Using MotifScan to Discover Transcription Factor Binding Sites in Non-human Genomes

Jason Harris
jharris@30doradus.org

Biochemistry 218 Final project

1 Introduction

Transcription Factor Binding Sites (TFBSs) are short DNA sequences (typically 5-20 bp) upstream of protein-coding genes that bind transcription factor proteins, which play an important role in regulating the expression of the gene (Pennacchio and Rubin 2001). Several algorithms have been published to perform *de novo* identification of TFBSs; e.g., Consensus (Hertz et al. 1990), MEME (Bailey and Elkan 1994), MDScan (Liu et al. 2002), MotifCut (Fratkin et al. 2006), and BioProspector (Liu et al. 2001). These methods are generally based on identification of conserved blocks of nucleotides in the upstream sequences of a group of genes in a single species which are known to be co-expressed. While these methods have been successfully applied to simple organisms like bacteria and yeast, *de novo* discovery of regulatory motifs is significantly more difficult in higher eukaryotes (such as *mammalia*), because the expression relationships among genes are complex, and the upstream intervals over which TFBSs can be found are very long.

Comparative genomics can provide leverage into the challenging problem of identifying TFBS motifs in the genomes of more complex species. Liu et al. (2004) introduced CompareProspector, which can identify regulatory motifs in the genomes of non-human species by first collecting known TFBSs in the human genome, and then searching for those known motifs in regions of the target genome that are highly conserved with respect to the human genome.

In this paper, we introduce a new comparative genomics technique for identifying regulatory sequences in non-human genomes. Like CompareProspector, our method starts with a sample of known TFBS motifs, and searches for k-mers in the non-human genomes that match the motif patterns of the known TFBSs. Unlike CompareProspector, we do not restrict our search to highly conserved sections of the genome; we uniformly select a 10-kbp upstream sequence from an orthologous gene in each genome, and use the graph-based MotifScan algorithm (Naughton et al. 2006) to search for novel TFBS instances. As a pilot application to demonstrate the method, we examine regulatory segments for the dopamine receptor D2 gene (DRD2), which has been identified in the genomes of many species. We use a BLAST search to identify the DRD2 gene in 18 non-human mammalian species, and obtain a 10-kbp sequence of the genome upstream from each orthologous gene from the UCSC Genome Browser (http://genome.ucsc.edu). Next, we collect 271 known TFBS motifs from the JASPAR project (Bryne et al. 2008), and use a MATLAB implementation of MotifScan to identify a subset of these that are statistically enriched in the upstream sequence of DRD2_HUMAN. Finally, we apply the MotifScan code

to search for novel instances of the human-enriched TFBS motif patterns in each of the non-human upstream sequences.

2 Implementation of MotifScan

I implemented the MotifScan algorithm described by Naughton et al. (2006) in MATLAB. MotifScan's function is to identify new instances of a known motif pattern, using a graph-based method which does pairwise comparison of the candidate k-mer with each member of the pattern, rather than employing a model which averages over the sequences in the pattern, such as a position-specific scoring matrix (PSSM). As argued by Naughton et al., models based on averaging inevitably wash out the true diversity present in the motif pattern, which can lead to false-negative rejection of true members of the pattern.

The MATLAB implementation of MotifScan takes advantage of the matrix arithmetic built into the language to provide an orders-of-magnitude speed improvement, compared to a straightforward implementation based on nested for-loops.

2.1 Obtaining the TFBS motif patterns

The JASPAR database (http://jaspar.genereg.net) is a curated and annotated collection of hundreds of known TFBSs (Bryne et al. 2008). While the emphasis of JASPAR is to present PSSM models for each motif, they do also make available the raw aligned sequences of all known instances in each motif pattern. I downloaded these "sites" files for 271 vertebrate TFBSs to serve as the definition of the motifs in my MotifScan analysis. The MATLAB parses these JASPAR "sites" files, extracting for each motif: (1) a matrix of the *unique* instances in the motif pattern, (2) a simple PSSM based on counting the fractional representation of each nucleotide at each position (which is not used in the present analysis), and (3) a vector which holds the number of times each unique instance was represented in the original pattern.

