

Prediction of Cis-Regulatory Elements of Coregulated Genes

Szymon M. Kielbasa¹

s.kielbasa@itb.biologie.hu-berlin.de

Nils Blüthgen¹

nils@itb.biologie.hu-berlin.de

Christine Sers²

christine.sers@charite.de

Reinhold Schäfer²

reinhold.schaefer@charite.de

Hanspeter Herzel¹

h.herzel@biologie.hu-berlin.de

¹ Institute for Theoretical Biology, Humboldt University, Berlin, Germany

² Institute of Pathology, Charité, Humboldt University, Berlin, Germany

Abstract

We present a computational pipeline to predict cis-regulatory elements composing results based on different algorithms: Clover, Cluster-Buster, an own implementation of human/rat/mouse sequence identity and our ITB algorithm. The procedure uses information from the human genome sequence, NCBI gene annotations, verified eukaryotic promoters (EPD), experimentally proven binding sites (Transfac) and homologies to mouse and rat (HomGL/HomoloGene).

We test the approach on 18 upstream regions of experimentally verified AP-1 target genes. About a half of the known sites belong to high-scoring candidates. Three top-scoring elements are confirmed by Cluster-Buster and high homologies.

The same analysis we applied to genes found to be up- or downregulated due to mutated RAS. We performed a detailed literature and computational search for promoter regions. Indications of overrepresented Elk-1 and AP-1 motifs are found via a comparison with shuffled sequences. In some promoters consistent predictions of clustered binding sites were obtained.

Keywords: promoter, binding site, weight matrix, RAS target genes, phylogenetic footprinting

1 Introduction

Growing databases of mRNA expression profiles increase the need for biocomputational methods to predict regulatory rules in sets of coregulated genes. A typical analysis of a large scale microarray experiment allows at the final stage to classify genes into several clusters, grouping genes which share similar pattern of expression in different experimental conditions or time points. Such nonrandom classification suggests the existence of common regulatory mechanisms.

Yeast (*Saccharomyces cerevisiae*) was one of the first organisms, whose genome was studied using the approach mentioned above. Several groups predicted or verified short DNA sequences bound by transcription factors in the upstream regions of genes assigned to a common cluster.

Unfortunately, problems arise when these methods are directly applied to higher eukaryotic organisms. The number of incorrectly predicted binding sites grows proportionally to the total length of DNA sequences in which regulatory elements are expected to be detected. This observation emphasizes the role of careful preselection of short promoter regions (instead of full upstream sequences in yeast) for further search.

Furthermore, one cannot ignore local properties of the DNA sequences selected. Noticeable changes of GC content (CpG islands versus CpG suppression) or poly-A sequences, known to occur frequently close to transcription start sites, may dominate the scoring procedure and produce unexpected results.

Certainly, to increase prediction quality all potential sources of information should be used. Here we present a method composing a search for similarities to already known factor sites with counting

statistically overrepresented short DNA motifs in the coregulated set. To improve the results further, we combine them with results of cross-species comparison of the promoter regions of homologues.

Last but not least, the presented schema of the regulatory motifs detection pipeline requires compilation of information stored in several databases (Ensembl for human, mouse and rat, Homologene, RefSeq, DBTSS, Transfac, EPD, NCBI, Unigene). Even though the task seems to be well defined, its practical implementation is difficult due to different standards for expressing gene names and base positions within genomic sequences. In order to overcome these technical issues, we have linked our programs to the locally developed HomGL database [1], which provides direct access to homologues, their upstream regions and handles a variety of accession identifiers.

2 Study of AP-1 Regulated Genes

In order to elucidate properties of the prediction algorithms an artificial collection of genes was constructed. From the Transfac database (version 6.0) [12] the table “genes” was scanned and all human entries (having the “HS\$” pattern in the gene identifier) annotated to be recognised by the AP-1 factor were selected. Next, these genes were identified using the human Ensembl service. DNA sequences from 1500 bp upstream to 200 bp downstream around Ensembl reported transcription start site were taken for further analysis. Additionally, using the HomGL database [1] the human genes were mapped to find their mouse and rat homologues.

