Characterization of the Effects of TF Binding Site Variations on Gene Expression Towards Predicting the Functional Outcomes of Regulatory SNPs

Michal Lapidot and Yitzhak Pilpel

Department of Molecular Genetics, Weizmann Institute of Science,
Rehovot, 76100, Israel
michal.lapidot@weizmann.ac.il, pilpel@weizmann.ac.il
http://longitude.weizmann.ac.il/

Abstract. This work addresses a central question in medical genetics – the distinction between disease-causing SNPs and neutral variations. Unlike previous studies that focused mainly on coding SNPs, our efforts were centered around variations in regulatory regions and specifically within transcription factor (TF) binding sites. We have compiled a comprehensive collection of genome wide TF binding sites and developed computational measures to estimate the effects of binding site variations on the expression profiles of the regulated genes. Applying these measures to binding sites of known TFs, we were able to make predictions that were in line with published experimental evidence and with structural data on DNA-protein interactions. We attempted to generalize the properties of expression-altering substitutions by accumulating statistics from many substitutions across multiple binding sites. We found that in the yeast genome substitutions that abolish a G or a C are on average more severe than substitutions that abolish an A or a T. This may be attributed to the low GC content of the yeast genome, in which G and C may be important for conferring specificity. We found additional factors that are correlated with the severity of a substitution. Such factors can be integrated in order to create a set of rules for the prioritization of regulatory SNPs according to their disease-causing potential.

1 Introduction

The identification of disease-causing mutations is a central objective of medical genetics. So far most efforts to distinguish disease-causing single nucleotide polymorphisms (SNPs) from neutral variations have focused on coding SNPs [1-7]. Regulatory region variants are also known to cause diseases through altering the expression profiles of their downstream genes [8, 9]. Estimates show that the human population contains thousands of *cis*-regulatory variations [10]. Such high numbers set a clear need for the development of computational means for the identification of potentially deleterious regulatory SNPs. The present work lays the foundations for the development of such tools. It does not yet address actual SNPs, but rather studies a comprehensive collection of genome wide transcription factor binding sites (TFBS), in order to characterize the effects of binding site variations on the expression profiles of the regulated genes. TFBS

E. Eskin et al. (Eds.): RECOMB 2005 Ws on Regulatory Genomics, LNBI 4023, pp. 51 – 61, 2007.

are short (~6-20 bases) redundant sequences. Transcription factors (TFs) recognize a range of binding sites that may differ at several positions. Substituting a single position within a binding site may thus, in many cases maintain the site in the recognition domain of the same TF. There is however a possibility that the substitution will result in binding site loss or in the acquisition of a binding site recognized by another TF (Figure 1 left panel). The aim of this work was to develop computational methods for distinguishing between these three possible scenarios without the need to systematically mutate each binding site. These methods were first applied to individual binding sites, and the results were generalized to deduce universal properties of expression-altering substitutions. We focused here on the *Saccharomyces cerevisiae* (*SC*) genome because vast knowledge already exists regarding TFBS in this organism; however the presented work can be easily applied to other organisms and specifically to human.

As a first step, we have compiled a comprehensive dataset of TFBS in the *SC* genome ([11], Lapidot et al in prep.). Each binding site was defined by its DNA sequence (its syntax) and assigned a likely regulatory function (its semantics), in terms of the expression profiles of the genes it controls and the experimental conditions in which it operates. Our set had a good coverage of a recently published TFBS set ([12]), as well as many novel binding sites (see [11] for more details). An analysis of this binding site collection revealed a non trivial relation between syntax and semantics: Binding sites with similar syntax may yield different expression patterns of the regulated genes, or operate at different conditions (differ on the semantic level), whereas binding sites with different syntax can have similar semantics (i.e. dictate similar expression patterns). We next developed computational measures to estimate the semantic consequence of substituting a single binding site position. We applied these measures to binding sites of known TFs and were able to make predictions that were in line with published experimental evidence and with structural data on DNA-protein interactions. We further attempted to generalize the properties of expression-altering substitutions by accumulating statistics from many substitutions across multiple binding sites. Finally we tested out additional factors that are correlated with the severity of a substitution, such as the Information Content (IC) of the substituted position. These factors can be integrated to form a prioritization scheme that will enable the prediction of potentially deleterious regulatory SNPs.

