BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 14 2007, pages 1737–1743
doi:10.1093/bioinformatics/btm249

Sequence analysis

STOP: searching for transcription factor motifs using gene expression

Libi Hertzberg[1,2], Shai Izraeli[1] and Eytan Domany[2,*]

[1]Department of Pediatric Hemato-Oncology, The Sheba Cancer Research Center, Sheba Medical Center, Tel Hashomer Israel and Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978 and [2]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

Received on December 14, 2006; revised and accepted on May 2, 2007

Advance Access publication May 11, 2007

Associate Editor: Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Existing computational methods that identify transcription factor (TF) binding sites on a gene's promoter are plagued by significant inaccuracies. Binding of a TF to a particular sequence is assessed by comparing its similarity score, obtained from the TF's known position weight matrix (PWM), to a threshold. If the similarity score is above the threshold, the sequence is considered a putative binding site. Determining this threshold is a central part of the problem, for which no satisfactory biologically based solution exists.

Results: We present here a method that integrates gene expression data with sequence-based scoring of TF binding sites, for determining a global score threshold for each TF. We validate our method, STOP (Searching TFs Of Promoters), in several ways: (1) we calculate the average expression values of groups of human putative target genes of each TF, and compare them to similar averages derived for random gene groups. The groups of putative targets show significantly higher relative average expression. (2) We find high consistency between the induced lists of putative targets in human and in mouse. (3) The expression patterns associated with human and mouse genes (ordered by PWM scores for each TF) exhibit high similarity between human and mouse, indicating that our method has firm biological basis. (4) Comparison of results obtained by STOP and PRIMA (Elkon *et al*., 2003) suggests that determining the score threshold using gene expression, as is done in STOP, is more biologically tuned.

Availability: Software package will be available for academic users upon request.

Contact: eytan.domany@weizmann.ac.il

Supplementary information: Supplementary data are available on *Bioinformatics* online.

1 INTRODUCTION

Elucidation of the regulatory mechanisms that control gene expression is a central issue in biology. One of the key elements in regulation of transcription are the binding sites of transcription factors (TFs) which are located along the DNA. The availability of whole genome sequences gave rise to the development of computational methods that search for these TF binding sites. Development of such methods, that are reliable and have low false positive and false negative rates, is of paramount importance.

We present here a method that searches for binding sites of TFs with known position weight matrices (PWMs). Our method follows the standard strategy: given a DNA sequence and a PWM denoted by M, the method uses a score that measures the quality of the match between the sequence and the matrix. The target (say, promoter) sequence is scanned and the binding motif of maximal score found on it is identified. Suppose that the value of this maximal score is s. We then need to decide whether s is high enough to call the corresponding sequence a 'hit', i.e. predict that the TF binds at the corresponding position. This decision is taken on the basis of comparison of s with a threshold $T(M)$: if $s > T(M)$ there is a hit or match and a putative binding site of the TF is identified. All methods that use PWMs have to deal with the question of determining an optimal threshold for each TF.

In PRIMA (Elkon *et al*., 2003), the value of the threshold is determined in a way that keeps the number of matches in a big background promoter set (of 12 981 promoters of length 1200 bp) approximately fixed at ~10% of the whole set. MatInspector (Quandt *et al*., 1995) calculates for every PWM an 'optimized threshold', 'minimizing the number of false positive hits in non-regulatory sequences' (http://www.genomatix.de/products/MatInspector/). However, the manner in which this threshold was computed was not published. In UCSC TFBS Conserved Track Settings, http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=80108084&g=tfbsConsSites, only those binding sequences that are conserved in the human/mouse/rat alignment are considered putative binding sites; a binding site is required to score above threshold in all three species. The value of the threshold is calculated for each PWM as follows. For each of the 24 635 RefSeq genes (taken from genome assembly hg 17) calculate the score (of the PWM) at every location on the promoter, from the transcription start site (TSS) to 5000 bps upstream. This generates (approximately) 24 635 times 5000 scores s. The threshold T is set so that a fixed fraction of these satisfy $s > T$.

In Match (Kel *et al*., 2003), a different threshold value is used for each TF, tuned so that the number of matches in a set of

*To whom correspondence should be addressed.

exon sites is minimized, because these sequences are presumed to contain no biologically relevant TF binding sites. Indeed, so far no functional, experimentally proven binding site was reported in the exon 2 sequences used by (Kel *et al*., 2003). However, the choice of thresholds based on such negative results is obviously limited by the scope of the search for counter examples that has been performed. F-match (Kel *et al*., 2006) chooses an optimal threshold for each PWM in the following way. Given an input group of genes and a background group of genes it scans a set of thresholds and chooses the one that maximizes the enrichment of the PWM in the given group in respect to the background group. Similarly, it chooses a threshold that minimizes this enrichment. The threshold chosen for each PWM is not global and depends on the input group of genes.

