BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 16 2011, pages 2239–2247
doi:10.1093/bioinformatics/btr366

*Gene expression*

Advance Access publication June 20, 2011

# RIP: the regulatory interaction predictor—a machine learning-based approach for predicting target genes of transcription factors

Tobias Bauer[1,2], Roland Eils[1,2,*] and Rainer König[1,2,*]

[1]Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), INF 280, 69120 Heidelberg and [2]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, and Bioquant, INF 267, University of Heidelberg, 69120 Heidelberg, Germany

Associate Editor: Joaquin Dopazo

**ABSTRACT**

**Motivation:** Understanding transcriptional gene regulation is essential for studying cellular systems. Identifying genome-wide targets of transcription factors (TFs) provides the basis to discover the involvement of TFs and TF cooperativeness in cellular systems and pathogenesis.

**Results:** We present the regulatory interaction predictor (RIP), a machine learning approach that inferred 73 923 regulatory interactions (RIs) for 301 human TFs and 11 263 target genes with considerably good quality and 4516 RIs with very high quality. The inference of RIs is independent of any specific condition. Our approach employs support vector machines (SVMs) trained on a set of experimentally proven RIs from a public repository (TRANSFAC). Features of RIs for the learning process are based on a correlation meta-analysis of 4064 gene expression profiles from 76 studies, *in silico* predictions of transcription factor binding sites (TFBSs) and combinations of these employing knowledge about co-regulation of genes by a common TF (TF-module). The trained SVMs were applied to infer new RIs for a large set of TFs and genes. In a case study, we employed the inferred RIs to analyze an independent microarray dataset. We identified key TFs regulating the transcriptional response upon interferon alpha stimulation of monocytes, most prominently interferon-stimulated gene factor 3 (ISGF3). Furthermore, predicted TF-modules were highly associated to their functionally related pathways.

**Conclusion:** Descriptors of gene expression, TFBS predictions, experimentally verified binding information and statistical combination of this enabled inferring RIs on a genome-wide scale for human genes with considerably good precision serving as a good basis for expression profiling studies.

**Contact:** r.koenig@dkfz.de

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

Received on December 10, 2010; revised on May 4, 2011; accepted on June 13, 2011

## 1 INTRODUCTION

Human gene regulation involves numerous mechanisms comprising protein–protein interaction, DNA binding and transcription, epigenetic DNA modifications, RNA interference and translation. Central to this is the specific binding of transcription factors (TFs) to promoters of genes to regulate their transcription, and the discovery of such regulatory interactions (RIs) to reconstruct large-scale regulatory networks is a main focus of systems biology research. So far, several hundreds of TFs have been identified for many species (Matys *et al.*, 2006). Some TFs bind exclusively to distinct DNA sequence motifs at specific conditions, whereas others are ubiquitously active (Farnham, 2009). Chromatin immunoprecipitation (ChIP) assays have been used to infer TF binding to the promoter of the investigated gene. This was scaled up by ChIP-on-chip technology to obtain the location of specific TF binding genome wide. However, results from such investigations strongly depend on the studied cellular system and treatment. Besides this, computational methods were developed and applied to predict transcription factor binding sites (TFBSs) independent from the samples under study (Stormo, 2000; Valen *et al.*, 2009). These predictions were mainly based on motif searches with position weight matrices (PWMs). PWMs are probabilistic representations of a frequency distribution of nucleotides at each position of a binding site. In contrast to ChIP-on-chip assays, genome-wide PWM searches detect potential TFBSs for any TF (for which a PWM has been assembled from experimentally discovered binding sites) independent of conditional restrictions and thus provide information about TFBSs in an unbiased manner. Predictions with PWMs have been effectively applied to identify the relevant TFs and their sets of regulated genes in gene expression data (Segal *et al.*, 2003a; Sinha, 2006). However, TFBS predictions are rather unspecific and therefore come along with high false positive rates (Stormo, 2000).

Several methods have been developed to construct large-scale regulatory networks using gene expression data, genome-wide ChIP profiles, PWM-scans and a combination of these (Bonneau, 2008). The availability of abundant experimental data for various model organisms enabled to infer regulatory networks for microorganisms explaining and predicting gene expression, e.g. for *Escherichia coli* (Faith *et al.*, 2007), *Saccharomyces cerevisiae* (Bar-Joseph *et al.*, 2003; Joshi *et al.*, 2009; Segal *et al.*, 2003b) and the *Halobacterium* NRC-I (Bonneau *et al.*, 2006). Furthermore, methods were designed to infer significant RIs between TFs and genes using Pearson's correlation and mutual information of gene expression, e.g. the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE; Margolin *et al.*, 2006) and the Context Likelihood of Relatedness (CLR; Faith *et al.*, 2007). ARACNE and CLR were jointly applied to identify target genes for Nrf2 (nuclear factor

*To whom correspondence should be addressed.

