Distribution of Transcription Factor Binding Sites in the Yeast Genome Suggests Abundance of Coordinately Regulated Genes

Andreas Wagner
Distribution of transcription factor binding sites
in the yeast genome suggests abundance of coordinately
regulated genes.

Andreas Wagner

The University of New Mexico, Department of Biology

and

The Santa Fe Institute

Address:
The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, U.S.A.

Phone: +1-505-984-8800 Ext. 231; FAX: +1-505-982-0565;

E-mail: aw@santafe.edu
The availability of whole genomic DNA sequences makes it possible to analyze regulatory DNA regions on a genome-wide scale. Characterization of such regions will be crucial to understand how the “parts” of an organism cooperate to build the whole. This is a difficult undertaking, as the example of eukaryotic promoter (enhancer) regions shows. It is common practice among molecular biologists to search for binding sites of transcription factors (TFs) in the non-coding regions surrounding a gene, in order to form hypotheses about transcriptional regulation of the gene. This method often fails, because many TF binding sites are frequent in genomic DNA, and the occurrence of one binding site alone does not imply its functional significance. When analyzing multi-megabase regions for regulatory regions, one would hope that it is possible to devise more sensitive techniques that take advantage of the large amounts of statistical information extractable from whole genome sequences. For the analysis of regulatory regions, an important question in this regard is whether the genome-wide distribution of TF binding sites could be used to make inferences about transcriptional regulation of individual genes.

The transcription of many, if not most, eukaryotic genes is cooperatively regulated via one kind (homotypic cooperativity) or several kinds (heterotypic cooperativity) of TFs. Such cooperativity requires binding of transcriptional regulators to closely spaced binding sites in regulatory regions. While the exact spacing and order of TF binding sites is often irrelevant, the fact alone that TF binding sites are spaced more closely in these regions than expected “by chance alone” may be used to detect regulatory regions on a genome-wide scale. In other words, if a TF shows cooperativity for at least some of the genes it regulates, closely spaced TF binding sites in a region of the genome may indicate that a nearby gene is regulated by the TF, a hypothesis that can then be tested experimentally. This approach requires some statistical notion of statistically significant groups (“clusters”) of TF binding sites, that is, what does it mean to say that a group of binding sites is linked more tightly than expected by chance alone? Recently, a statistical technique for whole genome analysis was proposed (Wagner, 1997) that makes the notion of significant clustering of TF binding sites precise, and that allows detection of significant binding
site clusters in a genome. Given one or more TFs with well-defined binding sites, the method identifies the best candidate genes for regulation by the TFs in a genome, based on significant binding site clusters. When applied to the genome of *Saccharomyces cerevisiae* for homotypic site clusters, the technique detects binding site clusters (i) upstream of genes experimentally known to be regulated by the respective TFs, and (ii) upstream of genes that play a role in the same biochemical process as the TF (e.g., cell-cycle regulation), but for which it is not known whether the TF regulates their expression. Notably, while there can be thousands of binding sites for a specific TF in the genome, the number of significant clusters of binding sites is typically small (Wagner, 1997). When analyzing the distribution of individual sites belonging to such clusters, one finds that they occur with strong preference in non-coding regions, as would be expected for yeast TF binding sites that play a role in transcriptional regulation. Although base composition may be a rather poor predictor of site frequency (Karlin and Macken, 1991), it may be worth pointing out that this bias can not be explained on the basis of different base compositions in coding and non-coding regions. Further validation of the method is currently limited by published information on specific TFs and the genes they regulate, but will soon become easier as vast amounts of expression data from transcript arrays will become available for yeast.

While a technique like this can help generate hypotheses on particular TFs and genes they regulate, application to a larger sample of TFs can reveal patterns that can not be detected by looking at individual factors, but may reveal information about genome organization or global regulatory patterns. The information displayed in Table 1 shows a potential example of such a pattern. It is based on analysis of the yeast genome for homotypic binding site clusters of a larger sample of TFs than the previous study. TFs represented in this list were chosen from a public database (Wingender *et al.*, 1996) on the basis of two criteria for their DNA binding sites. First, their binding sites are among the best characterized in yeast. Second, their binding sites occur at an intermediate frequency (> 50 copies) in the genome. Very frequent binding sites, such as the heat shock factor (HSF) binding site with > 10⁵ occurrences were excluded, because they are closely spaced
throughout the entire genome. The resulting list of sites in Table 4 includes TFs involved in particular cellular processes (e.g., SBF: cell-cycle regulation; Ogas et al., 1991), but also general TFs involved in the regulation of many different kinds of genes (REB1; Ju et al., 1993), and factors that also have functions other than transcriptional regulation (e.g., CBF1, implicated in centromer function; Mellor et al., 1990).