Several methods choose TFs' thresholds in a way that tunes the number of putative targets to be about the same for each TF, as was described earlier. However, determining the threshold according to the assumption that all TFs have a similar number of targets is problematic, since TFs differ in their affinity and selectivity in binding DNA motifs and in their number of targets. For example, according to (Odom *et al*., 2006), HNF4A has 5-fold more target genes than FOXA2.

In order to determine the thresholds in a more biologically sensible way, we developed a method that uses gene expression data in the threshold determination process. The idea to use gene expression for this purpose is based on the biological fact that co-regulated genes are co-expressed, i.e. have similar expression profiles over a set of samples. For each TF the threshold is determined separately, based on human gene expression datasets. This way the number of putative targets of each TF is independent of the other TFs.

2 METHODS

2.1 Gene expression datasets

The following three human gene expression datasets were measured on the GeneChip Human Genome HG-U133A Array (Affymetrix, Santa Clara, CA) and are used in this article.

(1) St Jude dataset containing 132 ALL samples (Ross, 2003).

(2) Colon dataset containing 202 samples (Tsafrir *et al*., 2006).

(3) Scripps dataset containing 158 samples obtained from 79 normal tissues (Su, 2004).

We also use a mouse gene expression dataset of 36 cardiomyocyte samples (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76) that was measured on the GeneChip Mouse Genome MG-U74A Array.

Low-level processing and normalization of the datasets were carried out using the Affymetrix Microarray Suite 5.0 (MAS 5.0) software.

2.2 Gene promoter sequences

DNA sequences upstream of human TSSs were downloaded from the GoldenPath server at UCSC http://hgdownload.cse.ucsc.edu/goldenPath/hg16/bigZips/. Putative regulatory regions of 300 bps upstream of the TSSs for the different genes were obtained. See Supplementary Material for explanation for taking 300 bps.

We used as our working set of genes the GeneChip Human Genome HG-U133A Array that represents more than 20 000 transcripts derived from approximately 17 000 well-substantiated human genes. We chose this array because of its coverage of the human genome and because of the availability of variable gene expression datasets measured on it.

DNA sequences upstream of mouse TSSs were downloaded from the GoldenPath server at UCSC http://hgdownload.cse.ucsc.edu/goldenPath/mm5/bigZips/. Putative regulatory regions (of 300 bps upstream of the TSS) for the different genes were obtained.

For the consistency *P*-value analysis (see Sections 2.6 and 3.3), we used as our working set of genes the GeneChip Mouse Genome 430 2.0 Array that represents over 39 000 transcripts. We chose this array because of its extensive coverage of the mouse genome.

For the comparative *Z*-score analysis (see Section 3.4), we used as our working set of genes the GeneChip Mouse Genome MG-U74A Array that represents over 12 000 transcripts. We chose this array because of the availability of variable gene expression datasets measured on it.

2.3 Calculating a score and a *P*-value (*P*) for a motif and a PWM

A common representation of TF binding sites is a PWM. A PWM is built out of all the known motifs to which the TF binds, and it counts the number of appearances of every nucleotide in every position of the motif. We use 392 human TF PWMs downloaded from the TRANSFAC database (Matys, 2003).

For a given PWM M and a given motif, we calculate the standard log-likelihood ratio to be the score of the motif. See Supplementary Material for the exact calculation.

We search a given promoter sequence of length N for the location with the maximal score for a given PWM. Then we calculate a *P*-value: the probability to get such a maximal score or higher in a random promoter of length N. We denote this *P*-value by P. P is calculated using the algorithm developed in (Hertzberg *et al*., 2005).

2.4 *Z*-score calculation using gene expression data

Suppose we have n genes on a gene expression array a. Let $g_1^a, \ldots, g_n^a$ be their log expression values; μ^a is their average and σ^a is their SD. Now suppose we suspect that a subgroup G^i of k genes, with indices $i_1, \ldots, i_k$, are targets of a particular TF. Their *Z*-score,

$$Z(TF, G^i, a) = \left(\frac{1}{k}\sum_{j=1}^{k} g_{i_j}^a - \mu^a\right)/(\sigma^a/\sqrt{k})$$

expresses the extent to which the average expression of G^i differs from the general average expression. Assuming that $g_1^a, \ldots, g_n^a$ are independent and identically distributed with mean μ^a and SD σ^a, the central limit theorem (CLT) predicts that the average expression of a large group of k randomly selected genes is normally distributed with mean μ^a and SD $\sigma^a/\sqrt{k}$, and thus the calculated *Z*-scores will have a normal distribution with mean 0 and SD 1. However, the earlier assumptions may not be valid since: (1) the expression values of different genes may be correlated; (2) the genes' expression distributions are not identical. In this case, the mean value of the *Z*-score is still expected to equal 0, but the SD will have a different value.