2.2 Obtaining the upstream DNA sequences

DNA sequences for 19 mammalian species (including *Homo sapiens*) were obtained using the UCSC genome browser (http://genome.ucsc.edu). After loading the genome for a particular species, the genome can be searched for any gene of interest. In the resulting genome map view, one can download the sequence data by first clicking on the RefSeq row for the gene, and then clicking on "Genomic sequence from assembly" in the page that follows. That opens a page entitled "Get Genomic Sequence Near Gene", from which one can download the DNA sequence encoding the gene, including upstream and downstream sequences. Sequences downloaded for this study each included 10kbp upstream.

The MATLAB code parses each of the FASTA-format sequences obtained from the above procedure, extracting: (1) the numerically-encoded 10kbp upstream sequence (A=1, C=2, G=3, T=4), (2) the name or number of the chromosome on which the sequence is found, (3) the starting chromosomal index for the upstream sequence, and (4) a flag indicating from which DNA strand the sequence was read (+1 = forward strand, -1 = reverse strand). This flag is important for getting the indexing right, because on the reverse strand, the chromosomal index *decreases* as you read forward in the sequence.

2.3 The MotifScan algorithm

The core of the MotifScan implementation takes the parsed TFBS and the upstream-sequence data structures as input, and searches for matches to each instance in each TFBS using the MotifScan graph-based scoring algorithm. At its core, the MotifScan algorithm assigns a score to any k-mer which indicates the likelihood that it is an instance of a given motif pattern. The score is described by Equation 3 from Naughton et al.:

$$\text{Score} = \sum_{i=1}^{N} \Theta_{\text{SS}}^{d} \Theta_{\text{NS}(b1,b2)} \sum_{j=1}^{n_i} \Theta_{\text{IK}}^{j}.$$

N is the total number of unique instances ("identity groups" in Naughton et al.) in the motif pattern. Note that in practice, we search each unique instance in the pattern as well as its inverse-complement, to account for the possibility that the TFBS is located on the opposite strand. d is the Hamming distance between the candidate k-mer and the *i*th instance in the motif pattern. n_i is the number of instances in the current identity group (i.e., the number of instances in the pattern that have the identical sequence). ThetaSS and ThetaIK are tunable parameters whose values are intended to be somewhere between zero and one. ThetaSS controls how strongly dissimilarity is penalized in the score. In other words, when ThetaSS has a smaller value, then a candidate k-mer with a larger Hamming distance will have a lower score. ThetaIK controls how strongly preference is given to matches with identity groups that have many members. In my implementation, I simply set both ThetaSS and ThetaIK to 0.5. ThetaNS(b1,b2) is a value representing the substitution matrix for JASPAR motifs (Table 1c in Naughton et al.). When d=1 (i.e., the candidate k-mer is different from the current motif instance at only one position), ThetaNS is simply the appropriate element from the substitution matrix. When d>1, we use the average substitution value from among those represented in the candidate k-mer comparison. If d=0, we simply set ThetaNS to 1.0.

Our MotifScan implementation assigns the score described above to each possible k-mer in the input sequence, starting with the k-mer covering positions 1 to k, then 2 to k+1, then 3 to k+2, and so on until we reach the end of the input sequence. We then need a way to determine statistical significance from the measured scores. Naughton et al. discuss several strategies for this, but the method I chose is based on randomly resampling the upstream sequence, and then analyzing the randomized sequence against each TFBS pattern with MotifScan. When the upstream sequence includes instances of a given motif, we expect to find a higher rate of occurrence of high-scoring sites than in the randomized sequence. We can evaluate the statistical significance of TFBS enrichment in a given sequence in several ways. Below, I will use a simple Kolmogorov-Smirnov test to determine whether the MotifScan score distributions are different between the real sequence and the resampled sequence. In addition, I will measure the "enrichment fraction": the fraction of MotifScan scores in the real sequence, which are larger than the 95th percentile MotifScan score in the randomized sequence.