2 Results

2.1 Compiling a Comprehensive TFBS Collection

This study was conducted in the *SC* genome, for which vast TFBS knowledge already exists. However in order to both broaden this knowledge and form a quantifiable connection between binding site sequence and the expression profiles of the regulated genes, we compiled our own comprehensive dataset. This dataset is unbiased by prior knowledge and is based on the premise that any nucleotide sequence that resides in a promoter of a gene may contribute to its expression regulation. We applied the previously described Expression Coherence (EC) score [13-15] to assess the effect of a promoter sequence motif on the related gene's expression profile. The EC score measures the extent to which a set of genes (in this case the set is defined by a common motif sequence) display similar expression profiles at a given condition.

The dataset was produced by integrating whole genome SC promoter sequences with expression patterns of the corresponding genes in 40 experimental conditions including cell cycle, sporulation and various stress responses (see http://longitude.weizmann.ac.il/TFLocation/conditions_explist.html for a full list of conditions). We systematically scanned all k-mers (k ranges from 7-11) that appear in *SC* promoters. For each k-mer, we computed the EC score of the set of genes that contain it in their promoter across the 40 conditions. A p-value was assigned to each EC score, which estimates the probability of obtaining the observed or higher EC score by chance [15] and false discovery rate (FDR) [16] of 0.1 was applied to correct for multiple hypotheses. A total of 8610 sequence motifs appeared significant in at least one of the tested experimental condition. These comprise the core of the dataset (hereafter referred do as the 'core dataset').

2.2 Method Validation and Comparison to Published Datasets

To validate the ability of our method to identify biologically significant motifs we tested out whether previously published regulatory motifs score highly using the same method. 89/102 (87%) of the TF binding sites published by Harbison et al. [12] passed FDR of 0.1 in at least one experimental condition, and thus would have been discovered by our method. For comparison only 15/102 (15%) random gene sets (identical in size to the gene sets containing each of Harbison's motifs) appeared significant in at least one condition. Additionally we assessed our coverage of Harbison's set by comparing each core motif to all of Harbison's positional weight matrices (PWMs) [11]. We devised a score between 0-100% that denotes how likely a given string is to be generated from a given PWM (see methods). Requiring a match score of 99% we obtain a coverage of 89/102 (87%), and of 70/77 (91%) of the non redundant Harbison set (The 102 PWMs fall into clusters of highly similar motifs). By relaxing the similarity requirement to 90% the coverage increases to 100/102 motifs, falling into 76/77 distinct clusters (table 1).

Table 1. Coverage of the Harbison motif set by our dataset. A scoring method was devised to assess how likely a given string is to be generated from a given PWM. The score is on a scale of 0 to 100 (see methods). We computed this score for all 8610 core motifs over the 102 Harbison PWMs. The coverage of Harbison's motif set was assessed for several different score cutoffs. Note that a single string may match more than one Harbison PWM, because of redundancy in Harbison's dataset. There are two very long (17 and 18 positions) gapped motif in Harbison's set, for which we have no match, because our set only covers motifs of length 7-11.

Score Cut-off	Coverage of our dataset	Coverage of Harbison	Coverage of unique Harbison clusters
99%	1402/8610=16%	89/102=87%	70/77=91%
98%	1528/8610=18%	93/102=91%	73/77=95%
97%	1719/8610=20%	96/102=94%	73/77=95%
95%	2198/8610=25%	99/102=97%	75/77=97%
92%	3251/8610=38%	99/102=97%	75/77=97%
90%	4161/8610=48%	100/102=98%	76/77=99%

2.3 Exploiting Our Dataset to Predict the Outcome of a Binding Site Substitution

In the process of producing our core dataset, we assigned EC scores, corresponding p-values and likely expression effects to all k-mers residing in yeast promoters, regardless of whether they were ultimately included in the dataset. This provided a unique source of information for our analysis: By comparing the EC scores and the induced expression profiles of k-mers differing in a single position we could predict the outcome of a substitution that transforms one k-mer into the other. Three main scenarios were observed (i) Two k-mers differing in a single position both belong to the core dataset (passed FDR) and regulate genes with a similar expression profile. This implies that the k-mers are recognized by the same TF and a substitution from one to the other will have a very mild effect (Figure 1, green arrows). (ii) The two k-mers belong to the core dataset but regulate genes with a different expression profile. This may imply that they are recognized by different TFs, thus a substitution from one k-mer to the other will cause binding site switching (Figure 1, blue arrows). (iii) One k-mer belongs to the core set whereas the other did not pass the FDR constraint. This implies that substituting the former to the latter will result in binding site loss without acquisition of a new site (Figure 1 red arrows).