erythroid 2-related factor) in the lung of mice in response to oxidative stress (Taylor *et al*., 2008). Recently, the third Dialogue on Reverse Engineering Assessment and Methods (DREAM3) compendium has set up a synthetic data compendium to benchmark several methods inferring RIs. The top five performing algorithms integrated both the provided steady-state (unperturbed, knockdown and knockout) and time-series data (multifactorial perturbations). The performance of the best method apparently depended mainly on predictions reconstructed from steady-state levels of the gene knockout datasets (Marbach *et al.*, 2010) and prediction algorithms for steady-state conditions are mainly required for predicting regulation in tumor samples. A modified version of CLR was among the top five performers and its optimal performance was achieved when using comprehensive knock-out data alone (Madar *et al.*, 2010). The compendium data resembled small subnets of *E.coli* and yeast and provides synthetic transcriptional data. However, the data—like many of the prediction methods—neglect post-transcriptional regulation of TFs. Besides this, the underlying presumption that expression of the target genes depends mainly on the mRNA gradients of their regulating TFs is often violated, specifically in higher eukaryotes. In turn, regulation of TFs on the protein level plays a substantial role, e.g. for hypoxia-inducible factors (HIFs; Kaelin *et al.*, 2002), p53 (Harris and Levine, 2005) and retinoblastoma 1 (RB1; Chen *et al.*, 2009). Additionally, some prediction algorithms are tailored to infer condition-specific RIs for a single or a few TFs rather than predicting RIs for numerous TFs and a wide range of cellular systems and conditions.

To address these limitations, we developed the regulatory interaction predictor (RIP), a supervised machine learning approach that predicts RIs between a large number of human TFs and genes, independent of any specific condition. Our approach distinguishes between TFs and genes and does not presume any dependency of target genes on the gene expression gradients of their regulating TFs. It bases on the knowledge of experimentally derived regulatory interactions in human. For deriving RIs for regulation studies of human cells, RIP has been implemented in a package for the statistical software R and is available for download at http://www.ichip.de/software/RIP.html.

## 2 METHODS

### 2.1 Gene expression analysis

Gene expression data was taken from the CAMDA 2007 dataset containing 5896 gene expression profiles collected from a wide range of human cancer types comprising normal and disease tissue samples which were performed with Affymetrix HG-U133A microarrays (ArrayExpress, www.ebi.ac.uk/arrayexpress, accession E-TABM-185). To get unbiased datasets, we disregarded cell line experiments and experiments with $<10$ samples. Finally, we used gene expression data from 4064 primary human tissue samples of 76 experimental subsets (=conditions) for our meta-analysis. Microarray probe-sets were included in the analysis if they mapped to exactly one gene from the EntrezGene database (Maglott *et al.*, 2007) according to Affymetrix annotations. For each of these probe-sets, the raw expression values were used from the probes located at the 3′ end of their target sequence to minimize RNA degradation effects and reverse transcriptase errors. With this, we obtained expression levels of 13 069 genes for all 76 conditions. Only these genes were considered. Each subset (microarrays of one condition) was normalized using the Robust multi-array average (RMA) method as implemented in the affy R-package (Bioconductor release 2.4, www.bioconductor.org). Pearson's correlation coefficients were computed for each gene pair and condition by a correlation meta-analysis. To account for anticorrelation due to inhibitory signaling propagation, the absolute correlation values were used. We employed a filtering approach adapted from Zhou *et al.* (2005) to select highly correlated gene pairs. The filter consisted of two parameters: The filter consisted of two parameters: correlation coefficient (CC) and fraction of conditions (FoC). When co-applied, they select gene pairs that exceed a defined minimum correlation (absolute value) in a defined minimum percentage of the 76 conditions. CC and FoC each take values between 0 and 1. Applying CC = 0.6 and FoC = 0.25 therefore selected those gene pairs that correlated $>0.6$ (or $<-0.6$) in $>19$ conditions (25%).

### 2.2 Identification of functionally related gene pairs using Gene Ontology

To estimate the functional relatedness between genes in a gene pair, we compared their Gene Ontology (GO) terms. The mapping of GO terms of biological processes was downloaded from EntrezGene (http://www.ncbi.nlm.nih.gov/gene). The GO term hierarchy was taken from the GO.db R-package in Bioconductor release 2.4. Following an approach described elsewhere (Zhou *et al.*, 2005), we constructed 81 functional categories. A GO term was selected as a functional category if its annotation contained $\geq 150$ genes of our analyzed genes and if each of its children contained $<150$ genes, resulting in 81 midrange GO terms. These 81 GO terms described functional categories that were used to estimate the functional relation of gene pairs from the correlation meta-analysis. We assessed a pair of genes as functionally related if they shared at least one of these functional categories.

### 2.3 The gold standard

The TRANSFAC database v2009.2 (Wingender *et al.*, 1996) provided PWMs used for TFBS predictions as well as a collection of TFs and their target genes derived from published experiments. TRANSFAC contained redundant entries for a number of TFs. Therefore, we manually corrected this by pooling TF entries if all their subunits were encoded by the same genes (the same EntrezGene IDs). We further discarded TFs with less than two RIs. This was done because we could not define appropriate validation sets for such TFs (to estimate their performance). Additionally, 72% of these TFs did not have any PWM motif in TRANSFAC (which was needed for several machine learning features, see below). This yielded 303 TFs with 2896 RIs for 949 regulated genes. For all these genes, we had the respective probes on the Affymetrix microarrays and promoters for the TFBS predictions. For the machine learning approach, we defined a gold standard comprising true positive and true negative regulatory interactions (True RIs and True non-RIs). True RIs (2896) were extracted from TRANSFAC and based on published experiments. The remaining 284 641 possible combinations of the 303 TFs and 949 genes were defined as True non-RIs. This is based on the assumption that regulatory networks are sparse and therefore a vast majority of unknown TF–gene pairs are unlikely to interact. It is to note that a large number of interactions may not have been discovered yet. Still, even if one assumes that e.g. only 10% of interactions have yet been discovered, our 'True' non-interactions would comprise ∼26 000 wrongly labeled interactions. This would still be acceptable compared to the much larger amount of remaining ∼258 000 real True non-RIs (out of 284 641 non-RIs).