3 Pilot application of the method: Dopamine Receptor D2

3.1 Identification of known TFBSs upstream of DRD2_HUMAN

I applied the MotifScan analysis to the 10kbp sequence upstream of the DRD2_HUMAN gene, using the full set of 271 vertebrate TFBS patterns from JASPAR. Of these, 68 are labeled as Human TFBSs in the JASPAR database, and four (NF-kappaB; CREB1; Egr-1; and NRSF form 1, named REST in JASPAR) are known TFBSs for DRD2_HUMAN, according to genecards.org and sabiosciences.com. Figures 1 a-d show the MotifScan scores for all k-mers in the upstream sequence, when using one of the four known TFBSs for DRD2_HUMAN. The MotifScan scores for all k-mers in the randomized sequence are also shown, for reference.

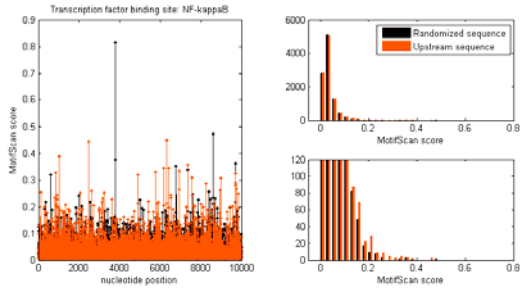


Figure 1a: Left: the MotifScan score for all k-mers in the upstream sequence for DRD2_HUMAN (orange) and in a randomly resampled sequence (black), using a TFBS motif that is known to be associated with DRD2_HUMAN (NF-kappaB). Right, top: the distribution of MotifScan scores for the upstream and randomized sequences. Right, bottom: the same distribution with the y-axis expanded to show the high-score tail.

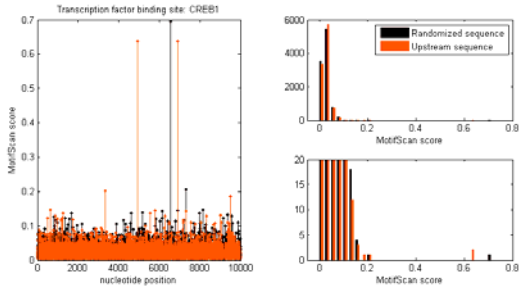


Figure 1b: Same as Figure 1, but for a different TFBS motif that is known to be associated with DRD2_HUMAN (CREB1).

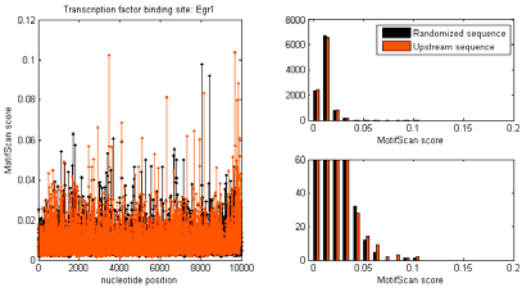


Figure 1c: Same as Figure 1, but for a different TFBS motif that is known to be associated with DRD2_HUMAN (Egr1).

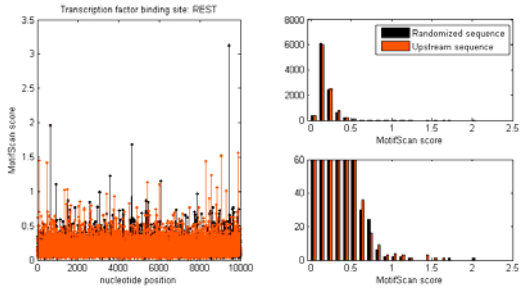


Figure 1d: Same as Figure 1, but for a different TFBS motif that is known to be associated with DRD2_HUMAN (NRSF form 1/REST).