In (Chiang, 2001) it was shown that although the assumptions are not strictly valid for gene expression data, the SD of the *Z*-score distribution varies by <5% from the values given by the CLT. We use the *Z*-score calculation of a group of genes as a measure for the 'randomness' of the group; i.e. if it differs a lot from zero we conclude the gene group is not random. In our context this means that the genes are probably co-regulated by a common TF.

2.5 Determining the score threshold for a TF of interest, using *Z*-score calculation

Given a gene expression dataset and a TF's PWM M, we determine a score threshold for M using the Z-score calculation. Genes that contain a position in their promoter with a score higher than the threshold $T(M)$ will be considered as putative targets for this TF.

Suppose n genes are represented on the expression array. We search the genes' promoters for the location with the maximal score for M, and calculate P, the probability for getting such a value (or higher) for the maximal score. The n genes are ordered (largest score first) according to their maximal score values.

Next, we calculate the Z-score for gene groups of increasing sizes. The simplest way to do this would be to take say the top $100, 200, 300, \ldots$, etc., scoring genes. However, since the PWM assigns to different motifs one of a set of discrete score values, in general there may be several genes with equal maximal scores. We define our gene groups G^k so that when a certain gene is included in it, then the whole set of equal-score genes is also included. Denote by $s_1 > s_2 >, \ldots, > s_x$ the x different maximal score values measured on the n genes ($x \leq n$), and let p_i be the P associated with s_i ($p_1 < p_2 <, \ldots, < p_x$). Let k_i be the number of genes with maximal score $\geq s_i$; clearly $k_1 < k_2 <, \ldots, < k_x$. The general idea is to scan the different values of k, $k_1 < k_2 <, \ldots, < k_x$, and search for the $k_{\max}$ at which the Z-score is maximal (and choose the corresponding score as the score threshold). In order to provide a rough limit on the number of putative targets, we scan only a subgroup of the values of k, $k_j < k_{j+1} <, \ldots, < k_b$, where j is the minimal index such that $k_j > 20$, and b is the maximal index such that $p_b < 0.5$. For each $m = j, j+1, \ldots, b$ the Z-score is calculated for the correspondding group of k_m genes. The Z-score $Z(TF, G^{k_m}, a)$ is calculated for each of the N samples a in the gene expression dataset. Then the average Z-score over the different samples is calculated:

$$\bar{Z}(TF, G^{k_m}) = \sum_{a=1}^{N} Z(TF, G^{k_m}, a)/N$$

We now plot $\bar{Z}(TF, G^{k_m})$ versus k_m and locate m', the value of m ($m = j, j+1, \ldots, b$) at which $|\bar{Z}(TF, G^{k_m})|$ is maximal. The corresponding value of the score, $s_{m'}$ (of the gene ranked $k_{m'}$) is chosen as the threshold $T(TF)$, as determined from this gene expression dataset.

This procedure was performed for each TF PWM and each of our three human gene expression datasets. Thus, for each gene expression dataset and for each TF PWM we get a score threshold.

The final score threshold is now determined by combining the results of n different expression datasets (in our case $n = 3$). The idea is to find a subgroup of gene expression datasets that yield, for a given TF, similar score thresholds and choose this value as our $T(TF)$. This procedure generates more robust score thresholds. In addition, we want to choose a score threshold where the absolute value of the average Z-score is 'high enough', generating score thresholds that are biologically meaningful. Note that a P threshold is associated with each score threshold. The value of the final score threshold is determined in the following way, for each PWM:

(1) Choose a subgroup of datasets with 'close' P thresholds as follows: for each value x between 0 and 0.1 (scanned using steps of 0.01 in increasing order), search for the subgroup of datasets of maximal size k, such that for each pair from the subgroup these conditions are satisfied: (a) their P threshold difference is lower than x, (b) their $\bar{z}$ values (at their respective thresholds) have the same sign and (c) the value of $|\bar{z}|$, averaged over the k datasets is larger than 2.5. If there is such a subgroup of size $k > 1$, stop increasing x. If there is more than one subgroup, choose the one with higher average $|\bar{z}|$. If we reached $x = 0.1$, choose $k = n$, i.e. all datasets are included in the 'subgroup'.