### 2.4 TFBS predictions with PWM motifs

Promoter sequences were extracted from EntrezGene using the biomaRt package for R (Bioconductor release 2.4). Sequences from the annotated transcriptional start site up to 1 kb upstream were considered. To detect putative TFBS, PWMs were taken from TRANSFAC v2009.2 and PWM-scans were performed as described previously using the cureos R-package v0.3 (Westermann *et al.*, 2008). *P*-values for each prediction were obtained

by comparing its score to 10 000 randomly generated sequences (for details, see Westermann *et al.*, 2008). A motif was considered to be significant if $P < 0.1$. Hits with a $P \geq 0.1$ were discarded.

## 2.5 Defining the features for the classifier

Ten features were calculated to describe discriminating properties of a pair of a TF and a gene (putative RI) based on TFBS predictions from PWM-scans, the correlation meta-analysis and information about co-regulation of genes from the gold standard. The correlation was used to (i) identify all genes that correlate at high levels with a given gene (defining sets of *correlation neighbors*, see below) and to (ii) compare the average correlation of a candidate target gene to known TF targets (or non-targets). The features, therefore, do not presume any dependency between the expression profiles of a target gene and the expression of TF encoding genes.

All possible gene pairs from the gene expression analysis were filtered by applying the two filters CC = 0.6 and FoC = 0.25. For each remaining gene pair, we defined the two genes to be *linked by correlation*. The set of genes with correlation links to a given gene were then designated to be its *correlation neighbors*. Six features for a putative RI were based on correlation neighbors:

(1) Feature one was the number of correlation neighbors of the corresponding gene.

(2) Feature two was the number of correlation neighbors (including the corresponding gene) which were known to be regulated by the corresponding TF (True RIs, taken from the gold standard).

(3) Feature three was $-\log 10(P)$ in which $P$ was the estimated significance for the enrichment of known regulated genes (True RIs in the training sets, taken from the gold standard) in the correlation neighbors (including the corresponding gene) in comparison to all other genes ($P$ was calculated by a Fisher's exact test).

(4) Feature four was the number of correlation neighbors with a significant PWM hit of the corresponding TF.

(5) Feature five was $-\log 10(P)$ in which $P$ was the estimated significance for enrichment of PWM-hits of the TF within the correlation neighbors including the corresponding gene ($P$ was calculated by a Fisher's exact test).

(6) Feature six was the number of correlation neighbors that were known to be regulated (True RIs in the training sets, taken from the gold standard) and which had a significant PWM hit of the TF.

Two features were added describing TFBS predictions and knowledge about co-regulation:

(7) Feature seven was $-\log 10(P)$ in which $P$ was the significance of the PWM hit of the corresponding TF.

(8) Feature eight was the number of genes known to be regulated by the TF (True RIs in the training sets, taken from the gold standard). This feature was added to enable differentiation between universal and specific TFs.

Furthermore, two features were defined considering the correlation of known co-regulated or non-co-regulated genes:

(9) We selected all genes known to be regulated by the corresponding TF (True RIs in the training sets, taken from the gold standard). For each of these target genes, we calculated the median correlation to the corresponding gene of the RI over all conditions (76 conditions). Feature nine was the average of these medians.

(10) We selected all genes which were not known to be regulated by the corresponding TF (True non-RIs in the training sets, taken from the gold standard). Feature 10 was then calculated in analogy to feature nine.

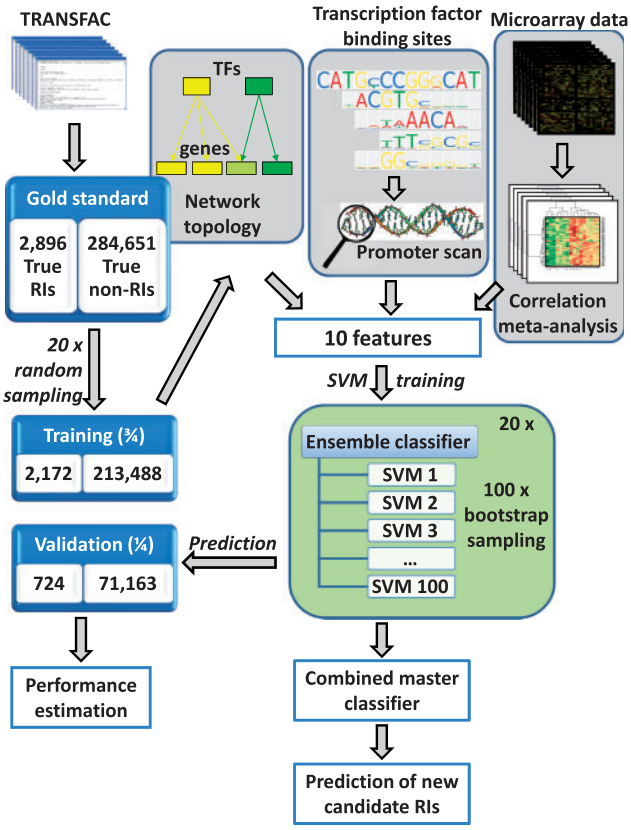


**Fig. 1.** General workflow. Features for inferring RIs between TFs and genes were derived from three different aspects: co-regulation of genes derived from a meta-analysis of gene expression profiles, TF binding site predictions and statistics about a combination of both including experimentally validated binding information from the training set (gold standard). The information of the gold standard was also used to define True RIs and True non-RIs. SVMs were trained with True RIs and True non-RIs and then used to predict new RIs. For training, True RIs and True non-RIs were divided into a training set and a validation set. An equal number of True RIs and True non-RIs were randomly drawn (bootstrapping approach) 100 times and used to train 100 different SVMs yielding one ensemble classifier. Each ensemble classifier was evaluated with its validation set. This procedure was repeated 20 times yielding an averaged estimate about their performances. The classifiers were combined to one master RIP classifier containing 2000 SVMs and applied to predict new RIs.