The differences in MotifScan score distribution between the real upstream sequences and the randomized sequences are subtle. While we generally cannot point to specific upstream sites as being unequivocally identified as a TFBS instance, we can say whether the high-score tail of the score distribution is statistically overabundant with respect to the randomized sequence. A Kolmogorov-Smirnov test of the four pairs of distributions in Figure 1 shows that NF-kappaB and Egr1 are not significantly enriched in the upstream sequence (p-values of 0.89 and 0.08, respectively), while CREB1 and REST do appear to be enriched (p-values of 0.01 and 0.006).

We now turn our attention to the full set of 271 vertebrate TFBSs and ask whether any of these appears to be significantly enriched in the DRD2_HUMAN upstream sequence. Figure 2 summarizes the overabundance of high-scoring k-mers in the upstream sequence for DRD2_HUMAN, using all of the JASPAR TFBS motifs. It shows, for each of the binding-site motifs, the fraction of MotifScan scores from the upstream sequence which were larger than the 95th percentile score from the randomized sequence. If the two MotifScan score distributions are identical, the plotted value should be 0.05, modulo sampling noise.

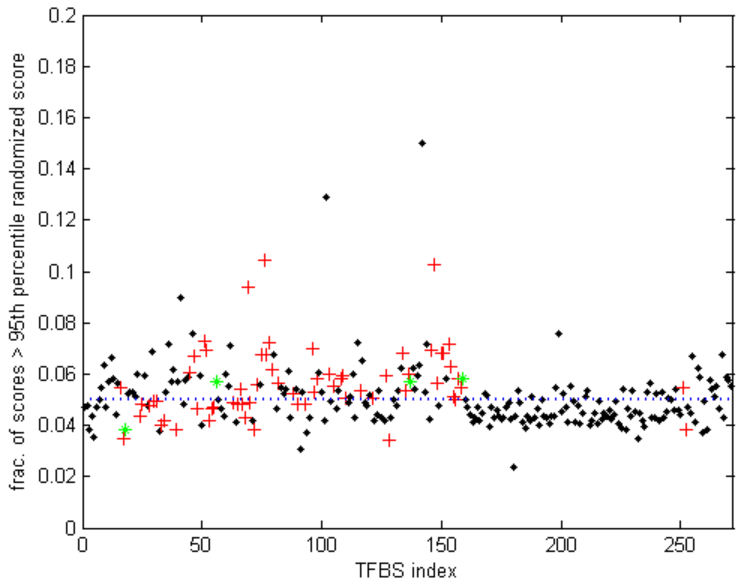


Figure 2: For each TFBS, the fraction of MotifScan scores from the real upstream sequence, which exceeded the 95th percentile score from the randomized sequence. TFBSs which were labeled Homo sapiens at sabiosciences.com are shown with red crosses, except for the four TFBSs known to be associated with DRD2_HUMAN, which are shown as green stars. An unenriched motif will have a value near 0.05, indicated with a blue dotted line.