(2) Out of the k datasets in the subgroup choose the score threshold of that dataset which has the maximal value of $|\bar{z}|$.

We compared values of the Z-score thresholds that were obtained using different gene expression datasets, and found that the dependence on the dataset used is weak. See Supplementary Material for a discussion of this approximate independence.

2.6 Consistency *P*-value (P_c)

We study the promoters of 20 248 human probe sets represented on the HG-U133A array and of 29 771 mouse probe sets of the Mouse 430 2.0 array. 8350 probe sets that appear on both chips are associated with identical gene symbols in human and mouse; they constitute 8350 orthologous (OR) gene pairs, common for human and mouse. Denote by OR_{Hu} (and OR_{Mous}) the 8350 human (and mouse) genes.

For each TF PWM, we identified its group of putative target genes in human using the threshold $T(M)$. We used the same values of thresholds to get for each TF its putative target genes in mouse as well. Let $OR_{Hu\text{-}T}$ represent the human putative targets out of OR_{Hu} and let $OR_{Mous\text{-}T}$ represent the mouse putative targets out of OR_{Mous}. The hyper-geometric P-value for the size of the intersection between $OR_{Hu\text{-}T}$ and $OR_{Mous\text{-}T}$ is denoted as the consistency P-value (P_c) of the TF.

3 RESULTS

3.1 Determining the score threshold using *Z*-score

For each TF we determined the score threshold using three different human gene expression datasets. We plot in Figure 1 $\bar{z}$, the average Z-score calculated for the TF NFAT, using the three human gene expression datasets, and demonstrate how the score threshold is determined. It can be seen that for this

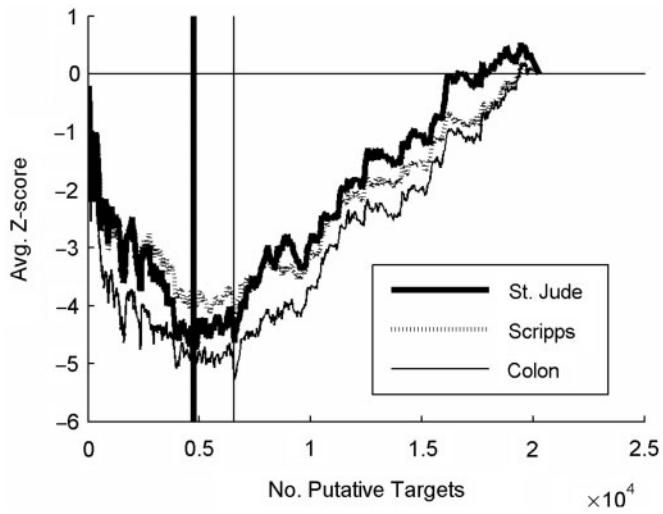

Fig. 1. Score threshold determination of the TF NFAT. The horizontal axis represents the genes on the Affymetrix HG-U133A array, ordered in a descending order of their maximal score of NFAT PWM. For each x, the x genes with the highest maximal score of NFAT are included in the group whose Z-score is calculated. The vertical axis represents $\bar{z}$, the average Z-score for each such group of genes, for each of the three gene expression datasets (bold line for the St Jude, dashed for Scripps and nonbold for the Colon dataset). For example, when $x = 4000$, $\bar{z}$ is calculated for the 4000 genes with the highest maximal score. The vertical lines represent the threshold point of each of the gene expression datasets. The point of maximal $|\bar{z}|$ is identical for two of the three datasets for this TF.

TF the absolute value of $\bar{z}$ attains its maximal value at the same point for two of the three datasets. The PWM score value at this point, $s = 5.8$, was chosen as the threshold $T(NFAT)$. $\bar{z}$ at the threshold point is between -4 and -6 in the three datasets. This suggests that NFAT functions as a repressor since the average expression level of its putative target genes is lower than the general expression level. Indeed, NFAT is known to be a repressor TF (Rao, 1997).

The list of all TFs with their calculated score thresholds and their $\bar{z}$ at that point is given in Supplementary Table 1.