During training the classifiers, information of known RIs from the validation sets was not used, and all RIs from the validation sets were considered as True non-RIs.

## 2.6 Training, validating and applying the classifier

We performed a 20× 100-fold stratified cross-validation (overview of the workflow is given in Fig. 1). The classifications were performed using support vector machines (SVMs) from the R-package MCRestimate of Bioconductor release 2.4. SVMs with Gaussian kernels were employed. For the training sets, we randomly selected 75% of all True RIs and True non-RIs (2172 True RIs, 213 488 True non-RIs). The remaining of 25% of RIs served for validation to estimate the performance of the classifiers. To optimize the parameters (kernel width and cost function), 75% of True RIs of the training set were drawn with replacement from the 2172 True

RIs and the same amount from the 213 488 True non-RIs of the training set. One SVM classifier was trained with these samples and kernel width $\gamma$ and cost function $c$ were optimized by a grid search employing $\gamma = 2^i$, $i \in \{-10, -8, -6, -4, -2, 0, 2, 4\}$ and $c = 2^j$, $j \in \{-6, -4, -2, 0, 2, 4, 6, 8, 10\}$ using the rest of the training set for validation. This was done by a 10-fold inner cross-validation of the MCR estimate package. To obtain variability and to account for the high amount of True non-RIs, we performed this procedure 100 times yielding 100 SVM classifiers by using different sets of randomly selected True RIs and True non-RIs from the training set. All 100 trained machines were combined and used as one ensemble classifier. The ensemble classifier was validated with the validation set by a voting scheme. Each SVM classifier voted for an RI of the validation set and the minimum number of positive votes was set to define the stringency for the ensemble classifier. During training, True RIs from the validation sets were set as True non-RIs to leave any class-label information of the validation sets untouched (for feature 10, this was done *vice versa* with True non-RIs from the validation sets). The whole process was repeated 20 times using different randomly selected training and validation sets. The overall performance was then estimated from the average of all 20 cross-validations for each stringency threshold. To predict new RIs, all information from the gold standard was employed to train the machines. All trained machines (2000) were taken as one combined master classifier (RIP master classifier) using again the voting scheme in which each SVM contributed one vote. Confidence values of the predictions were calculated from the number of positive votes according to the averaged precision values of the cross-validation.

### 2.7 Enrichment tests for differentially expressed genes of the case study

In the case study, interferon $\alpha$ (IFN$\alpha$) induction was examined using human monocytes and Affymetrix HG-U95Av2 microarrays (Tassiulas *et al.*, 2004). Data were downloaded from the NCBI Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo/, accession GSE1740) and normalized as described above. Differentially expressed genes were determined between samples with and without IFN$\alpha$ stimulation (six samples each) using the significance analysis of microarrays (SAM) and a false discovery rate (FDR) of $<0.01$ (Tusher *et al.*, 2001). One-sided Fisher's exact tests were performed to identify overrepresented differentially expressed genes among the predicted TF-modules (vote cutoff 1600). *P*-values were corrected for multiple testing by the (Benjamini and Hochberg, 1995).

### 2.8 Associating TF-modules to pathways

Predicted TF-modules were associated with pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000). For each TF-module, genes were selected from predicted RIs with $\geq 1600$ votes from the RIP master classifier. One-sided Fisher's exact tests were performed to identify pathways being significantly enriched in genes of the predicted TF-modules. To focus on major signatures, pathways and TF-modules with less than three genes as well as pathways assigned to diseases were not considered. This resulted in 176 pathways of signaling and metabolism for the analysis. All *P*-values were corrected for multiple testing using the Benjamini–Hochberg method.

## 3 RESULTS AND DISCUSSION

### 3.1 Predicting regulated genes of TFs with a machine learning approach

Figure 1 gives an overview of the workflow. True TF–gene interactions (True RIs) were derived from TRANSFAC (Wingender *et al.*, 1996) database (the gold standard) comprising 2896 experimentally well-studied RIs of 303 TFs and 949 genes. All other 284 651 pairs of TFs and genes from these sets were considered as True non-Regulatory Interactions (True non-RIs). For each RI, 10 features were calculated to separate True RIs from True non-RIs. These features were combined from results of a correlation meta-analysis of gene expression (4064 microarrays covering 76 conditions), PWM-scans and information about co-regulation of genes by common TFs (TF-modules). We used these features for training and validation of classifiers (SVMs). As the number of True RIs was sparse compared with the number of True non-RIs, we performed a stratified $20\times$ 100-fold cross-validation for training the classifiers and estimating their performances. Finally, one master classifier was used combining 20 ensemble classifiers (consisting of 100 SVMs each) to predict new RIs.

