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

Andreas Wagner[1]

Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131; and The Santa Fe Institute, Santa Fe, New Mexico 87501

Received October 29, 1997; accepted March 10, 1998

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 noncoding regions surrounding a gene 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 indicate 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 are 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 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 (9) 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 (9). When analyzing the distribution of individual sites belonging to such clusters, one finds that they occur with strong preference in noncoding 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 (4), it may be worth pointing out that this bias cannot be explained on the basis of different base compositions in coding and noncoding 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 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 cannot 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 (10) 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 $>10^5$ occurrences, were excluded, because they are closely spaced throughout the entire genome. The resulting list of sites in Table 1 includes TFs involved in particular cellular processes (e.g., SBF, cell-cycle regulation; 6), but also general TFs involved in the regulation of many different kinds of genes (REB1; Ref. 3), and factors that also have functions other than transcriptional regulation (e.g., CBF1, implicated in centromere function; Ref. 5).

The number of individual binding sites that are part of significant clusters is listed separately for coding and noncoding regions in columns 2 and 3 of Table 1. Given that approximately 72% of the yeast genome is protein-coding (2), one would expect a 72:28 percentage ratio of sites in coding and noncoding regions, respectively. When testing the null hypothesis that the observed number of sites is consistent with this expectation by a χ^2 test, one finds that clusters for

[1] Address correspondence to author at University of New Mexico, Department of Biology, 167A Costetter Hall, Albuquerque, New Mexico 87131-1091. Telephone: (505) 984-8800, Ext. 231. Fax: (505) 982-0565. E-mail: aw@santafe.edu.

ARTICLE NO. GE985303
0888-7543/98 $25.00

TABLE 1

Distribution of Transcription Factor Binding Sites Belonging to Tightly Linked Binding Site Clusters

	No. sites belonging to significant clusters			No. clusters in noncoding region			
					By orientation of adjacent gene pair		
Transcription factor[a]	In coding region	In noncoding region	P value[b] (χ^2 test)	Total	CW	WW/CC/WC	P value[c] (exact test)
ABF1	73	63	1.94×10^{-6}	4	4	0	ND
CBF1	61	56	1.71×10^{-6}	6	5	1	4.9×10^{-3}
DAL82	92	56	7.68×10^{-3}	4	2	2	ND
yE2F	129	86	8.90×10^{-5}	8	6	2	4.5×10^{-3}
GCN4/yAP-1	139	49	5.50×10^{-1}	1	0	1	ND
GCR1	133	31	9.46×10^{-3}	1	0	1	ND
MBF	40	60	1.03×10^{-12}	11	7	4	8.1×10^{-3}
MIG1	106	77	2.22×10^{-5}	7	4	3	7.4×10^{-2}
PHO4	80	72	1.05×10^{-7}	9	8	1	1.2×10^{-4}
PPR1-dep. protein	77	34	5.38×10^{-1}	1	1	0	ND
R-CAR1	18	17	6.72×10^{-3}	1	1	0	ND
RAP1	99	204	$>10^{-14}$	9	6	3	1.1×10^{-2}
REB1	124	288	$>10^{-14}$	14	7	7	4.1×10^{-2}
ROX1	148	72	1.20×10^{-1}	2	1	1	ND
SBF	65	46	1.16×10^{-3}	2	0	2	ND
STE12	80	25	3.40×10^{-1}	4	2	2	ND
YEB3	68	20	2.70×10^{-1}	1	0	1	ND

[a] Consensus binding site (number of mismatches tolerated per site): ABF1, RTCRYBNNNNACG (0); CBF1, RTCACRTG (0); DAL82, GAAAATTGCGTT (2); yE2F, GCGCGAAA (1); GCN4/yAP-1, RRTGACTCA (1); GCR1, WNYNRNCWTCCWNWWK (1); MBF, ACGCGTNA (0); MIG1, WWWWWNSYGGGG (1); PHO4, CACGTG (0); PPR1, TTCGGNRNTYNCCGAA (2); R-CAR1, AGCCGCCGA (0); RAP1, WRMACCCATACAYY (3); REB1, CCGGGTAA (2); ROX1, YYYATTGTTCTC (2); SBF, CACGAAAA (0); STE12, TGAAACA (0); YEB3, CAGGTCATGTGGC (3). All consensus site are taken from Ref. (1) and from the TRANSFAC database (http:\\transfac.gbf-braunschweig.de/TRANSFAC/; release 3.0; Ref. 10). Allowed mismatches maximize site count, while ensuring an exponential intersite distance distribution in genomic DNA, as assessed by likelihood ratio and χ^2 goodness-of-fit tests to an exponential distribution.

[b] Based on a χ^2 test of the observed ratio of binding sites in coding:noncoding region against the expected ratio of 72:28 (Ref. 2). All significant deviations are in the direction of increased number of binding sites in noncoding regions, except for GCR1.

[c] Probability of observing a deviation from the expected 1:3 ratio equal to or larger than that observed, based on an exact test (one-tailed) for TFs with five or more clusters (Ref. 8, Chap. 17). ND. not determined, because number of clusters is too small. Chromosomal locations of individual clusters are available upon request.