3.2 Comparison to random permutations of genes

In order to validate the determination of the score threshold according to the Z-score calculation, we check the behavior of $\bar{z}$ on a random permutation of the genes. Instead of ordering the genes in a descending order of their maximal score for a TF PWM, we take a random permutation of them. Again, $\bar{z}$ is calculated for increasing sizes of gene groups. The results for the Scripps dataset, for a random permutation, as well as for ranking according to the maximal score of the TF ELK1, are plotted in Figure 2. It can be seen that $|\bar{z}|$ is significantly lower for the random ordering. This suggests that the high value of $\bar{z}$ seen for ELK1 is not the result of a random fluctuation; i.e. the measure of the Z-score reflects some biological meaning. In this case a plausible explanation is that the group of genes with high maximal score for ELK1 is transcriptionally co-activated by it, causing this group's relatively high average expression level and correspondingly, high $\bar{z}$.

We use this observation to further estimate which values of the maximal absolute value of $\bar{z}$ can be designated as 'high' and not due to random fluctuations. For this purpose, we generated 1000 random permutations and repeated the analysis described earlier for each of the resulting orderings of the genes. The results are plotted in Figure 3, together with the histogram of the maximal $|\bar{z}|$ values that were obtained when the genes were ordered according to the scores of 392 TFs. Both histograms were obtained from the Scripps expression data.

The maximal $|\bar{z}|$ is significantly higher when the genes are ordered according to their maximal score motif for TFs than when the ordering is by random permutations. This suggests that the high values of $|\bar{z}|$ seen for the TFs are not random-high absolute Z-score values of the high-scoring genes probably reflect their co-regulation, which causes their average expression level to be different from the average expression level of all genes.

3.3 Comparison to the mouse genome

In order to test and validate further our method for threshold selection, we measured for each TF the consistency of its putative targets between human and mouse. P_c (see Methods Section) expresses the extent to which the human putative target genes tend to overlap with the mouse putative target genes. Conservation of regulatory motifs along different species is known to strengthen the likelihood that the motifs are biologically significant.

The P_c values of all 392 TFs studied were calculated; the results are plotted in Figure 4. 390 TFs passed FDR of $Q = 0.05$ (these TFs have P_c values lower than 0.049). Since 99% of the TFs have statistically significant P_c values, we conclude that the proposed determination of the score threshold according to the Z-score analysis produces biologically meaningful results.

There is a clear correlation between the value of $|\bar{z}|$ at the score threshold and the value of P_c. The higher the value of $|\bar{z}|$ at the threshold, the higher is the consistency between human and mouse. This observation provides further biological support to the determination of the score threshold according to the Z-score analysis.

3.4 Comparison to the mouse gene expression data

To complete the comparison between human and mouse, we used a mouse gene expression dataset. The Z-score procedure described earlier for human was performed on a mouse gene expression dataset. For each TF a score threshold was chosen according to the maximal $|\bar{z}|$ on the mouse data. We present in Figure 5 the $\bar{z}$ values at the score threshold points, as obtained from mouse versus human (Scripps) expression data.

For ~90% of the TFs the sign of $\bar{z}$ at the threshold is identical for human and mouse and the correlation between the two $\bar{z}$ values is high (more than 0.7, for $|\bar{z}|$ higher than 2.5). The biological meaning of this correlation in our context is that TFs which function as activators in human function as activators also in mouse; the same holds for repressors.

It should be noted that the Z-scores were measured here for completely different sets of genes: human versus mouse. The genes of each genome were ordered according to the maximal score motif in their promoter for the TF, using the same PWMs. Note that this ordering is not identical between human and mouse since in human the search for the maximal score motif was done on the human genes' promoters, and in mouse it was done on the mouse genes' promoters. The fraction of TFs that have $\bar{z}$ with different signs for human and mouse is

Table 1. Comparison between STOP and PRIMA results on E2F4 targets

PWMs (TFs)	STOP % matches (rank)	STOP *P*-value	PRIMA % matches (rank)	PRIMA *P*-value
V$E2F4DP2_01 (E2F4:DP2)	58 (1)	$3.11E-32$	15 (2)	$3.77E-17$
V$E2F1_Q3 (E2F1)	62 (2)	$6.55E-32$	22 (13)	$6.24E-11$
V$E2F1DP1_01 (E2F1:DP1)	77 (3)	$1.84E-26$	15 (2)	$3.77E-17$
V$E2F1_Q6_01	83 (5)	$4.16E-25$	16 (1)	$1.73E-20$
V$NFY_01	30 (17)	$1.28E-17$	29 (2)	$3.77E-17$

30% for those TFs with $|\bar{z}| \leq 2.5$ at the threshold (in the Scripps dataset), while for TFs with $|\bar{z}| \geq 2.5$ this fraction is only 4%. This suggests that when $\bar{z}$ is higher, the sign of $\bar{z}$, which reflects the activity type of the TF (activator/repressor) is more likely to be similar in human and mouse.