The table “sites” of the Transfac database was used to find short DNA sequences experimentally found to be bound by the AP-1 factor. In 18 cases it was possible to find an exact match of the substring representing the factor binding site in the selected range of 1700 bp around the transcription start site. It should be noted, that there were no data available, whether an extracted site was the only AP-1 site per gene or the one with the highest affinity.

The last piece of information taken from Transfac database comes from the “matrix” table. Since more than one weight matrix is linked to the AP-1 factor, the one corresponding to the largest number of sequences aligned was taken (matrix identifier M00174, alignment of 56 observations). This single matrix was used as the input file for the regulatory sites prediction programs.

Using this data set the following algorithms were studied:

- the Clover program [4] calculating a similarity score between a weight matrix and a position in a DNA sequence taking into account several advanced background models;
- the Cluster-Buster program [5] detecting multiple and close occurrences (clusters) of matches of weight matrices from a provided set;
- our implementation of a percent identity [11] procedure reporting percent of similarity between homologous DNA sequences [3];
- the ITB algorithm [6] calculating overrepresentation of short DNA words (5,6 or 7 base pairs) in the whole family of DNA sequences in comparison to another (background) set of DNA sequences.

The Clover program was requested to perform calculations with all three background methods provided (mononucleotide DNA sequence shuffling, dinucleotide shuffling, and shuffling of the weight matrix). The score thresholds were lowered (min. motif score 3.0, max. P-value 0.15), since experimentally known hits were often not found with the default values. A similar approach was used with the Cluster-Buster program – the motif score threshold was reduced to 3.0 and the cluster score threshold to 2.5. The ITB algorithm was run with the shuffled AP-1 sequences as the background set and for motif lengths 5, 6 and 7. No significant overrepresentation of DNA words was detected by the ITB algorithm. A detailed analysis of the AP-1 sites revealed that there was no significant excess of 5-mers due to the considerable variety of the sites.

Table 1: Results of AP-1 site search in genes containing an experimentally found AP-1 site. The Clover column shows the rank at which the known site was predicted (– stands for no prediction at all). The Cluster-Buster column contains the “+” symbol, when the experimental site belongs to a reported cluster, “–” when other cluster(s) were reported not containing the site, “0” when no clusters were found. The meaning of %-identity column: “h” – high identity at the experimental site, “m” – average, “l” – low, “0” – no data available.

Gene	Clover	Cluster-Buster	%-identity
Hs.1349	4	–	0
Hs.1832	1	+	h
Hs.1905	> 10	–	m
Hs.202453	5	–	l
Hs.253495	6	0	m
Hs.25647	1	–	0
Hs.375129	–	0	h/m
Hs.385521	10	–	l
Hs.408312	7	–	h
Hs.418083	4	0	0
Hs.435800	1	+	m
Hs.446641	> 10	0	0
Hs.511899	1	+	h
Hs.694	1	–	0
Hs.78465	7	–	m
Hs.83169	> 10	–	h
Hs.856	6	–	0
Hs.89679	3	–	0

The results of the mentioned programs as well as experimentally known sites were summarized together on graphs like Fig. 1, describing each DNA sequence of the AP-1 collection separately. The summary for all 18 sequences is given in the Table 1. Even in this artificial set with enriched AP-1 sites and with a single weight matrix, only half of the sites were detected in the top five match candidates list of the Clover. The top-ranked Clover motifs belong to groups reported by the Cluster-Buster program, i.e. such clusters may be used as an additional indicator of true sites. The human, mouse, rat percent identity signals tend to have values higher than average at the experimentally known sites, and can be used also as an additional indication of true sites.

3 Study of the RAS-Dependent Genes

3.1 The Gene Sets

In order to find candidate genes involved in tumor development two cell lines have been studied [13]: preneoplastic rat 208F fibroblasts and its malignant HRAS-transformed derivative FE-8. Subtractive suppression hybridization technique was used to find gene fragments upregulated or downregulated upon neoplastic transformation. Out of more than 1200 subtracted cDNA clones 244 have been recognized as already known genes.