12 of the 17 sites occur with high preference in the noncoding regions (column 4). Five of the remaining TF binding sites show no significant bias (at $P = 0.01$), and one site (GCR1) shows a bias toward 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 noncoding regions with respect to the orientation of adjacent gene pairs. If a binding site cluster occurs in the noncoding 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 case the cluster lies upstream of both genes. Because transcriptional regulators in yeast function in general only when bound upstream of a gene, the TF that 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 cannot 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 (2; A. Wagner, unpublished results), 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, unpublished results). 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 concerns 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 noncoding regions. Only 2 of those (one each for DAL82 and REB1, a multifunctional TF) occur between genes in WC orientation ($\chi^2 = 23.25$, 1 df, significant at $P \ll 0.001$). The second pattern concerns 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 ($\chi^2 = 67.30$, 1df, $P \ll 0.001$) deviation from the null hypothesis of a 1:3 ratio, under the assumption that clusters for different TF binding sites are not colocated. Thus, a clear bias is observed toward a large number of clusters between genes in the 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 must be approached with extreme caution. With this caveat in mind, an exact test (Ref. 8; Chap. 17) for the deviation from the null hypothesis was carried out for those individual TFs whose binding sites showed more than 5 clusters (column 8). Four of seven TFs for which an exact test was carried out show a deviation from the 1:3 ratio significant at $P < 0.01$. The other 3 show a nonsignificant deviation. However, the direction of this deviation is toward CW oriented genes, a pattern that also holds for 6 of the 10 TFs for which less than 5 clusters were found. The remaining 4 TFs for which the pattern does not hold account for only 5 of the 85 clusters observed. Notably, 3 of these 4 TFs are ones for which the binding sites do not accumulate in noncoding 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, noncooperative regulation cannot be detected by this approach. The robustness of the signal is therefore surprising. It does not, of course, 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 coregulated units reminiscent of operons in prokaryotes are even more frequent in yeast than is currently appreciated (7). For the time being, the result suggests a possible avenue toward 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.

REFERENCES

1. Dhawale, S. S., and Lane, A. C. (1993). Compilation of sequence-specific DNA-binding proteins implicated in transcriptional control of fungi. *Nucleic Acids Res.* **24:** 5537–5546.
2. Dujon, B. (1996). The yeast genome project: What did we learn? *Trends Genet.* **12:** 263–270.
3. Ju, Q., Morrow, B. E., and 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. *Mol. Cell. Biol.* **10:** 5226–5234.
4. Karlin, S., and Macken, C. (1991). Assessment of inhomogeneities in an *E. coli* physical map. *Nucleic Acids Res.* **19:** 4241–4246.
5. Mellor, J., Jiang, W., Funk, M., Rathjen, J., Barnes, C. A., Hinz, T., Hegemann, J. H., and Philippsen, P. (1990). CPF1, a yeast protein with functions in centromers and promoters. *EMBO J.* **9:** 4017–4026.
6. Ogas, J., Andrews, B. J., and Herskowitz, I. (1991). Transcriptional activation of CLN1, CLN2, and a putative new G1 cyclin (HCS26) by SWI4, a positive regulator of G1-specific transcription. *Cell* **66:** 1015–1026.
7. Olson, M. V. (1992). Genome structure and organization in *Saccharomyces cerevisiae. In* "The Molecular and Cellular Biology of the Yeast *Saccharomyces*" (E. W. Jones, J. R. Pringle, and J. R. Broach, Eds.), Cold Spring Harbor Laboratory Press, New York.
8. Sokal, R. R., and Rohlf, F. J. (1981). Biometry. Freeman, New York.
9. Wagner, A. (1997). A computational genomics approach to the identification of gene networks. *Nucleic Acids Res.* **25,** 3594–3604.
10. Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24,** 238–241.

Mapping of Two Mouse Membrane-Type Matrix Metalloproteinase (MT-MMP) Genes, *Mmp15* and *Mmp16,* to Mouse Chromosomes 8 and 4, Respectively

Michael F. Seldin,* Mark D. Gustavson,†,[1] and Suneel S. Apte†,[2]

**Rowe Program in Genetics, Department of Biological Chemistry and Department of Medicine, University of California, Davis, Davis, California 95616; and †Department of Biomedical Engineering, Lerner Research Institute, Cleveland Clinic Foundation, (Wb3), 9500 Euclid Avenue, Cleveland, Ohio 44195*

Received February 10, 1998; accepted April 7, 1998

The matrix metalloproteinases (MMPs)[3] are a family of proteolytic enzymes with broad specificity for components of the extracellular matrix and with important implications for devel-

The mouse genes encoding MMP-15 (MT2-MMP) and MMP-16 (MT3-MMP) are *Mmp15* and *Mmp16*, respectively. MMP15 and MMP16 are the corresponding human genes.

[1] Present address: Department of Cell Biology, Vanderbilt University, Nashville, TN 37232-2175.

[2] To whom correspondence should be addressed. Telephone: (216) 445-3278. Fax: (216) 445-4383. E-mail: aptes@bme.ri.ccf.org.

[3] Abbreviations used: MT-MMP, membrane-type matrix metalloproteinase; MMP, matrix metalloproteinase.

GENOMICS **50,** 295–297 (1998)
ARTICLE NO. GE985331
0888-7543/98 $25.00