In continuation to the earlier analysis, we compared the average *Z*-score graphs (see Supplementary Fig. S2) of the TFs between human and mouse. For 50% of the TFs the average *Z*-score graphs between human and mouse are highly correlated (correlation >0.82). See Supplementary Material for a full description of this analysis.

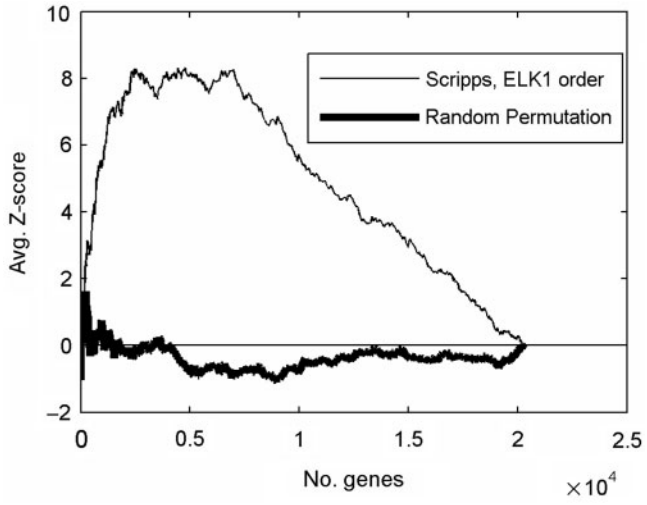

Fig. 2. Average *Z*-score of a random gene permutation. The horizontal axis represents the genes on the Affymetrix HG-U133A array. The vertical axis represents $\bar{z}$ for each size of group of genes for the Scripps expression dataset. The non-bold line represents $\bar{z}$ when the genes are ordered in a descending order of their maximal score for TF ELK1. The bold line represents $\bar{z}$ when the genes are ordered at random.

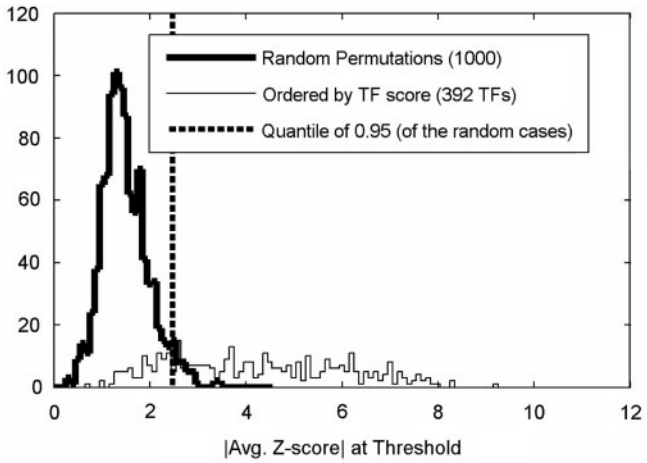

Fig. 3. Histogram of maximal average *Z*-score for genes ordered by either the TF scores or randomly. The horizontal axis represents maximal value of $|\bar{z}|$. The bold line represents the number of random gene permutations with such a maximal $|\bar{z}|$. The non-bold line represents the number of TFs with such a maximal $|\bar{z}|$. In 95% of 1000 random permutations, the maximal $|\bar{z}|$ was lower than 2.5.

3.5 Comparison to prima

We compared our method with PRIMA, a program for searching for common TFs in a set of genes using PWMs from the TRANSFAC database (Elkon *et al.*, 2003). Given a group of genes, PRIMA calculates a *P*-value for the enrichment

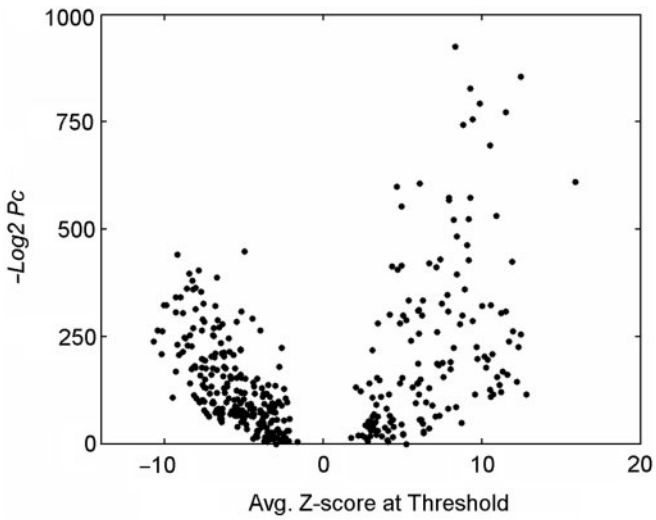

Fig. 4. Human-mouse consistency *P*-values (P_c). Each point corresponds to a TF. The horizontal axis represents $\bar{z}$ at the point of the score threshold (from the Scripps data). The vertical axis represents the minus log (basis 2) of P_c.