In the next step human homologues of these genes were searched. Using the BLAST algorithm we accepted 216 alignments with the human Unigene set. Afterwards, based on the whole human genome assembly provided by Ensembl, upstream regions of the found human homologues were extracted using

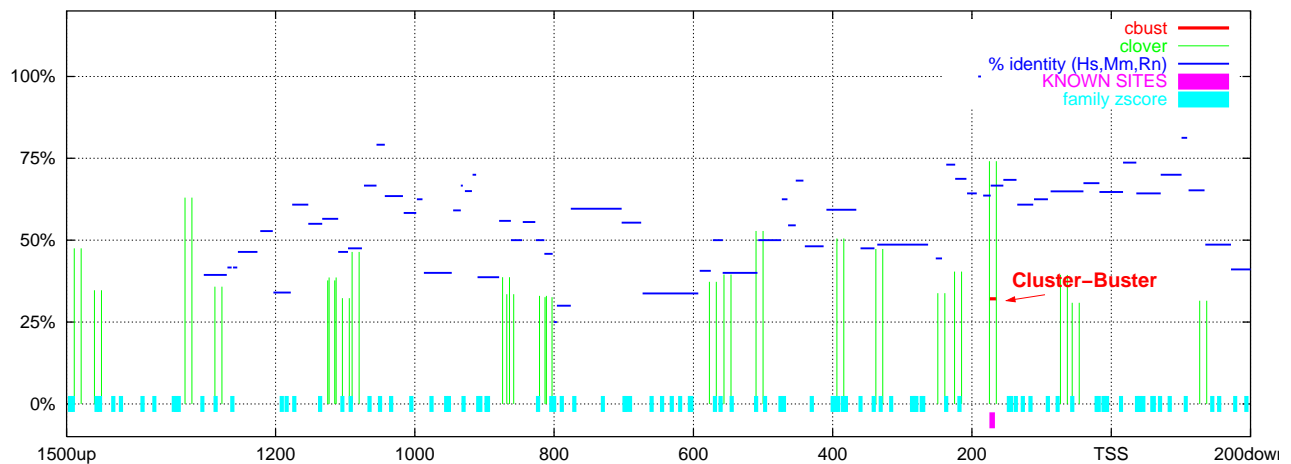


Figure 1: A plot of results of searching for AP-1 matrix in the Unigene sequence Hs.511899 (EDN1, endothelin 1) known to be bound by AP-1 factor. The vertical thin lines bound Clover hits; the arrow shows a cluster reported by Cluster-Buster. The broken horizontal lines represent ungapped stretches of sequence similar to rat and mouse, with similarity provided on the vertical axis. The bars on the horizontal axis mark locations of overrepresented 6-bp long words. The single bar at the bottom of the picture shows the experimentally known AP-1 site. Note: the scales for different programs are not related.

the HomGL system. We downloaded for further analysis DNA fragments from 1500 bp upstream to 200 bp downstream around the transcription start site reported by Ensembl.

For the upstream regions of genes, whose experimentally verified promoter locations could not be found by literature search (or Eukaryotic Promoter Database [9], NCBI databases), several promoter prediction programs available in the internet (First Exon Finder [2], McPromoter [8], CONPRO [7]) were applied. After this step the analysis was limited to genes, for which at least two of the programs predicted transcription start sites (TSS) not further than 300 bp from each other and within the region of 1700 bp downloaded from Ensembl. From those genes we constructed a list of 30 RAS suppressed genes (the “RAS-P” set) and 20 RAS activated genes (the “RAS-Y” set).

3.2 Weight Matrices

The Transfac 6.0 database [12] provides a collection of more than 300 (partially redundant) weight matrices describing sites bound by vertebrate transcription factors. Unfortunately, while searching for binding sites the amount of false signals grows proportionally to the number of matrices studied. Due to this problem we limited our search by preselecting weight matrices mentioned in the literature in the context of RAS regulation. We constructed a subset of Transfac containing 52 weight matrices for our studies below.