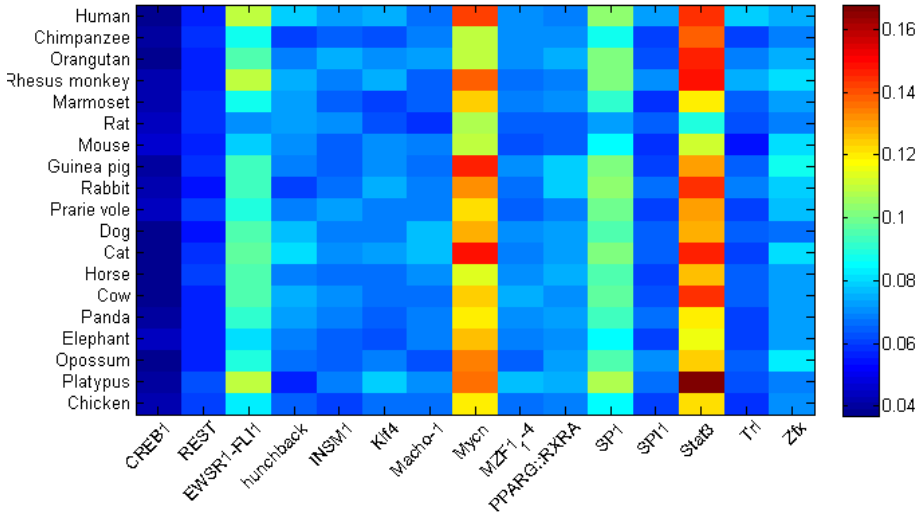
From Figure 2, we selected the 13 TFBSs with the largest enrichment fractions, and we also include the two TFBSs from our known-associated sample that had significant p-values in our K-S tests (CREB1 and REST). In the next section, we will determine whether these 15 TFBSs are also enriched in the genomes of 18 mammalian species, in the 10kbp segment upstream of each DRD2 gene. See Table 1 for a summary of the 15 TFBSs used in this study. Appendix A plots the full MotifScan score distributions for these motifs.

3.2 Discovery of TFBSs upstream of non-human DRD2 genes

Table 2 lists the 17 non-human mammals (plus one non-mammal: *Gallus gallus*) for which I downloaded a 10kbp sequence upstream of the DRD2 gene from the UCSC Genome Browser. I performed MotifScan analysis on each of these sequences using each of the 15 TFBS motifs which were found to be enriched in the human sequence in the previous section. The resulting enrichment fractions are shown as a heat map in Figure 3. Each row in the heat map shows the enrichment fraction pattern for the 15 TFBS

motifs in a particular species. The rows are ordered in (approximately) increasing phylogenetic distance from *Homo sapiens*.

The most significant feature in the heatmap is that the four TFBS patterns that were most overabundant in the Human sequence are generally the most overabundant in the other species as well. We do not detect any significant trends across rows in Figure 3 that would suggest evolutionary relationships reflected in the distribution of TFBS instances in the genome.



4 Summary and Discussion

Using a MATLAB implementation of the MotifScan algorithm, we identified 15 TFBS motifs (out of a pool of 271) that were statistically enriched in the 10-kbp sequence upstream of DRD2_HUMAN, and then examined the enrichment fractions of these motifs in the sequences upstream of DRD2 in 18 non-human genomes. We found that the pattern of TFBS enrichment in the human genome was remarkably well conserved across the other species' DRD2 upstream sequences; while there is some inter-species variation in the enrichment fractions, it is not obviously correlated with the phylogenetic relationships of these organisms.

Still, the pilot application explored here was fairly narrowly targeted. A follow-up study with a wider scope might yield more fruitful results. For example, one possible explanation for the lack of a strong detection of any of the four known-associated TFBS motifs in the upstream sequence for DRD2_HUMAN might simply be that 10 kbp is too short to capture the full complement of regulatory sites for this gene. Similarly, we could have included downstream sequences as well as upstream. It would also be informative to examine the enrichment patterns of all 271 TFBS motifs across all 19 genomes, rather than just a selection of 15 that were found to be enriched in the upstream sequence for DRD2_HUMAN. Finally, the selection of DRD2_HUMAN was fairly arbitrary for this pilot project. A more careful selection might try to identify a gene which exhibits a wider variety in expression or biological function across the target species, under the expectation that more expression diversity implies more diversity in the pattern of regulatory TFBSs.

