Supplementary Information for

Transcriptional Regulatory Code of a Eukaryotic Genome

Christopher T. Harbison[1,2*], D. Benjamin Gordon[1*], Tong Ihn Lee[1], Nicola J. Rinaldi[1,2], Kenzie Macisaac[3], Timothy Danford[3], Nancy M. Hannett[1], Jean-Bosco Tagne[1], David B. Reynolds[1], Jane Yoo[1], Ezra G. Jennings[1], Julia Zeitlinger[1], Manolis Kellis[1], Alex Rolfe[3],Ken T. Takusagawa[3], David K. Gifford[3], Ernest Fraenkel[1,3†] and Richard A. Young[1,2†]

[*]These authors contributed equally to this work

[1]*Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA*

[2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[3]*MIT Computer Science and Artificial Intelligence Laboratory, 200 Technology Square, Cambridge, MA 02139, USA*

†To whom correspondence should be addressed:

RAY: E-mail young@wi.mit.edu, EF: E-mail efraenkel@wi.mit.edu

Transcriptional Regulatory Code of a Eukaryotic Genome

Various authors[1,2]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142

[2] MIT Computer Science and Artificial Intelligence Laboratory, 200 Technology Square, Cambridge, MA 02139

Motif Discovery

Overview

Putative binding motifs were identified by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data. The resulting specificity predictions were filtered for significance using uniform metrics and then clustered to yield representative motifs. For cases in which multiple significant binding motifs were found for a factor, information from specificity databases was used to identify which motifs represented the true binding preference.

Motif Discovery Programs

Motif Discovery Programs have different strengths with respect to finding specificities. To gain as comprehensive an analysis as possible, we applied five different motif-finding programs to the binding data: AlignACE[1], MEME[2], MDscan[3], the conservation-based method described in Kellis et al.[4] , and a new conservation-based method called CONVERGE (described below). The MEME program was also used to analyze a modified input that incorporated conservation information (see "Probe Sequences").

To make the search more thorough, we ran each of these programs multiple times with different parameters. AlignACE was run using the default settings ten times with different random number seeds, in order to increase the motif space it sampled. The results from the AlignACE runs were grouped together for analysis. MEME was run using the supplied 5th-order Markov background model, the "ZOOPS" motif model, and the "-minsites 20 -dna -revcomp" options. MEME runs were repeated using motif width ranges of 7 to 11 and 12 to 18. To run MDscan, seqeuences were ranked according the p-value of binding, and the program was run with the "-s 30 -r 5 -t 10" options. To compensate for the fact that MDscan searches only for motifs of fixed width, the program was run repeatedly, once with each width in the range 8 to 15 bases. The method of Kellis et. al was applied to the data as described[4]. CONVERGE was run twice using motif widths of 8 and 15.

MEME_c

We tested whether we could improve the performance of AlignACE, MEME and MDscan by modifying the input sequences to convey the conservation of each base in the *sensu stricto Saccharomyces* species. Using ClustalW[5] alignments for the *sensu stricto* species [4], we replaced a base in the *Saccharomyces* genome with the letter "N" if it was not conserved in 2/3 or 3/4 of the other genomes. Of the programs we tested, only MEME was able to use process the modified sequences.

CONVERGE

We designed CONVERGE to identify motifs that are both over-represented in a set of input sequences and conserved across multiple genomes. CONVERGE input sequences consists of an ungapped DNA sequence corresponding to the primary genome, as well as one or more optional aligned sequences, which may contain gaps. The algorithm is based on the ZOOPS model of MEME and uses a 5th-order Markov background model. However, whereas MEME searches for matches to a motif model across a set of input sequences, CONVERGE searches across the multiple-sequence alignments for each

sequence. Specifically, CONVERGE treats the probability of a motif occurring at a site in the alignment as the product of the probabilities of the motif occurring at the same site in each of the aligned sequences. Thus, CONVERGE defines a site as conserved in a flexible manner that depends on the motif being discovered. Full details will be presented elsewhere.

Probe Sequences

Motif discovery programs were applied to the sequences of probes bound with a p-value ≤ 0.001. We found that some intergenic regions were highly homologous over their entire length, and consequently skew the results of motif discovery since all subsequences are overrepresented. To remove this bias, we used BLAST[6] to identify pairs of probes with high sequence similarity over 50% of their lengths. For each pair, the shorter intergenic region was omitted from motif discovery computations.

To determine the sequences present on the microarrays, we computed the expected products of the PCR used to construct the arrays. Research Genetics primer sequences were obtained from http://www.resgen.com/products/YeIRP.php3 and the March 2002 revision of the yeast genome was obtained from SGD[7]. Probes that were predicted to amplify more than two different genomic sequences were omitted from the calculations. Twenty five probe sequences neighboring repetitive, non-transcribed features (e.g. telomeric repeats, X elements and Y' elements) were also omitted.