3.3 Analysis with Clover, Cluster-Buster and Percent Identity

The prepared RAS suppressed and RAS activated sets of genes were searched for sites matched by the RAS related matrices using exactly the same procedure as in the AP-1 case (characteristic examples are shown in Fig. 2). The number of Clover hits increased significantly in comparison to the AP-1 study due to the larger number of weight matrices analysed. On the other hand in the whole set of 50 genes we could find only 6 regions with strong signals given by all three algorithms. In the case of

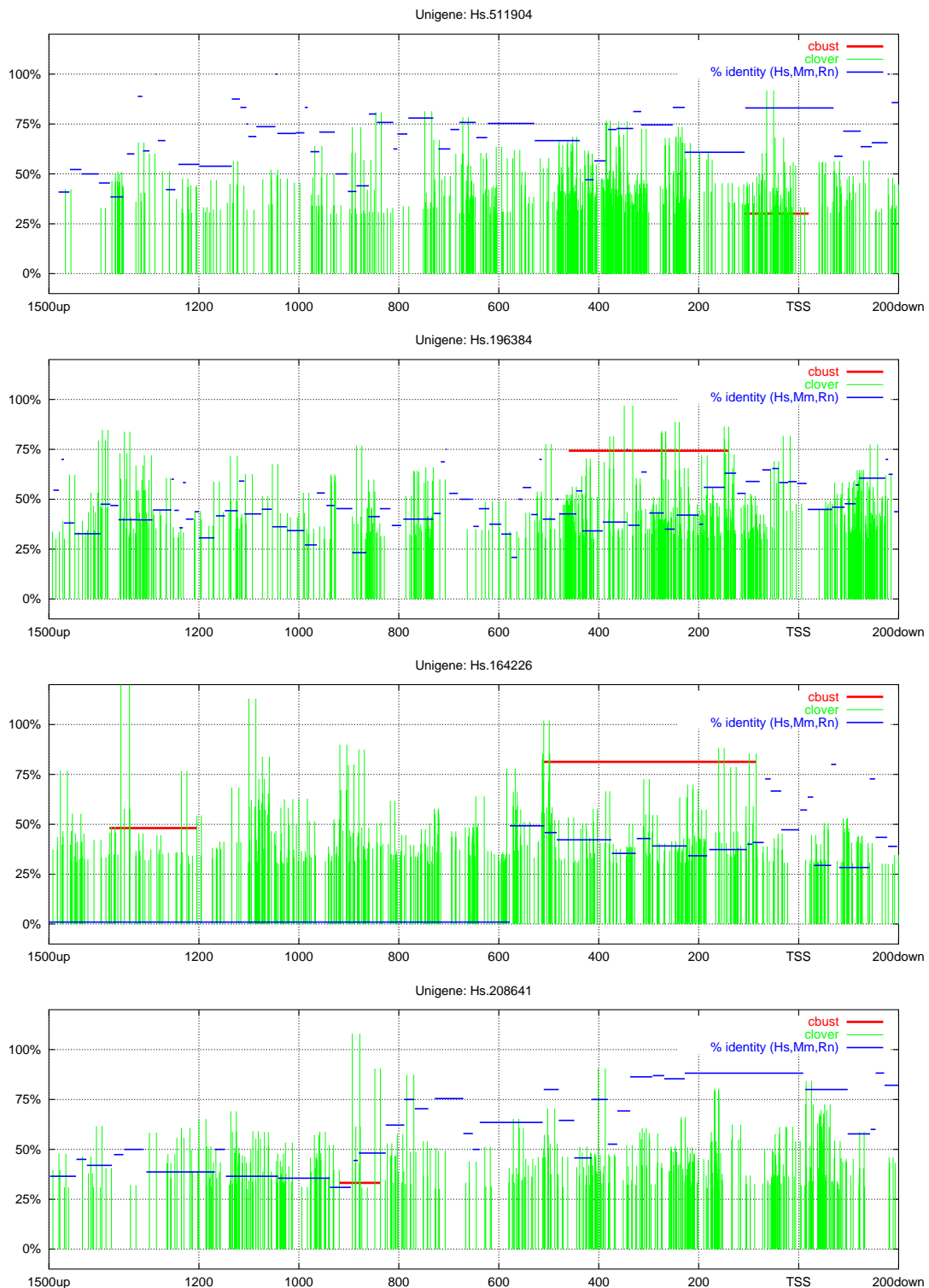


Figure 2: Typical results of regulatory elements search. On the two top charts, (corresponding to genes PTGS2 and EIF4A2 classified to the RAS-P family) it is possible to see regions of higher human-mouse-rat homology and clustered occurrences of predicted binding sites: E2F, TBP (in PTGS2); NF- κ B, AP-4, Sp1, E2F (in EIF4A2). The third graph (THBS1 gene) shows a region with a cluster, but with a lower homology level. On the bottom graph (ACTA2 gene) the regions of high homology and good clustering are mutually exclusive.

AP-1 study 3 out of 8 strong signals given by all methods were related to the experimentally verified sites. Therefore it is likely that some of the six cases of RAS target genes represent true signals.

3.4 Overrepresentation of Binding Sites in Coregulated Genes

The last method studied here considers the possibility that the whole set of genes is a target of the same factor. In such a case the distribution of weight matrix match scores should be different in the original sequence set compared to a random sequence set.

We used the relative entropy score to model factor affinity to a position of a DNA sequence [10]. The score background probabilities were counted locally in a sliding window of 201 bp centered at the studied position.