Table 1: 15 TFBS motifs that were found to be enriched in the 10kbp sequence upstream of DRD2_HUMAN

TFBS name	n-mer	KS-test p-value
CREB1[a]	8	0.01
EWSR1-FLI1	17	$1x10^{-29}$
hunchback	10	$3x10^{-6}$
INSM1	12	$1x10^{-5}$
Klf4	10	$1x10^{-7}$
Macho-1	9	$4x10^{-6}$
Mycn	26	$2x10^{-66}$
MZF11-4	6	$8x10^{-10}$
PPARG:RXRA	15	$4.x10^{-11}$
NRSF form 1/REST[a]	11	0.006
SP1	10	$4x10^{-25}$
SPI1	7	0.0004
Stat3	19	$8x10^{-72}$
Trl	10	$2x10^{-6}$
Zfx	20	0.0004

[a]:known to be associated with DRD2_HUMAN

Table 2: Mammal organisms included in the MotifScan DRD2 analysis

Species (common name)	Notes
Ailuropoda melanoleuca (panda)	D2HZY0_AILME gene identified from BLAST search
Bos taurus (cow)	DRD2_BOVIN gene identified from BLAST search
Callithrix jacchus (marmoset)	Upstream sequence obtained from alignment with DRD2 gene in other species
Canis Familiaris (dog)	DRD2_CANFA gene identified from BLAST search
Cavia porcellus (guinea pig)	Upstream sequence obtained from alignment with DRD2 gene in other species
Equus caballus (horse)	Upstream sequence obtained from alignment with DRD2 gene in other species
Felis catus (cat)	Upstream sequence obtained from alignment with DRD2 gene in other species
Gallus gallus (chicken)	A9YZQ5_CHICK gene identified from BLAST search
Loxodonta africana (elephant)	Upstream sequence obtained from alignment with DRD2 gene in other species
Macaca mulatta (rhesus monkey)	Upstream sequence obtained from alignment with DRD2 gene in other species
Microtus ochrogaster (prairie vole)	E0V889_MICOH gene identified from BLAST search
Monodelphis domestica (opossum)	Upstream sequence obtained from alignment with DRD2 gene in other species
Mus musculus (mouse)	DRD2_MOUSE gene identified from BLAST search
Ornithorhynchus anatinus (platypus)	Upstream sequence obtained from alignment with DRD2 gene in other species
Oryctolagus cuniculus (rabbit)	Upstream sequence obtained from alignment with DRD2 gene in other species
Pan troglodytes (chimpanzee)	DRD2_PANTR gene identified from BLAST search
Pongo Abelii (orangutan)	Upstream sequence obtained from alignment with DRD2 gene in other species
Rattus norvegicus (rat)	DRD2_RAT gene identified from BLAST search

References:

Bailey, T. L. and Elkan C. *Proc. Int. Conf. Intelligent Systems Molecular Biology* 1994 2:28

Bryne J. C., Valen E., Tang M. H., Marstrand T., Winther O., da Piedade I., Krogh A., Lenhard B., and Sandelin A. *Nucleic Acids Res.* 2008 36:D102

Fratkin, E., Naughton, B. T., Brutlag, D. L. and Batzoglou, S. In *Proceedings of International Conference on Intelligent Systems and Molecular Biology* 2006

Hertz, G. Z., Hartzell III, G.W., and Stormo, G.D. *Comput. Appl. Biosci.* 1990 6:81

Liu, X., Brutlag, D.L., and Liu, J.S. *Pac. Symp. Biocomput.* 2001 127–138

Liu X., Brutlag D. L., and Liu J. S. *Nat Biotechnol.* 2002 20(8):835.