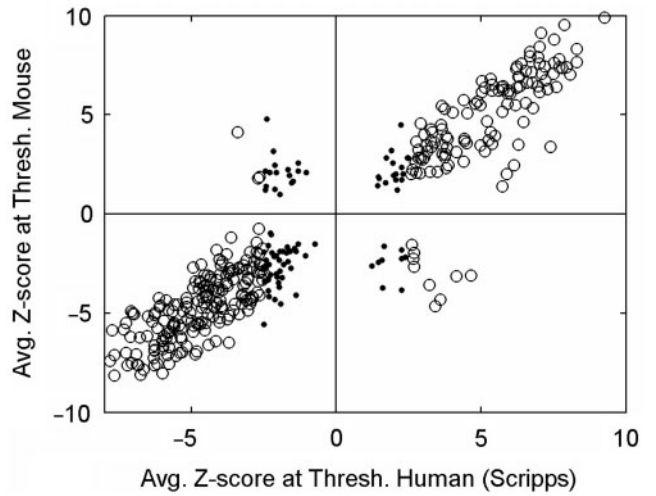

Fig. 5. Comparison of human-mouse *Z*-score analyses. Each point corresponds to a TF. The horizontal (vertical) axis represents $\bar{z}$ at the score threshold according to the human Scripps (or Mouse) gene expression dataset. Open circles correspond to TFs with Scripps $|\bar{z}| > 2.5$; solid points to those with $|\bar{z}| < 2.5$. The Pearson correlation is 0.73 for the points with $\bar{z} > 2.5$, and 0.72 for $\bar{z} < -2.5$. In the intermediate region $-2.5 < \bar{z} < 2.5$ the Pearson correlation is small.

of each PWM in the given group, with respect to a large background set. STOP calculates a hyper-geometric *P*-value for the enrichment of each PWM *M* in the given group in a similar way. The hyper-geometric *P*-value is calculated for the size of the intersection between the given group and the full set of putative target genes as induced by $T(M)$. The main difference between the two methods stems from the different way the score thresholds are determined for each PWM. Clearly, the score threshold one uses affects the number of putative target genes in both the selected group of genes and the background, and thus affects the measured enrichment.

In order to compare PRIMA and STOP, we chose as the given group of genes the targets of TF E2F4 that were identified by chromatin immunoprecipitation (ChIP) combined with DNA microarrays (Boyer *et al*., 2005). The genes for which E2F4 is bound in the first 300 bps upstream the TSS were extracted (361 genes). Except the PWM score thresholds, we would like all other parameters used by PRIMA and STOP to be equal. Thus, we used the same background set of genes (the genes represented on the HG-U133A chip) and the same 392 human PWMs, when running on the E2F4 target genes. Both methods searched for binding motifs on 300 bps upstream of the TSS. The comparison between the results of STOP and PRIMA is presented in Table 1.

Two PWMs associated with TF E2F4 exist in the TRANSFAC database: V$E2F4DP1_01 for E2F4:DP1 heterodimer and V$E2F4DP2_01 for E2F4:DP2 heterodimer. As listed in Table 1, STOP found the E2F4:DP2 heterodimer as the most statistically enriched PWM (lowest *P*-value). PRIMA found it as the second most enriched PWM. The score threshold STOP determined for E2F4 was 4.82; less than 5.9, the value used by PRIMA. This difference increased the percentage of identified targets in the group from 15% (PRIMA) to 58% (STOP). Thus, STOP's score threshold for E2F4 is high enough to make the percentage of targets in the gene group very high (58%), but in a way that keeps the background percentage low, so that the enrichment *P*-value calculated is very small ($3.11E{-}32$), 10^{15} fold smaller than that given by PRIMA. We conclude that for the TF E2F4, the determination of the score threshold according to the gene expression *Z*-score as is done in STOP is more biologically tuned than determining it by a constant cutoff as is done in PRIMA. This trend, of significantly more identified targets in the group by STOP and better enrichment *P*-values is shared by 4 of the 5 PWMs of Table 1.

The overall TF identification of STOP and PRIMA are quite similar: STOP identified 28 TFs with enrichment *P*-value lower than 0.01, while PRIMA identified 27 such TFs. About 70% of the TFs (20) are identified by both methods; these include E2F4:DP2, E2F, E2F1, NFY, STAT1 and ELK1.