### 3.2 Genes with correlated gene expression share biological processes

Our prediction method is based on the assumption that genes regulated by the same TF share common cellular processes and thus correlate in their gene expression. To get a quantitative estimate for this assumption, we performed a correlation meta-analysis and compared the correlation of gene pairs with similar function to randomly selected gene pairs. For the correlation meta-analysis, we calculated Pearson's correlation for all possible gene pairs for all 76 conditions. We selected gene pairs at different levels of correlation (adjusting CC and FoC, see Section 2) and estimated their functional relation using GO terms (Ashburner *et al.*, 2000) for biological processes. We considered 81 GO terms for the analysis adapting an approach described elsewhere (Zhou *et al.*, 2005). The 81 GO terms were retrieved by selecting terms that contained at least 150 annotated genes and that had only children with $<150$ annotated genes. This cross-section through the GO graph yielded a well-balanced selection of GO terms which were sufficiently descriptive and specific to describe particular biological functions on a broad range. To quantify the functional relatedness of gene pairs, we defined the functional similarity score (FS-score) as the percentage of gene pairs sharing at least one selected GO term out of all gene pairs. Figure 2 shows the results for different correlation stringencies. Notably, for a wide range of cutoffs (selecting $\leq 5000$ genes, see Fig. 2), the FS-score of our inferred gene pairs was higher than for gene pairs of the gold standard (pairs of genes known to be regulated by the same TF). Our inferred pairs showed FS-scores between 14.8% and 58.3% (stringency parameters CC = 0.6–0.9, FoC = 0.25–0.5, see Section 2) while the gold standard had an FS-score of 35.3%. The FS-score increased with higher stringency (up to CC = 0.8, FoC = 0.35) from 14.8% to 57.3%.

Interestingly, increasing the stringency further reduced the FS-score (at cutoffs for which the number of selected gene pairs was $<300$). We found that this behavior was due to an increased fraction of constitutively expressed gene families (e.g. hemoglobins, histones, immunoglobulins) that show high correlation of expression between each other without sharing any common biological processes. We compared these results to 100 000 randomly selected gene pairs sampled from the same gene-pool. For these random samples, the FS-score was distinctively lower (11.2%). These results demonstrate that genes with correlated expression were considerably often involved in similar cellular processes, which was comparable to gene pairs known to be regulated by the same TFs.

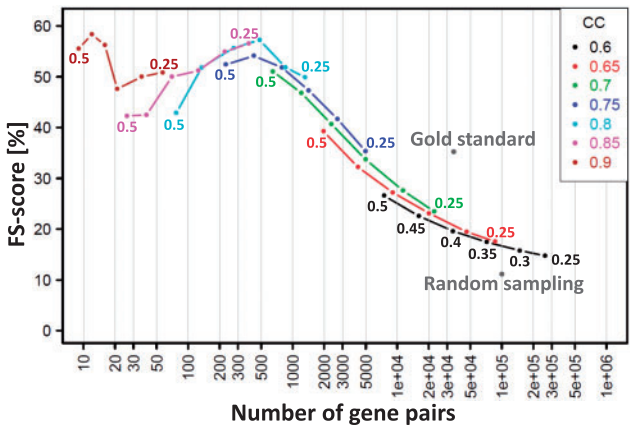


**Fig. 2.** Gene pairs with high expression correlation are functionally related. The graph shows the FS-score which is the percentage of gene pairs sharing at least one functional category for a variety of different stringency criteria [Pearson correlation (CC) and fraction of classes (FoC)]. For example, setting the threshold for CC to 0.85 in >25% (FoC = 0.25) of the datasets yielded 380 annotated gene pairs of which 56.6% (=215) shared the same functional GO category. For comparison, the gold standard comprised 35.3% (12 176 out of 34 538) pairs having at least one functional GO category in common and only 11.2% of 100 000 randomly selected gene pairs had common functional GO categories.

### 3.3 Groups of correlated genes are frequently regulated by common TFs

The correlation of gene pairs served as a good basis for deducing features of RIs for a machine learning approach to distinguish True RIs from True non-RIs. For this, we defined correlation links for gene pairs with sufficiently high expression correlation. We chose CC = 0.6 and FoC = 25% as a robust cutoff yielding the most correlation links for the highest number of genes, while having an FS-score that was still sufficiently higher than the FS-score of random gene pairs. For a gene of interest (of an RI), we defined genes as its correlation neighbors if they had a correlation link to that gene. We calculated six features for RIs based on correlation neighbors (features 1–6) and two additional features containing the averaged correlation over all conditions (features 9 and 10). For example, we considered the number of correlation neighbors that were known to have a True RI to the TF of interest. As expected, if there was a True RI between a gene and a TF, the number of correlation neighbors with True RIs of the same TF was higher compared with True non-RIs (Fig. 3a). The other features based on correlation neighbors yielded similar discriminative power between True RIs and True non-RIs. Table 1 contains the results for all features.

### 3.4 Predicted TFBSs discriminate known RIs from the bulk

To find putative TFBSs, the promoter of each gene was scanned for known binding motifs (using PWMs). We found that True RIs had more often a putative TFBS of the particular TF than True non-RIs ($P < 4.6E\text{-}86$ using a Wilcoxon test). Moreover, within the group of correlation neighbors of a gene with a True RI, putative TFBSs of the corresponding TF were more significantly enriched in comparison to True non-RIs (Fig. 3b). In total, we derived four features from