We devised three quantitative measures in order to compare the regulatory functions of two k-mers: (1) ΔEC – the difference in EC scores between the set of genes containing k-mer a in their promoters and the set of genes containing k-mer b in their promoters. (2) ΔPV – the difference in the logarithm of p-values assigned to the EC scores of the two gene sets (3) Distance in expression profiles – each k-mer is represented by the mean expression profile of all genes containing it in their promoters. This measure is the distance between the mean expression profiles of the two gene sets (calculated as 1-corelation coefficient of the two vectors representing the means).

Note that although these three measures may seem redundant, they capture slightly different phenomena; An unaltered EC score (low ΔEC) accompanied by a significant change in mean expression profile may imply TF switching, whereas the opposite case in which the expression profile is maintained, but there is a decrease in coherence (high ΔEC) may imply lower affinity to the same TF. The combination of these two measures could thus aid in differentiating between cases of TF switching and cases of a reduction in binding affinity of the same TF.

We have developed a computational tool termed 'motif landscape analysis' [15] that employs our comprehensive dataset in order to systematically predict the outcome of all possible single nucleotide substitutions within a given motif. For a motif of length L this tool examines all 3*L k-mers that are obtained by substituting the motif consensus at each single position. For each such k-mer it computes the three described measures ΔEC, ΔPV and distance in expression profiles between genes containing it in their promoters and genes containing the consensus motif. The results are graphically displayed using a modified version of the previously introduced Combinogram [13] showing the EC scores of the gene set including each of the 3*L motif variants in their promoters and the similarity of their averaged expression profiles.

Applying this tool to the yeast sporulation factor Ndt80 (Figure 1 right panel) using the *SC* sporulation expression data, predicted that two out of the three possible substitutions in the second position will have only a minor effect on expression whereas an A->G substitution at the same position will result in a severe effect (see figure 1

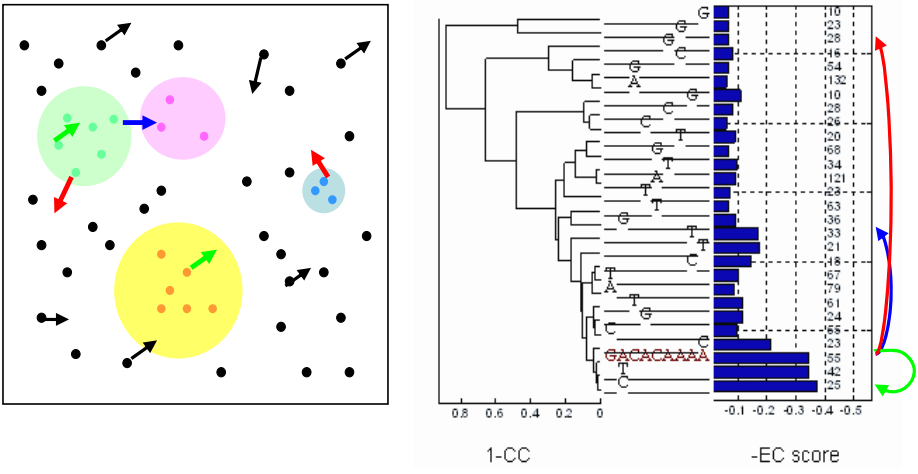

Fig. 1. Possible outcomes of binding site substitutions: Left panel a cartoon depicting possible effects of mutations in regulatory motifs. Points represent promoter elements and discs represent transcription factor recognition ranges. Points that are included within the disc of a given TF represent promoter elements that are bound by the TF. Arrows illustrate the result of single nucleotide substitution within a promoter element. Such a substitution, can cause binding site loss (*red arrows*), a change in affinity to the same TF (*green arrows*), or binding site switching - creation of a binding site with higher affinity to a different TF (*blue arrows*). The right panel illustrates the detection of the same outcomes using our motif landscape analysis tool (as described in detail in [15]). This display captures the effects of single nucleotide substitutions of a given motif on the expression profiles of the downstream genes. The analyzed motif is the yeast Ndt80 sporulation factor (*wild type motif marked in red*). The dendrogram on the left part of the display shows the similarity in mean expression profiles between gene sets baring variations of the motif in their promoters. The right side of the display shows the similarity within sets of genes that contain the same motif variation in their promoters, as measured by the EC score (the numbers correspond to the gene set sizes). The middle section displays the sequence of the motif variation studied in the corresponding row (with a '-' indicating same nucleotide as the wild type motif). A substitution, that is in the recognition range of the same TF, is expected to maintain a high EC score and a similar expression profile (*green arrow*), A substitution that causes binding site loss, is expected to be recognized by both loss of coherence and a change in the mean expression profile (*red arrow*). A substitution that creates a new motif, that is in the recognition range of a different TF, is expected to maintain high expression coherence, while altering the mean expression profile (*blue arrow*). The second motif position appears relatively tolerant to substitutions, 2 out of the 3 possible single nucleotide substitutions of this position do not alter TF recognition (green substitutions). This observation is supported by the recently published structural data of Ndt80 bound to DNA [17]. The second motif position does not form a contact with the protein[1].