PSSM Representation

Motifs from all programs were converted to a standard position-specific scoring matrix (PSSM) for subsequent analysis. AlignACE and MDscan produce alignments of binding sites, and these were first converted into matrices representing the frequency of each base (A, C, G, T) at each position of the alignments. The method of Kellis et al. represents motifs as text strings containing ambiguity codes, which were also converted to matrices of frequencies. (For example, if a motif contained the letter "S" at a particular position, a value of 0.5 would be assigned to both "C" and "G.") The matrices of base frequencies were converted to probabilities and then were adjusted with 0.001 pseudo-counts in proportion to the 0^{th}-order background probabilities (A:T 0.31; G:C 0.19). Log-likelihood scores were computed by dividing the estimated probabilities by the background probability for each letter and computing the base-2 logarithm. CONVERGE and MEME both provide probability matrices, which were used directly.

Motif Scoring

We tested the significance of each motif by comparing how often it was found in the bound and unbound probes. To encapsulate different approaches to measuring motif over-representation, we employed three different metrics: Enrichment, ROC AUC, and for motifs discovered by the method described in Kellis et al., the "CC4" score. The enrichment score is a direct measure of the occurrence of a motif among bound probes compared to all possible gene targets, but does not distinguish between the number of motifs occurrences within each intergenic region. The ROC AUC metric is more sensitive to cases in which the number of motif occurrences is a distinguishing factor.

Finally, the CC4 metric provides a way to account for the importance of the conservation of the motif among bound probes.

Motif score significance P<0.001 thresholds for "Enrichment" and "ROC a.u.c." specificity metrics obtained from calculations on randomized selections of intergenic regions as described in Methods. Entries containing "n/a" denote that the empirical distribution was not normal. The threshold for the CC4 metric (4.95) is not dependent on the number of sequences.

Enrichment score
To obtain the enrichment score, the hypergeometric distribution was used to compare the frequency of the motif in the bound probes to that which would be expected if the intergenic regions were selected at random from the genome. A sequence was considered to contain a motif if it contained at least one or more sites scoring at least 70% of the maximum possible score of the matrix.

A p-value for the enrichment was computed according to the formula:

$$p = \sum_{i=b}^{\min(B,g)} \frac{\binom{B}{i}\binom{G-B}{g-i}}{\binom{G}{g}} \qquad (5)$$

where B is the number of bound intergenic regions and G is the total number of intergenic regions represented on the microarray (or the genome). The quantities b and g represent the number of intergenic regions of B and G matching the motif. The quantity $-log_{10}(p)$ is referred to as the enrichment score.

*ROC AUC (*Receiver Operating Characteristic Area Under Curve)
The ROC AUC refers to the area under a receiver operating characteristic curve which is assembled by ranking the sets of bound and unbound probes according to the number of motif matches they contain, and plotting the fractional rankings against each other. We used the method and code described by Clarke and Granek[8].

Conservation CC4
Motifs discovered using the method of Kellis et al.[4] were judged according to the CC4 metric, in which the occurrence of a conserved motif among the bound probes is compared to the expected ratio observed among all 3-gap-3 motifs in among the same set of bound probes. The binomial probability of the observed ratio was computed, and is reported as in terms of the equivalent z-score.

Motif Significance
We observed that motif discovery programs produce motifs with high over-representation metrics (such as “Enrichment” and “ROC AUC”) even when applied to random selections of intergenic regions. To identify the true motifs we converted the scores from each metric into the empirical probability that a motif with a similar score could be found by the same program in randomly selected sequences. We accepted only those motifs

with a p-value ≤ 0.001. To estimate these p-values we ran each program 50 times on randomly selected sequences on sets of 10, 20, 30, 40, 50, 60, 70, 80, 100, 120, 140, and 160 probes.
The observed scores from these random runs were parameterized by a normal distribution. The critical values equivalent to a p-value of 0.001 are provided in Supplementary Table S7 for each program and each metric. If the empirical distribution was not normal (p-value of the Shapiro-Wilk test < 0.001), the corresponding metric was not used to evaluate motifs generated by the relevant program for regulators with a similar number of bound probes.

For example, suppose a motif found by performing ten runs of AlignACE on 30 intergenic sequences had an enrichment score of 20. The relevant score distribution is obtained by performing ten runs of AlignACE on 50 randomly selected sets of 30 intergenic sequences. The distribution of enrichment scores has a mean of 14.1 and standard deviation of 2.1. Thus the significance of the motif is estimated as 0.002.

Inter-Motif Distance
We constructed a distance metric to aid in the comparison of motifs. The distance D between two aligned motifs "a" and "b" is defined as,

$$D(a,b) = \frac{1}{w}\sum_{i=1}^{w}\frac{1}{\sqrt{2}}\sum_{L\in\{ACGT\}}(a_{i,L} - b_{i,L})^2 \qquad (1)$$

where w is the motif width, and $a_{i,L}$ and $b_{i,L}$ are the estimated probabilities of observing base L at position i of motifs a and b, respectively. The normalizations by w and $\sqrt{2}$ facilitate the interpretation as a fractional distance. For example, a distance of 0.20 indicates that the two motifs differ by about 20%.