4 DISCUSSION

Reliable determination of the binding motifs of a TF on the promoters of putative target genes is a problem of central importance. Most existing methods use a predetermined PWM of the TF to score candidate sequences. The promoters of genes of interest are scanned, searching for segments that score higher than some threshold. The reliability of the method, measured by the number of false positives and negatives, depends on the value of the threshold (and of course on the PWM) that is used. The threshold values are, as a rule, determined in a heuristic way, without using available experimental data. We introduce here STOP, a new method for determining threshold values for a large number of TFs, using gene expression data. We tested our method and demonstrated that our data-based thresholds determination procedure produces results, listed below, that make biological sense:

(1) The absolute values of the average *Z*-scores, $|\bar{z}|$, are significantly higher when the genes are ordered according to their maximal scores for a TF than when the averaging is over groups of genes assembled on the basis of some random ordering. Therefore, the genes with high $|\bar{z}|$ for a TF are more likely to be co-regulated and co-expressed, leading to average expression values that differ significantly from averages taken over a random group of the same size.

(2) The consistency between human and mouse is correlated with $\bar{z}$; e.g. the hyper-geometric *P*-value for the size of the common target gene group between human and mouse is lower for TFs with higher maximal $|\bar{z}|$. Therefore high $|\bar{z}|$, which is indicative of stronger co-expression and co-regulation, implies consistency between the two organisms. Furthermore, the frequency of occurrences of threshold $\bar{z}$ values with different signs in human and mouse is very low for TFs with high (>2.5) $|\bar{z}|$ values.

(3) For 50% of the TFs the average *Z*-score graphs between human and mouse are highly correlated (correlation >0.82).

(4) Comparison of STOP and PRIMA (Elkon *et al*., 2003) results suggests that determining the score threshold according to the gene expression, as is done in STOP, is more biologically tuned.

These observations and findings support the validity of determining the TF score threshold according to the *Z*-score analysis.

In addition, STOP provides an insight to the functionality of the TFs, i.e. which TFs function mostly as activators (those with high positive *Z*-score value) and as repressors (those with low negative *Z*-score value). Note that for these TFs the correlation between the human and mouse *Z*-score values was high (about 0.7) even though the determination of the threshold was performed separately for completely different sets of genes: human and mouse, and using different expression data. The only common element in the data of the two analyses was ordering the genes according to the maximal score motif of the TF that was found in their promoters.

This work can be extended in several directions. First, the determination of the score threshold according to the *Z*-score could be done in a more delicate way, that takes into account the trends in the *Z*-score graphs (and not just the maximal absolute value). Second, for a substantial fraction of the TFs the average *Z*-score absolute value at the threshold point is not high (e.g. in the Scripps dataset 86 TFs have value <2.5). This could happen, e.g. for TFs which function both as

activator and repressor. For such TFs it would be useful to separate the genes to two groups: up-regulated and down-regulated, and then determine the threshold separately for the two groups. Third, after searching for all putative binding sites of all TFs, cooperation between groups of TFs could be examined by comparing the locations of their binding sites.

ACKNOWLEDGEMENTS

This study was supported by grants from the Israel Ministry of Science, by the Wolfson Family Charitable Trust, London, on Tumor Cell Diversity, by the NCI/NIH grant P01 CA 65930-06, and by the Ridgefield Foundation. This study was done as a partial fulfillment of the requirement for the PhD Degree of Libi Hertzberg, Sackler Faculty of Medicine, Tel-Aviv University.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.*, **57**, 289–300.

Boyer,L.A. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Chiang,D.Y. *et al.* (2001) Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, **17**, S49–S55.

Elkon,R. *et al.* (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.

Hertzberg,L. *et al.* (2005) Finding motifs in promoter regions. *J. Comput. Biol.*, **12**, 314–330.

Kel,A.E. *et al.* (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.

Kel,A. *et al.* (2006) Beyond microarrays: Finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics*, **7** (Suppl. 2), S13.

Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Odom,D.T. *et al.* (2006) Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Sys. Biol.*, **2**, 2006.0017 2006.0017 doi:10.1038/msb4100059 (see http://www.nature.com/msb/journal/v2/n1/full/msb4100059.html).

Quandt,K. *et al.* (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Rao,A. *et al.* (1997) Transcription factors of the NFAT family: regulation and Function. *Ann. Rev. Immunol.*, **15**, 707–747.

Ross,M.E. *et al.* (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, **101**, 6062–6067.

Tsafrir,D. *et al.* (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res.*, **66**, 2129–2137.