The number of individual binding sites that are part of significant clusters is listed separately for coding and non-coding regions in columns 2 and 3 of Table 1. Given that approximately 72 percent of the yeast genome is protein-coding (Dujon, 1996), one would expect a 72:28 percent ratio of sites in coding and non-coding regions respectively. When testing the null-hypothesis that the observed number of sites is consistent with this expectation by a $\chi^2$-test, one finds that clusters for 12 out of the 17 sites occur with high preference in the non-coding regions (column 4). Five of the remaining TF binding sites show no significant bias (at $P = 0.01$), and one (GCR1) shows a bias towards occurrence in the coding regions. Computation of such global statistics on the genomic distribution of TF binding sites may provide a first indication that a TF displays homotypic cooperativity.

Columns 6 and 7 of Table 1 show the distribution of clusters located in non-coding regions with respect to the orientation of adjacent gene pairs. If a binding site cluster occurs in the non-coding region between two genes or (putative) open reading frames, the two genes can have four possible combinations of orientations. Both of them may be encoded on the Crick (C) strand or on the Watson (W) strand (WW or CC orientation), i.e., they are transcribed in the same direction. They may be convergently transcribed (WC orientation), in which case the cluster is located downstream of the transcription start site of both genes. Third, they can be divergently transcribed (CW orientation), in which the cluster lies upstream of both genes. Because transcriptional regulators in yeast function in general only when bound upstream of a gene, the TF which binds to its site(s) can regulate the transcription of only one gene if it lies between two genes in CC or WW orientation. It can not regulate expression of either gene in the WC combination, but it can coordinately regulate the expression of both genes in the CW combination. The orientation of individual
genes in the yeast genome does not show any obvious local correlation (Dujon, 1996; A. Wagner, unpubl.), i.e., the orientation of two consecutive genes in the yeast genome appears uncorrelated. The mean fraction of adjacent (putative) open reading frames in the four orientations averaged over all 16 yeast chromosomes is statistically indistinguishable from the ratio CC: CW: WC: WW = 1:1:1:1 (A. Wagner, unpubl.) One might expect a priori that TF binding site clusters are approximately randomly distributed among these four combinations. However, this is not the case. The first observed pattern regards the number of clusters between genes in WC orientation, where they are not likely to play a role in transcriptional regulation. For the 17 TFs listed here, a total of 85 homotypic clusters were found in non-coding regions. Only two of those (one each for DAL82 and REB1, a multi-functional TF) occur between genes in WC orientation ($\chi^2 = 23.25, 1 df$, significant at $P \ll 0.001$). The second pattern regards the distribution of clusters among genes in CW orientation on one hand (column 6), and in all other orientations on the other hand (column 7). Pooling information from all 17 TFs, one observes 54 clusters between genes in CW orientation vs. 31 clusters between genes in all other orientations, a highly significant ($\chi^2 = 67.30, 1 df, P \ll 0.001$) deviation from the null-hypothesis of a 1:3 ratio, under the assumption that clusters for different TF binding sites are not co-located. Thus, a clear bias is observed towards a high number of clusters between genes in CW orientation, where they could play a role in the coordinated regulation of both genes. Because the number of clusters for individual TFs is small, interpretation of any statistical test for individual TFs has to be approached with extreme caution. With this caveat in mind, an exact test (Sokal and Rohlf, 1981, Ch. 17) for the deviation from the null-hypothesis was carried out for those individual TFs whose binding sites showed more than five clusters (column 8). Four out of seven TFs for which an exact test was carried out show a deviation from the 1:3 ratio significant at $P < .01$. The other three show a non-significant deviation. However, the direction of this deviation is towards CW oriented genes, a pattern that also holds for seven of the 10 TFs for which less than five clusters were found. The remaining four TFs for which the pattern does not hold account for only five out of the 85 clusters observed.
Notably, three of these four TFs are ones for which the binding sites do not accumulate in non-coding regions.

It should be emphasized that the statistical signal detected here is likely to be seriously obscured by three factors. First, well-characterized binding sites are critical to the success of the technique. Some of the TF binding sites considered, although among the best characterized for yeast, are still not very well defined (e.g., through extensive mutagenesis). Second, heterotypic associations were not even considered. Third, non-cooperative regulation can not be detected by this approach. The robustness of the signal is therefore surprising. It does of course not imply biological significance. Experimental validation would involve testing the effect of changes in the activity of multiple TFs on the expression of multiple genes. While conventional techniques are not efficient at addressing such global questions about gene regulation, relevant data may become available soon with the ability to assess expression changes in thousands of genes simultaneously. If it is proven that at least some of the observed TF binding site clusters have a role in transcriptional regulation, one may have to conclude that gene pairs forming co-regulated units reminiscent of operons in prokaryotes are even more frequent in yeast than is currently appreciated (Olson, 1992). For the time being, the result suggests a possible avenue towards the identification of new functional yeast ORFs, namely to take a look at ORFs adjacent to any gene regulated by a known TF, if the respective pair of ORFs is in a head-to-head orientation.

LITERATURE CITED


Ju, Q., Morrow, B.E., Warner, J.R. (1993) *REB1, a yeast DNA-binding protein with many targets, is essential for cell growth and bears some resemblance to the oncogene myb*.  

6
Molecular and Cellular Biology 10, 5226-5234