legend for details). When averaging over all possible single nucleotide substitutions, the second position appears to be the most tolerant towards substitutions and the seventh position - the most sensitive (figure 2). These results are in agreement with the

[1] The figures of this paper appear in color in the online version of the book.

structural data of Ndt80 bound to its DNA target [17]; the second 'permissive' motif position is the only position which does not form a direct contact with the protein. It is also supported by recently published in vivo reporter expression experiments and in vitro binding essays of Ndt80 mutants, that showed that this position is the most permissive one, and that, as predicted here, G is the only nucleotide that when placed at this position weakens binding affinity and reduces expression level of the reporter gene [18]. This implies that our method can complement and in some cases replace time consuming mutation experiments.

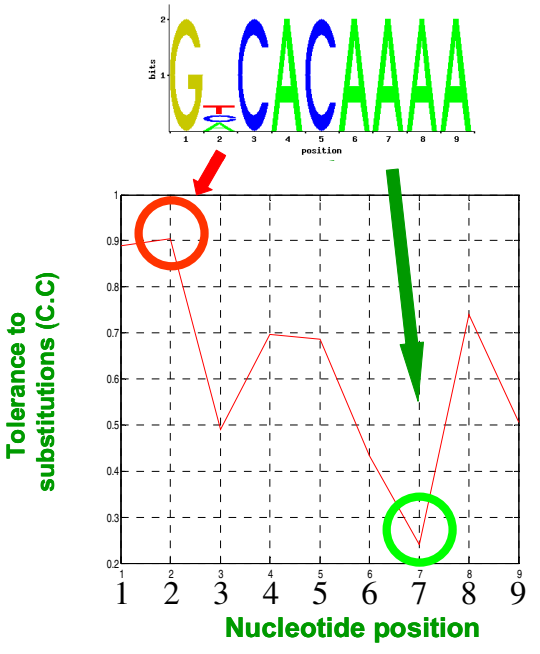

Fig. 2. The averaged tolerance to substitution for each nucleotide position within the Ndt80 motif was defined as the averaged correlation coefficient between the averaged expression profiles of the genes that have a perfect match to the consensus motif and the averaged expression profiles of the genes that have each of the three possible substitutions relative to the consensus in that position

2.4 Deducing General Properties of Expression-Altering Substitutions

Encouraged by our ability to predict the effects of binding site substitutions within a single motif, we attempted to generalize these predictions in order to define universal properties of substitutions that alter gene expression. We used the three measures described above to assess the severity of a substitution from base i to base j in a regulatory motif. Namely: change in EC score, change in p-value and change in mean expression profiles of genes regulated by a motif with nucleotide j versus genes regulated by the same motif with nucleotide i at the substituted position. This time, instead of analyzing a single motif we accumulated statistics from substitutions of different positions across multiple binding sites. There are twelve possible single nucleotide substitutions from base i to base j (when i can be A,C,G or T, and $j \neq i$). Each

severity measure was averaged over all substitutions of the type ni-> nj in any possible motif. The motifs used for this analysis were core dataset motifs that correspond to known TFBS from Harbison's set [12].

Our fist question was whether there were substitution types that are more radical than others (in analogy to amino acid substitutions where there are conservative changes that maintain the chemical properties of the residue versus radical changes that form a residue with different characteristics). Interestingly, although there was no single substitution type that appeared more radical than others, there were systematically higher penalties for substitutions that abolished a C or a G versus substitutions that abolished an A or a T (figure 3). Because the yeast genome is AT rich, this result may suggest that C and G are the nucleotides that confer specificity to a motif, and thus their substitution bears a greater effect on the motif's function. This raises a prediction that in other genomes with different GC content of the regulatory regions the penalties might be different, reflecting loss of information content with the elimination of different nucleotides. We intend to check this hypothesis in the human genome, which has a higher GC content than yeast, using the tools developed here.