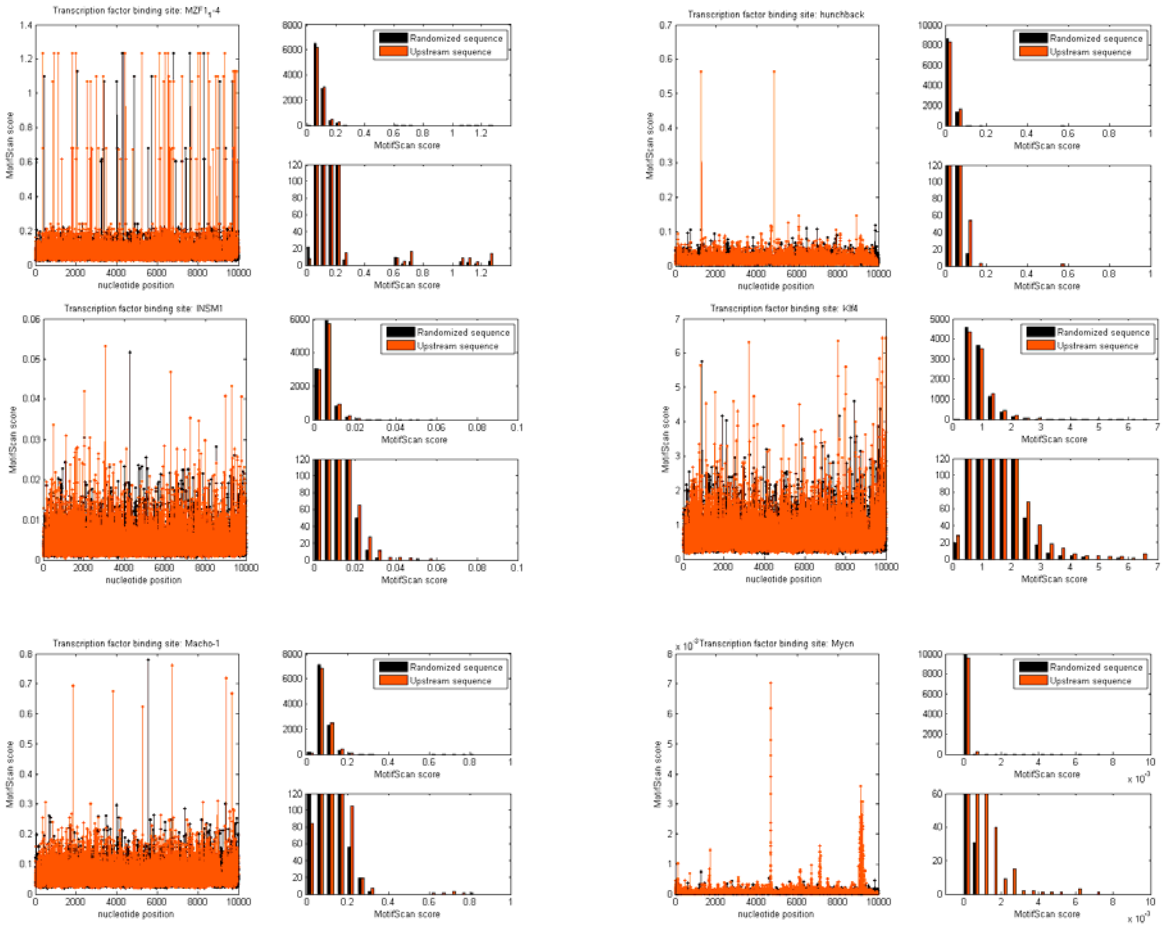
Liu, Y., Liu, X. S., Wei, L., Altman, R. B., and Batzoglou, S. *Genome Research* 2004 14:451

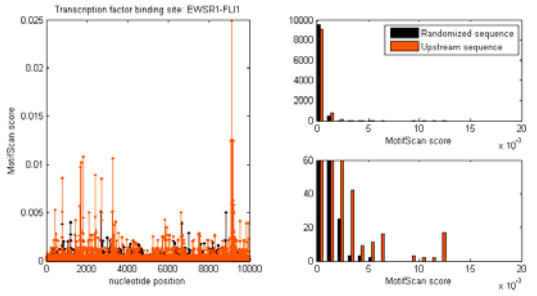
Naughton, B. T., Fratkin, E., Batzoglou, S., and Brutlag, D. L. *Nucleic Acids Res.* 2006 34:5730