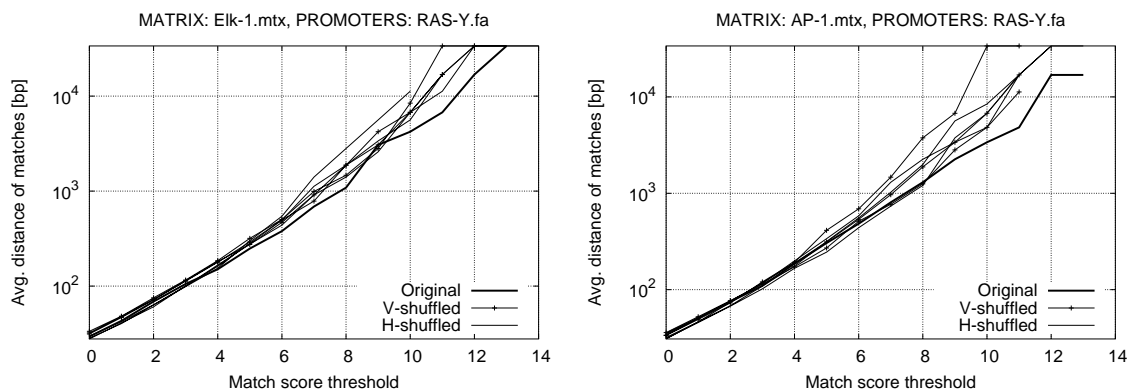


Figure 3: The average distance between weight matrix matches (factors: Elk-1 (left), AP-1 (right)) as a function of minimum match score in the RAS activated genes: original promoters (thick line), horizontally or vertically shuffled promoter sequences (thin lines).

Any analysis of predicted transcription factor binding sites depends strongly on the chosen threshold. In order to study the general case we vary the threshold and count the detected binding sites. Instead of showing the histograms of binding sites directly we plot in Fig. 3 the mean distances between matches of a matrix as a function of the threshold (thick line). Consequently, small distances correspond to many detected binding sites. Thus a lower thick line (obtained from the original sequences) compared to the shuffled sequences (thin lines) indicates overrepresentation of the predicted sites.

In order to find over- and underrepresented matches of Transfac matrices we compare the distance of matches for the original and shuffled RAS sequences. Two methods of shuffling have been studied: *horizontal* mixing of bases inside each promoter separately and *vertical* reordering bases at same positions in TSS-aligned promoters (so this method preserves the GC content as the function of TSS distance). The thick lines are essentially below the thin lines of shuffled promoters. This indicates, that particularly for large thresholds these binding sites are overrepresented.

4 Results and Discussion

The current implementation of the prediction pipeline accepts a list of human coregulated genes, extracts their upstream sequences through HomGL as well as the upstream regions of their mouse and rat homologues. Afterwards the different regulatory element prediction methods are called and their results are presented on charts for each gene of the input coregulated family.