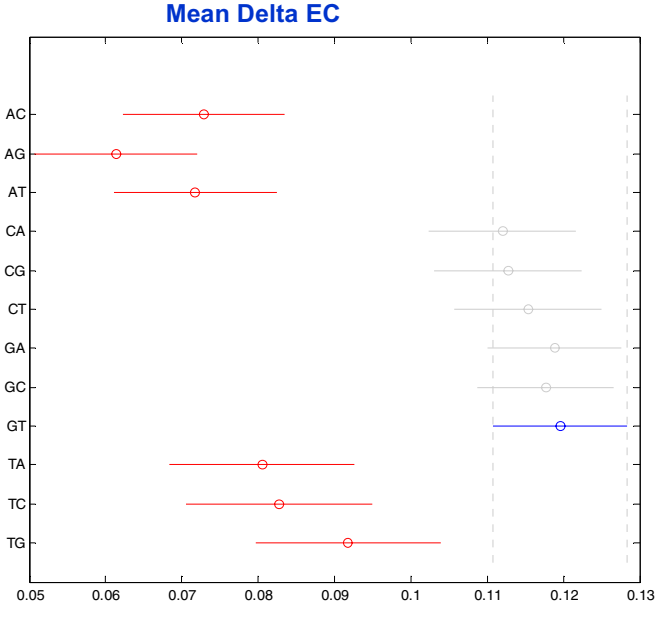

Fig. 3. Effects of each of the twelve possible single nucleotide substitutions on expression. The data was accumulated from all possible substitutions of each type in a dataset of highly scoring k-mers that correspond to known Harbison PWMs. The 'severity measure' applied is mean delta EC, thus high values correspond to sever substitutions. A clear trend is seen whereby substitutions that abolish an A or a T are less severe than substitutions that abolish a C or a G.

2.5 The Information Content of the Substituted Position

The degree of conservation of the substituted position in the PWM may also affect the severity of the phenotype. This was shown to be the case in protein coding SNPs [1, 2],

but parallel investigations were not carried out in promoter motifs. Substitutions of highly conserved positions are expected to have a more dramatic effect on expression compared to substitutions of positions with low conservation. To test this hypothesis we analyzed high scoring k-mers from our dataset which correspond to known Harbison PWMs. For these motifs both the expression measures obtained in the process of creating our dataset (EC, p-value, expression profiles) and the information content (IC) of all PWM positions are available. We could thus assess the correlation between the IC of a position and its sensitivity to substitution based on the previously described severity measures. Indeed a significant correlation exists between the mean expression distance and the IC of a position (table 2). The mean expression distance is also highly correlated to our other two expression based measures mean ΔEC and mean ΔPV.

Table 2. Correlations between the three expression measures mean ΔEC, mean ΔPV, mean expression distance and the IC of a PSSM position. Data was accumulated for 1867 positions. In each table cell, the first number is the correlation and the second number is the p-value on this correlation. The different expression measures are highly correlated. There is a correlation between the measure Mean Expression distance and the information content of a position. Note that the change in mean expression profile has a very significant but rather low correlation (0.3) with the other two expression based measures. This is because there are cases where high EC is maintained, but the expression profile changes (implying TF switching), and cases where the expression profile is maintained, but there is a decrease in coherence (implying lower affinity to the same TF).

	Mean ΔEC	Mean ΔPV	Mean expression Distance	Position IC
Mean ΔEC	1			
Mean ΔPV	0.5402 6.12e-142	1		
Mean Expression Distance	0.3505 4.34e-055	0.3053 1.41e-041	1	
Position IC	0.0827 3.47e-004	0.0531 0.0217	0.1252 5.7785e-008	1