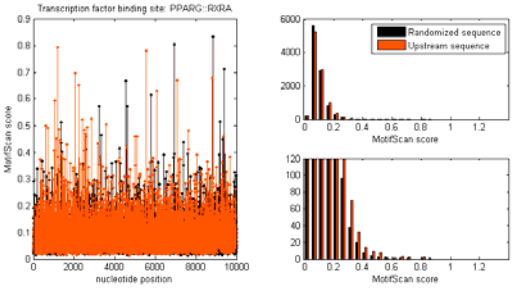
Pennacchio, L.A. and Rubin, E.M. *Nat. Rev. Genet.* 2001 2:100

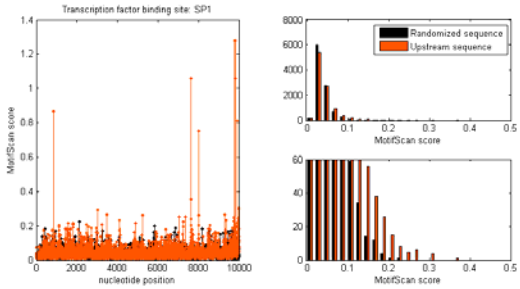
Appendix A:

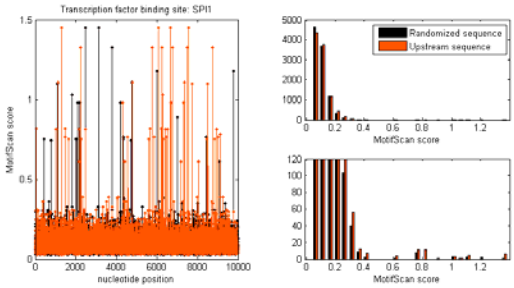
MotifScan score distributions for 13 TFBS motifs which were identified as enriched in the 10kbp sequence upstream of DRD2_HUMAN.

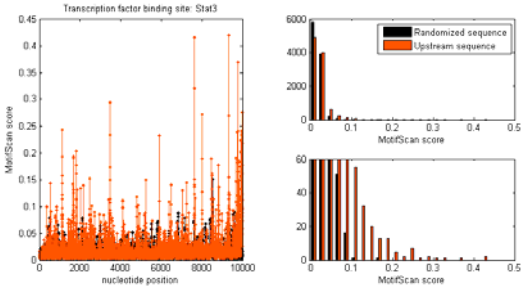


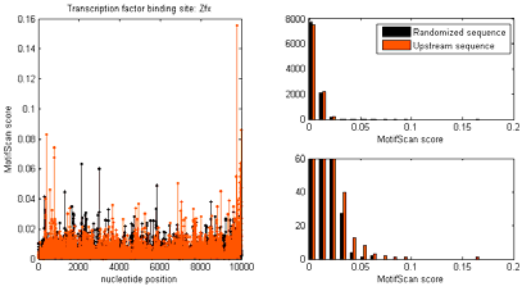

Transcription factor binding site: EWSR1-FLI1
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score
x 10^-3

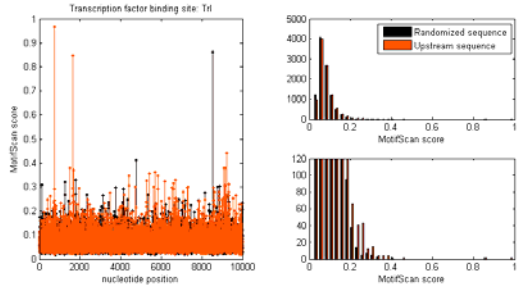

Transcription factor binding site: PPARG::RXRA
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score


Transcription factor binding site: SP1
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score


Transcription factor binding site: SPI1
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score


Transcription factor binding site: Stat3
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score


Transcription factor binding site: Zfx
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score


Transcription factor binding site: Tal
MotifScan score
nucleotide position
Randomized sequence
Upstream sequence
MotifScan score