenomenon. Firstly, multiplicity of isotypic sites could lead to high-affinity cooperative binding of the regulator molecules [1]. Secondly, site clusters could promote lateral diffusion of activator proteins from sites of low affinity to the high-affinity ones [2–4].

The eukaryotic cell of a unicellular organism is capable of fast changes in the metabolism depending on the external conditions. Thus, amino acid starvation leads to increased translation of the mRNA of the *GCN4* gene, the main regulator of the amino acid metabolism. The experimental data show that the Gcn4p protein activates from 9 to 30 genes whose products are involved in amino acid biosynthesis [5–7]. The regulatory regions of most of these genes contain binding sites of Gcn4p, called GCRE (*GCN4* responsive element) with consensus TGACTC [8], and in most cases there are several GCRE elements in one regulatory region [9].

There exist several algorithms for identification of monotypic site clusters. One such algorithm, CLUSTER [10], was applied to the analysis of Gcn4p regulons of yeasts *Saccharomyces cerevisiae* [ftp://ftp.ncbi.nih.gov/genomes/Saccharomyces_cerevisiae/] and *Candida albicans* [ftp://cycle.stanford.edu//pub/projects/candida/].

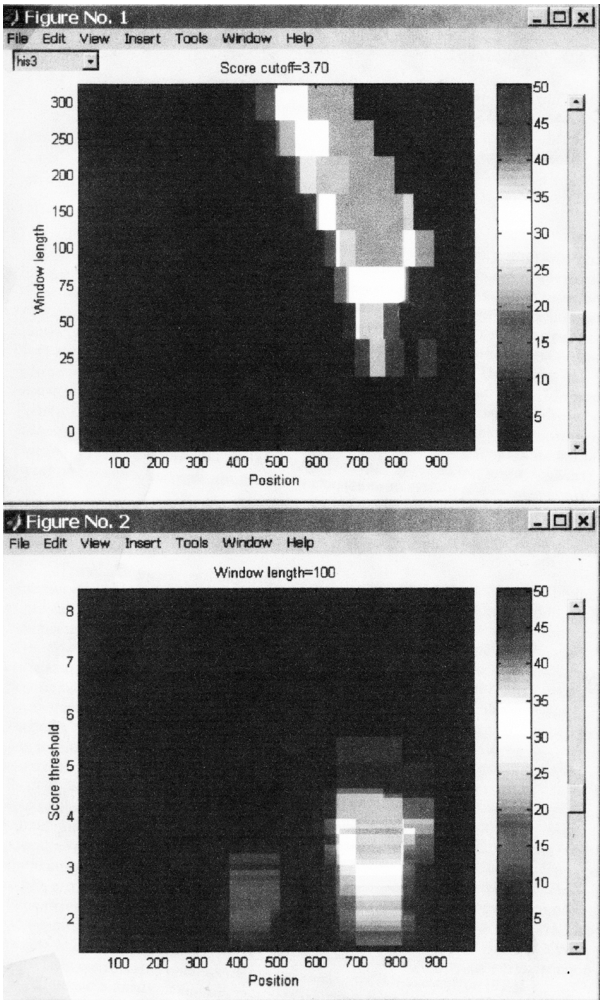
RESULTS AND DISCUSSION

In *S. cerevisiae* the regulatory regions of genes normally occupy up to 250 bp from the transcription start, although larger regulatory regions of length up to 1000 bp have been observed [11]. Experimental data show that most binding sites of Gcn4p occur

within 600 bp from the transcription start [12]. However, preliminary analysis identified candidate sites at larger distances. Besides, for most genes the transcription start point is not known. Therefore, we considered 1000-bp fragments upstream of the translation start codon. These regions were selected from the genome using GenomeExplorer [13].