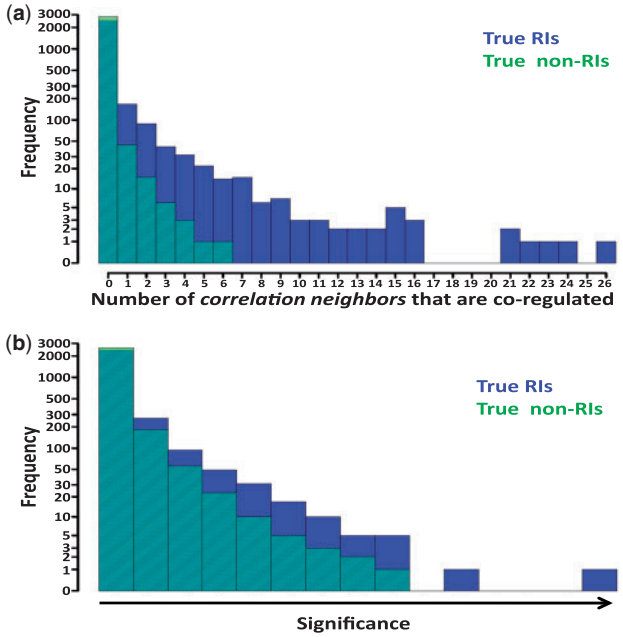


**Fig. 3.** Distributions for True RIs and True non-RIs of two selected features. (**a**) Frequency distributions are given for feature two of True RIs (blue bars) and True non-RIs (green bars, overlap with blue appears dark green). Feature two was the number of correlation neighbors which were known to be regulated by the corresponding TF (True RIs, taken from the gold standard). The higher the feature value, the more correlating genes are known to be regulated by the TF. Genes of True RIs had more correlation neighbors that were regulated by the same TF than genes of True non-RIs. (**b**) Frequency distribution of feature five: $-\log 10(P)$ and $P$ was the significance of enrichment of correlation neighbors with TFBSs of the TF. True RIs showed more significant enrichment of correlation neighbors with TFBSs than non True RIs. For comparability, counts for True non-RIs were stratified to the total number of True RIs in this figure.

**Table 1.** *P*-values of Wilcoxon rank sum tests for all features

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *P*-value | 5.1 E-03 | <4.6 E-86 | <4.6 E-86 | 1.5 E-50 | 4.6 E-86 | <4.6 E-86 | <4.6 E-86 | 1.5 E-50 | 7.2 E-55 | 1.1 E-02 |

TFBS predictions (features 4–7). All four features showed highly discriminative power distinguishing True RIs from True non-RIs (Table 1).

### 3.5 Classifier performance

The 20 training sets of True RIs and True non-RIs were assembled. For each training set, all features were calculated and 100 SVMs (denoted as one ensemble classifier in the following) were trained by a cross-validation. Each of the 20 ensemble classifiers predicted a separate validation set from the gold standard. The results of all 20 ensemble classifiers were averaged (as an estimate of their performance) and compared with the performance of conventional PWM-scans. We were specifically interested in correctly predicting RIs which was estimated by the precision (rate of true positives

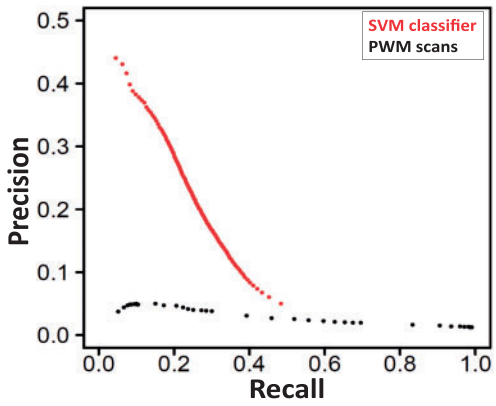


**Fig. 4.** Recall–precision curve of the ensemble classifiers (red) and the PWM-scans (black). Mean values of all 20 ensemble classifiers were calculated to estimate precision and recall for the master classifier. We obtained recall levels between 4.5% and 48.4% and precision levels between 44.0% and 5.0% by decreasing the threshold of required positive votes from 100 to 1. In contrast, PWM scans only yielded precisions of <5% at all stringencies.

out of all positively classified). Notably, the classifiers showed very good precision. At the most stringent cutoff (all 100 SVMs voted positively), the classifiers reached a precision of 44.0% (recall: 4.5%, accuracy: 99%, specificity: 99.9%). This level of precision is considerably good regarding that regulatory networks are sparse, i.e. the number of True RIs is substantially lower than the number of True non-RIs (∼1:99 in the gold standard). For the lowest stringency (only one classifier needed to vote positively), we got the best recall of 48.4% (precision: 5%; accuracy: 90.1%; specificity: 90.6%). The recall–precision curve for all stringencies is shown in Figure 4. We analyzed the features of True RIs that were never classified positively and compared them to True RIs that were always correctly classified. More than 50% of these misclassified True RIs did not have any significant PWM hits within the promoter regions, whereas >99% of the correctly classified True RIs had significant PWM hits. In contrast to our ensemble classifiers, PWM-scans alone were below 5% precision at all stringency settings, and were therefore considerably outperformed by the ensemble classifiers. It is to note that we estimated the performance of our classifiers rather conservatively. The actual prevalence of True RIs is likely to be much higher than the one obtained from the gold standard as many RIs may not have been discovered so far. Our approach was designed to discover such new RIs and the considerably good precision of the ensemble classifiers implied that combining them to a master classifier suited well to infer new RIs (Section 3.6).