3 Discussion

We have composed a comprehensive dataset of TFBS and developed measures for quantifying the effect of a binding site, present in the promoter, on the expression profiles of the regulated genes. These measures allowed us to compare the effects on gene expression of binding sites differing by a single nucleotide position, and to infer from the comparison what would be the severity of substituting one binding site into the other. We applied our tools to the yeast genome and were able to produce reliable predictions about the outcome of single nucleotide substitutions in a single binding site. By accumulating statistics for many substitutions across multiple binding sites we observed that not all nucleotide substitutions are similar in severity: In the *SC* genome abolishing a C or a G has a harsher effect on average than abolishing an A or a T. Although this result may be specific to the AT rich *SC* genome, the same measures

and tools can be easily applied to other genomes, and specifically to human. We have showed that other characteristics of a substituted motif position, such as its IC are in correlation with our measures of the effect on expression. We intend to test additional features including the evolutionary conservation of the substituted position and its vicinity to the protein in the DNA-protein co-crystal structure. All these features can be integrated to form a prioritization scheme that would allow the ranking of existing genome variations by their disease-causing potential.

The approach presented here demonstrates for the first time how a huge amount of data, on all known yeast TFs, using all genes, whose expression was monitored in multiple conditions, can be harnessed and utilized for taking the first step towards assessing the effects of nucleotide substitutions in TF binding sites. A conceptual analog of this endeavor for assessing the effects of amino acid substitutions on protein function could amount to mutating many proteins, say enzymes, in many different ways, and checking for each mutation reduction, or change, in biochemical activity and specificity. Since data for such effort is not even close to become available, the methodology presented utilizes in a unique way data that is available for its domain. While the main advantage of our methodology is the huge sample size, the disadvantage is that we are unable to control for other differences between promoters of analyzed genes (i.e. differences that are outside of the substituted position). The fact that we get statistically significant differences between the effects of different types of substitutions on expression likely indicates that despite uncontrolled sources of variation we extracted genuine signals.

An additional application of the present approach may be in algorithms that assign PWMs to promoters (e.g. PRIMA [19]) as it should provide means to weigh differently mismatches between the PWM preferences and the promoter sequence based on expected effect on expression.

4 Materials and Methods

4.1 Dataset Construction

Promoter sequences for 5651 *SC* genes were taken from SGD [20]. Expression data for 40 different time series experiments was downloaded from ExpressDB [21]. The promoters were systematically scanned for all occurrences of every possible k-mer (k varies from 7-11), resulting in an index file listing for each k-mer the set of genes that contain it in their promoters, along with the positions and orientations (strand). For the purpose of indexing each k-mer was combined with its reverse complement because it is well accepted that TFs bind double stranded DNA.

Following the k-mer indexing, EC scores in various experimental conditions were calculated for the sets of genes containing each of the k-mers in their promoters. A p-value was assigned to each EC score and false discovery rate (FDR) of 0.1 (allowing 10% false positives) was used to correct for multiple hypotheses.

4.2 Expression Coherence (EC) Score

The formal definition of the EC score is the fraction of pairs of genes in a given set S, for which the Euclidean distance between expression profiles falls bellow a threshold D.

$$EC(S) = \frac{\left|\left\{g_i, g_{j\neq i} \in S : ExpDist(g_i, g_j) < D\right\}\right|}{|S| * (|S| - 1) \div 2} \quad (1)$$

The threshold D is determined based on the distribution of pair-wise distances between expression profiles of all genes in the genome (or more precisely of all genes for which expression level was measured). The original definition of the EC score [13] used the 5th percentile as the cutoff for defining "close" expression profiles (D). This definition may create a bias towards TFs that exert a very tight regulation and miss regulatory motifs that correspond to factors exerting a more lose regulation. We therefore tested a range of EC definitions, with cutoffs corresponding to the 5th, 10th, 20th, 30th, 40th and 50th percentile of the pair-wise distance distribution. For each definition of EC cutoff we assigned a significance p-value separately. P-values were calculated by random sampling. For each of the 40 expression time series and for each gene set sizes (varying from 3-100 genes), we selected 100,000 random gene sets and computed an EC score for each such set at each cutoff definition. We define the p-value of a given EC score as the fraction of random sets (of the same size and condition) that scored similarly or higher (note that this sets a lower bound of 10^{-5} on the significance that can be assigned to a given EC score). Since we assume that for a given EC score, the probability to get the same score for random sets of genes drops with the set size, gene sets larger than 100 are assigned an upper bound approximated p-value, using the randomly sampled sets of size 100.