Based on published experimental data, we selected nine genes known to be regulated by Gcn4p: *HIS3*, *ARG8*, *ARG1*, *ADE4*, *ILV1*, *TRP4*, *HIS4*, *HIS7*, and *ILV2* [7]. Site clusters were identified in several steps. Experimentally found sites were selected from the TRANSFAC database [http://www.gene-regulation.com/index.html; 14], aligned, and used to construct the recognition profile. This profile is used to compute the score of a candidate site and to select sites passing the threshold. Then clusters of sites are determined and their significance is computed. Each cluster is characterized by its position in the sequence fragment, the threshold for individual sites, and window size. For convenience, two of these parameters may be fixed; then the cluster significance is a function of the remaining parameter [10].

The obtained significance values were visualized using MatLab^{TM}6.5. For each regulatory gene, two parametric portraits are constructed, allowing one to assess the dependence of the cluster significance on the three parameters. The portraits for the upstream region of the gene *HIS3* are shown in the figure as an example. The observed cluster in the region 250–300 bp upstream of the translation start is highly significant. Analogous plots were constructed for other genes (not shown).



Statistical significance of theGcn4p binding site clusters in the regulatory region of gene HIS3. Horizontal axis: position. Vertical axis: threshold (left) or window size (right). Shading denotes significance (white for high, gray for intermediate, black for low).

The goal of this step was to verify the existence of Gcn4p binding site clusters. It turned out that the significance of the observed clusters varies, the strongest site was observed in the regulatory region of *ADE4*, and the weakest sites were seen upstream of *ILV2* and*TRP4*. Nevertheless, all analyzed genes had significant Gcn4p site clusters.

To check that this is not due to random fluctuations, we considered 104 genes of known functions for which no data about Gcn4p have been published. Candidate clusters were sought using the same settings of the program. Upstream regions of most genes in this control sample did not contain sites clusters, although a minority (about 10%) had clusters of significance comparable to that in the training set. However, the parameters of these spurious clusters (in particular, position relative to gene starts, see below) were different from those of the real sites. In fact, a recent experimental study claimed that Gcn4p may regulate much more genes than it had been known, up to 10% of the yeast genome [10]. Thus, it might be possible that some of these genes were included into the control sample.

Then we attempted to create a recognition rule that would combine the values of all three parameters characterizing true sites. This would allow one to scan the entire genome in order to find additional genes most likely regulated by Gcn4p. Unfortunately, this proved to be impossible, as the parameters of significant clusters for all genes in the training set differed widely. In fact, the only common feature of the true clusters was their high statistical significance.

Unfortunately, computing the significance of clusters over the entire range of the parameter values by the CLUSTER algorithm is computationally difficult, and cannot be done for the full genome. On the other hand, we have observed that all true sites occupy the region 200–250 bp upstream of the translation start.

At the next stage, we examined whether clustering of Gcn4p binding sites is a universal feature of this regulator. For comparison, we considered the genome of yeast *Candida albicans*, which is quite distant from *S. cerevisiae* on the phylogenetic tree. The genome of *C. albicans* contains a gene orthologous to *GCN4*.

Eight out of nine considered genes had orthologs in *C. albicans* with 45% identity on the amino acid level. One more gene, *HIS4*, had a 21%-identical ortholog. All identified orthologs were retained for further analysis.

Since the considered genomes diverged fairly early, one should not expect either alignability of gene upstream regions or complete coincidence of site clusters. Thus the goal was to verify the existence of sites *per se*. This proved to be true indeed, as most *C. albicans* orthologs have candidate sites. However, it should be noted that the parameters of significant clusters differed more widely than in *S. cerevisiae.* In particular, as noted above, a common feature of the *S. cerevisiae* clusters was the distance to the translation start. For orthologs from *C. albicans* this rule is not universal: although most clusters are at the same distance from the translation start point, in some cases they are closer to the gene, in the region 200 bp upstream of the translation start. One should also mention that in the case of *ADE4*, the significance of the cluster changes considerably: in *S. cerevisiae* this gene has the strongest cluster, whereas in *C. albicans*, the cluster is rather weak. Still, these observations confirm the universality of the clustering mechanism: all orthologs from *C. albicans* have clusters of candidate Gcn4p-binding sites, and most of these clusters are characterized by comparable significance and approximately the same distance to the start.