### 3.6 Inferring new regulatory interactions

To infer new RIs on a large scale, we combined all trained 20 ensemble classifiers to one RIP master classifier. We investigated all combinations of the set of 303 TFs and all 13 069 investigated genes yielding 3 959 907 candidate RIs. Features were calculated with the entire training and validation set. The RIP master classifier was applied providing 2000 votes (one vote from each SVM machine) for each candidate RI. Confidence values were assessed from the precisions of the validation set (averages of 20 ensemble classifiers, see Section 4). With the most stringent cutoff, we predicted 6073 RIs with 44.0% confidence. With ≥31.5% confidence (cutoff of ≥1600 votes), we yielded 73 923 RIs for 301 TFs and 11 263 genes. Supplementary Table S1 contains all predictions with this cutoff. Supplementary Table S2 provides an overview of the numbers of predicted RIs for different cutoffs. The chosen cutoff of 31.5% confidence and 17.7% recall yielded a sufficient number of genes from newly predicted RIs while potentially avoiding too many false positives, and these predicted RIs were used for further investigations. We compared the performance of RIP to the established methods CLR (Faith *et al.*, 2007) and ARACNE (Margolin *et al.*, 2006) using the same expression data as input. However, neither of the algorithms reached acceptable precision levels at any stringency (details of the method and results can be found in Supplementary Table S3.) We were interested why they performed comparably poorly. These methods were based on direct relationships between the expression profiles of TFs and their target genes. This assumption may well hold in lower organisms, but it was not appropriate for our human expression data. Only 2.2% of all true RIs showed an absolute overall correlation (computed over all 4064 microarray experiments) >0.6 between the expression profiles of the TF coding genes and their targets (see Supplementary Table S3). These findings support the utility of RIP, in particular, for regulation analysis of human cells.

**Table 2.** Association of predicted TF-modules with differentially expressed genes upon IFN$\alpha$ induction in monocytes

| Transcription factor | Differentially expressed | Module size | Percentage | Corrected *P*-value |
|---|---|---|---|---|
| STAT1:STAT2:IRF9 | 20 | 28 | 71.4 | 6.95e-23 |
| IRF1 | 58 | 1187 | 4.9 | 5.72e-03 |
| IRF2 | 15 | 169 | 8.9 | 1.07e-02 |
| STAT1 | 67 | 1513 | 4.4 | 1.15e-02 |
| GAF | 3 | 5 | 60 | 1.15e-02 |
| NFKB1 | 36 | 681 | 5.3 | 1.59e-02 |
| STAT3 | 23 | 384 | 6 | 3.21e-02 |
| IRF7 | 4 | 17 | 23.5 | 3.53e-02 |
| ETS1 | 48 | 1065 | 4.5 | 3.53e-02 |
| RELA | 37 | 762 | 4.9 | 3.53e-02 |
| IRF3 | 4 | 18 | 22.2 | 3.53e-02 |
| ELF2 | 3 | 9 | 33.3 | 3.70e-02 |
| SPI1 | 24 | 439 | 5.5 | 4.63e-02 |

### 3.7 Applying the inferred regulatory interactions to a microarray gene expression study: identifying TFs responsive to IFN$\alpha$

A typical application of our predicted RIs is to investigate the association of TFs and their regulated genes (TF-modules) to a list of differentially expressed genes from a microarray study. A TF can be associated to this list of differentially expressed genes, if the genes of the TF-module are significantly enriched in the gene list. We used microarray data from a study investigating the effect of IFN$\alpha$ on monocytes (Tassiulas *et al.*, 2004). The dataset contained six samples treated with IFN$\alpha$ and six reference samples. We identified 241 significantly differentially expressed genes (FDR < 0.01). These genes were significantly enriched ($P < 0.05$ of a Fisher's exact test) in 13 of the predicted TF-modules (Table 2). On top of the resulting

list were TFs known to respond to IFN$\alpha$. The most significant enrichment was found for the heterotrimeric TF-complex IFN-stimulated gene factor 3 (ISGF3, $P = 6E\text{-}23$) for which 20 out of 28 predicted target genes were differentially expressed. ISGF3 consists of the subunits STAT1, STAT2 and interferon regulatory factor (IRF) 9. It is activated by cytokines and inflammatory factors (Tassiulas *et al.*, 2004). ISGF3 mediates the transcriptional activation of IFN-inducible genes dependent on IFN$\alpha$ treatment (Fu *et al.*, 1990). Furthermore, we found enrichments for IRF1, IRF2, IRF3, IRF7 and STAT3. Together with ISGF3, their response to IFN$\alpha$ treatment is mediated by the JAK–STAT signaling pathway. Two nuclear factor kappa b (NF$\kappa$B) subunits (NFKB1 and RELA) and three E-twenty six (ETS)-domain TFs (ETS1, ELF2, SPI1) completed the list of associated TF-modules. Specific roles for all these TF classes have been described in monocytes (Brach *et al.*, 1993; Friedman, 2007) and in IFN signaling of T-helper cells (Grenningloh *et al.*, 2005). SPI1 plays a central role in monocyte and granulocyte development. It interacts with IRF4 and IRF8 upon phosphorylation, and IRF8:SPI1 complexes bind to an ETS/IRF composite element containing an SPI1 binding site. NF$\kappa$B family members are key regulators of the inflammatory response in monocytes, and AP1 family members cooperate with SPI1 in gene regulation in erythroid cells (Friedman, 2007). We compared our predictions of ISGF3 target genes with predictions employing only PWM-scans. The stringency of the PWM-scans was chosen ($P \leq 0.005$) to get a number of predicted target genes that was comparable to the master classifier. The PWM-scan yielded 29 target genes for ISGF3, only 4 out of which (=13.8%) were differentially expressed in the IFN$\alpha$ study. Our predictions from the machine learning approach were considerably more sensitive to infer genes and TFs involved in the gene regulatory processes of IFN$\alpha$ response (our predictions: 20 differentially expressed out of 28 predicted ISGF3 target genes = 71.4%).