4.3 Comparing Our Binding Sites to Known PWMs

A scoring method was devised to assess how likely a given string is to be generated from a given PWM. The score is on a scale of 0 to 100. It is computed by summing up the frequencies corresponding to the observed nucleotides over all motif positions, and normalizing this score to a scale of 0-100. The scaling is done by subtracting the minimal possible score and dividing by the range of possible scores. For example for the PWM [A: 0.0191 0.0191 0.9733 0.9733 0.0120, C:0.9500 0.9500 0.0074 0.0074 0.0074, G: 0.0117 0.0117 0.0074 0.0074 0.0074 T:0.0191 0.0191 0.0120 0.0120 0.9733] the lowest possible score 0.0455 is obtained for the string GG(C/G)(C/G)(C/G), the highest possible score 4.8198 is obtained for the string CCAAT. After scaling GGCCC will score 0%, CCAAT will score 100% and CCATT will score 79.9% ((3.8585-0.0455)/(4.8198-0.0455)). Because our k-mers and the known PWMs may differ in length, we aligned them by sliding the shorter sequence along the longer. For each such alignment, we calculate the match score (in percentage 0-100%) and took the position with the best score as the true alignment.

References

1. Ng, P.C., Henikoff, S.: Predicting deleterious amino acid substitutions. Genome Res 11 (2001) 863-874
2. Ng, P.C., Henikoff, S.: Accounting for human polymorphisms predicted to affect protein function. Genome Res 12 (2002) 436-446

3. Ng, P.C., Henikoff, S.: SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31 (2003) 3812-3814
4. Wang, Z., Moult, J.: SNPs, protein structure, and disease. Hum Mutat 17 (2001) 263-270.
5. Sunyaev, S., Ramensky, V., Bork, P.: Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16 (2000) 198-200.
6. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A.S., Bork, P.: Prediction of deleterious human alleles. Hum Mol Genet 10 (2001) 591-597.
7. Vitkup, D., Sander, C., Church, G.M.: The amino-acid mutational spectrum of human genetic disease. Genome Biol 4 (2003) R72
8. Prokunina, L., Castillejo-Lopez, C., Oberg, F., Gunnarsson, I., Berg, L., Magnusson, V., Brookes, A.J., Tentler, D., Kristjansdottir, H., Grondal, G., Bolstad, A.I., Svenungsson, E., Lundberg, I., Sturfelt, G., Jonssen, A., Truedsson, L., Lima, G., Alcocer-Varela, J., Jonsson, R., Gyllensten, U.B., Harley, J.B., Alarcon-Segovia, D., Steinsson, K., Alarcon-Riquelme, M.E.: A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans. Nat Genet 32 (2002) 666-669
9. Zwarts, K.Y., Clee, S.M., Zwinderman, A.H., Engert, J.C., Singaraja, R., Loubser, O., James, E., Roomp, K., Hudson, T.J., Jukema, J.W., Kastelein, J.J., Hayden, M.R.: ABCA1 regulatory variants influence coronary artery disease independent of effects on plasma lipid levels. Clin Genet 61 (2002) 115-125
10. Rockman, M.V., Wray, G.A.: Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 19 (2002) 1991-2004
11. Shalgi, R., Lapidot, M., Shamir, R., Pilpel, Y.: A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. Genome Biol 6 (2005) R86
12. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., Young, R.A.: Transcriptional regulatory code of a eukaryotic genome. Nature 431 (2004) 99-104
13. Pilpel, Y., Sudarsanam, P., Church, G.M.: Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 29 (2001) 153-159
14. Sudarsanam, P., Pilpel, Y., Church, G.M.: Genome-wide Co-occurrence of Promoter Elements Reveals a cis-Regulatory Cassette of rRNA Transcription Motifs in Saccharomyces cerevisiae. Genome Res 12 (2002) 1723-1731
15. Lapidot, M., Pilpel, Y.: Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription. Nucleic Acids Res 31 (2003) 3824-3828
16. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy Stat Soc 57 (1995) 289-300
17. Lamoureux, J.S., Stuart, D., Tsang, R., Wu, C., Glover, J.N.: Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. Embo J 21 (2002) 5721-5732
18. Pierce, M., Benjamin, K.R., Montano, S.P., Georgiadis, M.M., Winter, E., Vershon, A.K.: Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol Cell Biol 23 (2003) 4814-4825
19. Elkon, R., Linhart, C., Sharan, R., Shamir, R., Shiloh, Y.: Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res 13 (2003) 773-780
20. SGD: http://www.yeastgenome.org/.
21. ExpressDB: http://salt2.med.harvard.edu/ExpressDB/.