Further analysis will aim at creating an exact recognition rule capable of identification of clusters of binding sites in the complete genome. An analogous study for *S. cerevisiae* has been performed [16]. The conditions established in this study were the existence in the 5'-region of at least three candidate Gcn4p-binding sites, so that one of them should fall in the region 40–300 bp upstream of the transcription start, another, 500 bp or closer, and the third, at most 600 bp. Candidate sites were identified by the MatInspector program [17]. Unfortunately, currently it is not available; instead the Match program [18] is suggested. This program was applied to nine genes in the training sample with the original parameters of [16]. Unfortunately, for none of the genes the results could satisfy the set conditions. This is most probably caused by the difference in the work of Match and MatInspector.

Making the conclusions, one should note the following points:

we demonstrated clustering of sites for the global amino acid biosynthesis regulator Gcn4p,

analysis of 100 *a priori* nonregulated genes demonstrated nonrandomness of the observed clusters in the 5′-regions upstream of regulated genes,

clustering is retained in a distant organism, *C. albicans*.

ACKNOWLEDGMENTS

This study was partially supported by grants from HHMI (55000309) and LICR (CRDF RBO-1268).

REFERENCES

1. Hertel, K.J., Lynch, K.W., and Maniatis, T., *Current Opinions in Cell Biology,* 1997, vol. 9, pp. 350–357.
2. Kim, J.G., Takeda, Y., Matthews, B.W., and Anderson, W.F., *J. Mol. Biol*., 1987, vol. 196, pp. 149–158.
3. Khory, A.M., Lee, H.J., Lillis, M., and Lu, P., *Biochim Biophys. Acta,* 1990, vol. 1087, pp. 55–60.
4. Coleman, R.A. and Pugh, B.F., *J. Biol. Chem.,* 1995, vol. 270, pp. 13850–13859.
5. Hinnebusch, A.G. and Natarajan, K., *Eukaryotic Cell,* 2002, vol. 1, pp. 22–32.
6. Drysdale, C.M., Duenas, E., Jackson, B.M., Reusser, U., Braus, G.H., and Hinnesbusch, A.G., *Mol. Cell Biol.,* 1995, vol. 15, pp. 1220–1233.
7. Natarajan, K., Meyer, M.R., Jackson, B.M., Slade, D., Roberts, C., Hinnebusch, A.G., and Marton, M.J., *Mol. Cell Biol.,* 2001, vol. 21, pp. 4347–4368.
8. Mosch, H.-U., Scheier, B., Lahti, R., Mantsala, P., and Braus, G.H., *J. Biol. Chem.*, 1991, vol. 266, pp. 20453–20456.

9. Jones, E.W. and Fink, G.R., The molecular biology of the yeast Saccharomyces: methabolism and gene expression, Strathern, J.N., Jones, E.W., and Broach, J.R., Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor N.Y., 1982, pp. 181–300.

10. Lifanov, A.P., Makeev, V.J., Nazina, A.G., and Papatsenko, D.A., *Genome Res*., 2003, vol. 13, pp. 579–588.

11. Struhl, K., *Annu. Rev. Biochem*., 1989, vol. 58, pp. 1051–77.

12. Bruin, D., Zaman, Z., Liberatore, R.A., and Ptashne, M., *Nature,* 2001, vol. 409, pp. 109–113.

13. Mironov, A.A., Vinokurova, N.P., and Gelfand, M.S., 2000, Molecular Biology, vol. 34, no. 2, pp. 222–231.

14. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E., *Nucleic Acids Res.,* 2003, vol. 31, pp. 374–378.

15. Jia, M.H., Larossa, R.A., Lee, J.-M., Rafalski, A., Derose, E., Gonye, G., and Xue, Z., *Physiol. Genomics,* 2000, vol. 3, pp. 83–92.

16. Schuldiner, O., Yanover, C., and Benvenisty, N., *Curr. Genet.,* 1998, vol. 33, pp. 16–20.

17. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T., *Nucleic Acids Res.,* 1995, vol. 23, p. 4878.

18. Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E., *Nucleic Acids Res.,* 2003, vol. 31, p. 3576.