In practice, the optimal alignment of motifs is not known. We therefore use the minimum distance between motifs among all alignments in which the motifs overlap by at least seven bases, or when the motifs are shorter, by 2 bases fewer than the shortest motif length. Alignments to the reverse complements of the motifs are included.

Motif Clustering
The set of significant motifs for each experiment was then clustered via k-medoids clustering[9] using the distance metric above. The k-medoids algorithm was performed 500 times to find a clustering with a minimal sum of inter-cluster distances. To find the optimal number of clusters, this process was first performed with 10 clusters, and then repeated with incrementally fewer clusters until all average distances between members of a cluster and medoids of other clusters were sufficiently large (greater or equal to 0.18).

Motif Averaging
A single motif representing each cluster was computed by averaging the probabilities at each matrix position of the aligned motifs comprising the cluster.

Motif Assignment
Often, motif discovery calculations produced several significant distinct motifs (3, on average). These motifs could represent the desired binding specificity of the protein, or they might arise from the specificity of binding partners or have other biological significance. To identify those motifs represent the binding specificity of the profiled transcription factor, we compared the specificities to binding data in the Transfac[10], YPD[11], and SCPD[12] databases. Motifs were also checked for similarity to known specificities of factors other than the profiled regulator that were found to bind a significant number of common targets ($p<10^{-12}$ by hypergeometric distribution).

Specificity data from these databases is sometimes available in the forms of raw sequences, ambiguity codes, and matrices. For each factor, we assembled a single consensus sequence to represent the body of experimentally determined specificity information and converted it to a PSSM as described above. A motif was considered to match the known PSSM if the distance (as described above) between them was less than 0.24.

Binding Site Map
Binding motifs were fused with location analysis data and conservation data to produce a map of active binding sites in intergenic regions. The map was constructed by finding all conserved occurrences of each motif within intergenic regions bound by the corresponding factor. For example, of the binding site of Bas1 (TGACTC) has 2368 matching occurrences in intergenic regions, and only 34 of these are conserved and found in intergenic regions bound by Bas1 in either rich media or starvation conditions.

We considered a sequence a match to a motif if it had a score of at least 60% of the motif maximum. We defined "conserved" to mean that the aligned sequence of at least two other *sensu stricto* species also matched the motif. In cases where fewer than two aligned sequences were available, the site was treated as "not conserved."

Conservation Test for Averaged Motifs
When the significance of a motif was close to our threshold, it was eliminated if it had fewer than three conserved instances among bound intergenic regions and at least 20 were bound.

Promoter Classification
Promoters were classified based on the aggregate binding data from all experiments. A promoter was defined as having multiple regulator architecture if more than one regulator bound in the aggregate data, regardless of the number of regulators that bound in any particular condition. Similarly, a promoter was assigned to the single regulator architecture if it was bound by exactly one regulator in the aggregate data.

Regulators that had a tendency to use the repetitive motif architecture were identified by χ^2 analysis. For each regulator, we calculated the number of promoters bound using a single site and the number using multiple sites. These values were then compared to the expected values based on the average for all factors.

1. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45 (1998).
2. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-9 (1995).
3. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* **20**, 835-9 (2002).
4. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).
5. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
7. Dolinski, K. et al.
8. Clarke, N. D. & Granek, J. A. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics* **19**, 212-8 (2003).
9. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of Statistical Learning; Data mining, inference and prediction* (Springer-Verlag, New York, 2001).
10. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-8 (2003).
11. Csank, C. et al. Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol* **350**, 347-73 (2002).
12. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-11 (1999).

Supplementary Table 7.
Motif score significance cutoffs (P ≤ 0.001)

Enrichment Score

Number of sequences	Converge	AlignACE	MDscan	MEME	MEME(c)
10	12.70	20.32	11.78	13.54	n/a
20	11.96	21.14	12.95	12.89	9.81
30	11.43	20.43	13.30	12.57	n/a
40	11.34	20.62	14.04	11.64	7.53
50	10.74	19.94	12.23	12.81	7.43
60	10.50	19.71	10.95	12.37	n/a
70	10.34	18.30	13.25	11.34	n/a
80	10.20	19.40	12.84	11.93	n/a
100	9.36	20.31	11.56	10.58	2.91
120	n/a	18.59	13.14	10.94	n/a
140	8.14	18.52	11.26	10.87	n/a
160	n/a	20.04	11.38	9.77	n/a

ROC a.u.c.

Number of sequences	Converge	AlignACE	MDscan	MEME	MEME(c)
10	n/a	n/a	n/a	n/a	n/a
20	0.812	0.842	0.857	0.925	n/a
30	0.758	0.773	0.793	0.831	0.785
40	0.720	0.713	0.758	0.764	0.737
50	0.687	0.674	0.719	0.737	0.711
60	0.670	0.662	0.688	0.706	0.654
70	0.663	0.641	0.686	0.684	0.664
80	0.643	0.626	0.670	0.675	0.648
100	0.634	0.615	0.664	0.633	0.606
120	0.624	0.604	0.629	0.624	0.602
140	0.608	n/a	0.634	n/a	0.590
160	0.594	0.580	0.613	0.593	0.588