To evaluate the prediction quality we studied the performance of the methods using an artificial data set of AP-1 target genes. Based on annotations provided by the Transfac database a list of

18 genes with known sites bound by the AP-1 transcription factor was compiled. The analysis with three different algorithms resulted in 8 regions with high scores shared by all algorithms – 3 of them corresponded to the experimentally known sites. The other 5 might be false positives or other sites not listed in the database.

The same strategy was applied to two sets of RAS target genes. The algorithms indicate 6 candidate regions in 50 genes, which might contain clusters of regulatory elements.

Further development has to focus on the preparation of more test data sets to improve and evaluate the scoring schemes. The final scoring system should incorporate the affinity of a transcription factor to genes of the family in comparison to appropriately chosen random sets.

In summary we stress that the computational identification of promoters and transcription factor binding sites in higher eucaryotes is still a difficult task. The upstream regulatory regions are large and quite heterogeneous with respect to their composition. Consequently, appropriate background sets are required. Due to the limited affinity of many sites and the high number of potential weight matrices a huge amount of false positive prediction occurs.

Only a careful combination of diverse informations can lead to a successful prediction. First, the search regions and the number of weight matrices should be minimized to reduce false positives. Local background models can minimize artefacts due to heterogeneous promoters. Many functional binding sites appear in clusters as analyzed by Cluster-Buster. Furthermore, homologies to other species can be exploited as another indicator of regulatory signals. Finally, information on coregulation of gene sets can be used to search for overrepresented DNA words and predicted bindings sites.

We have illustrated in this paper how these techniques can be combined to improve the reliability of transcription factor binding site predictions.

References

- [1] Blüthgen, N., Kielbasa, S., Čajavec, B., and Herzel, H., HomGL – comparing genelists across species and with different accession numbers, *Bioinformatics*, 20:125–126, 2004.
- [2] Davuluri, R., Grosse, I., and Zhang, M., Computational identification of promoters and first exons in the human genome, *Nature Genetics*, 29:412–417, 2001.
- [3] Dieterich, C., Wang, H., Rateitschak, K., Luz, H., and Vingron, M., CORG: a database for COmparative Regulatory Genomics, *Nucleic Acids Res.*, 31:55–57, 2003.
- [4] Frith, M., Fu, Y., Yu, L., Chen, JF., Hansen, U., and Weng, Z., Detection of functional DNA motifs via statistical over-representation, *Nucleic Acids Res.*, 32:1372–81, 2004.
- [5] Frith, M., Li, M., and Weng, Z., Cluster-Buster: Finding dense clusters of motifs in DNA sequences, *Nucleic Acids Res.*, 31:3666–3668, 2003.
- [6] Kielbasa, S., Korbelt, J., Beule, D., Schuchhardt, J., and Herzel, H., Combining frequency and positional information to predict transcription factor binding sites, *Bioinformatics*, 17:1019–1026, 2001.
- [7] Liu, R. and States, D., Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling, *Genome Research*, 12:462–469, 2002.
- [8] Ohler, U., Niemann, H., Liao, G., and Rubin G., Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition, *Bioinformatics*, 17:199–206, 2001.
- [9] Perier, R., Junier, T., and Bucher, P., The eukaryotic promoter database EPD, *Nucleic Acids Res.*, 26:353–357, 1998.

- [10] Schneider, T., Stormo, G., Gold, L., and Ehrenfeucht, A., Information content of binding sites on nucleotide sequences, *J. Mol. Biol.*, 188:415–431, 1986.
- [11] Schwartz, S., Zhang, Z., Frazer, K., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W., PipMaker – A web server for aligning two genomic DNA sequences, *Genome Research*, 10:577–586, 2000.
- [12] Wingender, E. Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., and Thiele, S., The TRANSFAC system on gene expression regulation, *Nucleic Acids Res.*, 29:281–283, 2001.
- [13] Zuber, J., Tchernitsa, O.I., Hinzmann, B., Schmitz, C., Grips, M., Hellriegel, M., Sers, C., Rosenthal, A., and Schäfer, R., A genome-wide survey of RAS transformation targets, *Nature Genetics*, 24:144–152, 2000.