Additionally, we analyzed the expression levels of the TF encoding genes. IRF7 from the listed 13 TFs showed strong differential expression at a significant level and was upregulated in the IFN$\alpha$-induced cells. IRF1, IRF2, STAT2, STAT3 and RELA had slightly (but consistently) elevated expression levels, whereas IRF3 was slightly decreased in the IFN$\alpha$-induced cells.

### 3.8 Genes were highly enriched in pathways which were associated with TFs regulating these genes

To investigate the functional relevance of our predicted RIs, we associated the predicted TF-modules with signaling and metabolic pathways from KEGG database. We selected RIs from regulated genes which were found in KEGG, yielding 220 TF-modules of 22 345 predicted RIs with 3276 genes. Each gene set of the TF-modules was tested to be enriched in each pathway of KEGG (Fisher's exact test). The results cover a variety of signaling and metabolic pathways (Table 3). Additionally, we performed the same analysis for 565 genes with known RIs in TRANSFAC (which was the overlap of KEGG and TRANSFAC) and provided the results in Supplementary Table S4. Most of the associations with our predicted RIs have been described previously and reflect the biology of the pathways. Potentially novel associations are marked in bold in Table 3. In the following, we describe the findings for cell cycle and proliferation-related signaling pathways (yellow in Table 3) in more detail (other pathways are described in Supplementary Table S5). We observed highly significant associations of E2F1 and E2F4 and their hetero-dimers with TFDP1 and TFDP2 to cell cycle and related pathways. Their function in cell cycle progression has been extensively reported in the literature (see e.g. Weinberg, 2006).

**Table 3.** Associations of predicted TF-modules with pathways (new associations are given in bold)

| Transcription factors | Pathway |
|---|---|
| IRF1, IRF2, IRF3, IRF5, **IRF7**, STAT1, STAT3 **NFATC2**, **NF-GMa**, CD28RC, HMGA1 **NF$\kappa$B**, **NFKB1**, **NFKB1:RELA**, **RELA** JUN, CEBPA, CEBPB | Cytokine–cytokine receptor interaction |
| IRF7, **STAT4**, STAT1:STAT2:IRF9 CD28RC, NFATC2, **NF-GMa**, POU1F1 | Jak–STAT signaling pathway |
| IRF1, **IRF3**, **IRF7**, **NF$\kappa$B**, **RBPJ** | Toll-like receptor signaling |
| **NFATC2**, **NF-GMa** | Fc epsilon RI signaling pathway |
| NF-AT, **NFATC2**, **NF-Gma**, SPI1 | Hematopoietic cell lineage |
| IRF2, **NF-AT**, **NF-AT1** | T cell receptor signaling |
| ELF1 | Natural killer cell-mediated cytotoxicity |
| IRF1, IRF2, LEF1, XBP1 **RFX2**, **RFX3**, **RFX5:RFXAP:RFXANK** | Antigen processing and presentation |
| IRF1, XBP1 **RFX2**, **RFX3**, **RFX5:RFXAP:RFXANK** | Cell adhesion molecules (CAMs) |
| ETS1, STAT1 | MAPK signaling pathway |
| IRF2, **NFKB1:RELA** | Apoptosis |
| SP4 | Calcium signaling pathway |
| **TCF7L2** | Wnt signaling pathway |
| **p53**, p73 | p53 signaling pathway |
| **E2F:DP**, E2F4, NFYA | Cell cycle |
| E2F:DP, E2F1:TFDP1/TFDP2, E2F4 | DNA replication |
| E2F:DP | Purine metabolism |
| E2F:DP, E2F4 | Pyrimidine metabolism |
| E2F:DP, E2F1:TFDP1/TFDP2, E2F4 | Nucleotide excision repair |
| E2F1:TFDP1/TFDP2, E2F4 | Mismatch repair |
| GATA4, **NR5A1** | C21-steroid hormone metabolism |
| **NR5A1** | Androgen and estrogen metabolism |
| NR1H4, **PPARA:RXRA**, **RXRA** | PPAR signaling pathway |
| **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3** | Retinol metabolism |
| **NR1I2**, **RXRA:NR1I2** | Linoleic acid metabolism |
| **HNF1A**, **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3** | Drug metabolism—cytochrome P450 |
| **HNF1A**, NFE2 **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3** | Metabolism of xenobiotics by cytochrome P450 |
| FLI1, HNF1B, SMAD3 | ECM–receptor interaction |
| HNF1B | Focal adhesion |
| **NFE2L2** | Cell junctions |
| NR1H3, SP4 | Glutathione metabolism |
| | Neuroactive ligand–receptor interaction |
| | Non-homologous endjoining |
| RARB | Proteasome |
| | Protein export |
| | Oxodative phosphorylation |

Nuclear transcription factor Y alpha (NFYA) was also associated with these pathways. The binding of E2F is dependent on an adjacent CCAAT site being occupied by NFY (Zhu *et al.*, 2004), and functionality of NFY has been shown for G2/M transition of the cell cycle in combination with p53 (Imbriano *et al.*, 2005). p53 and its family member p73 were associated with the pathway of p53 signaling. TCF7L2 was associated with the Wnt signaling pathway, which is in accordance with the functionality of TCF7L2 in that canonical pathway. In summary, these findings well support the functional relevance of the predicted RIs for these pathways.

### 3.9 Predicted RIs are supported by comparison with an independent database

To validate our predictions, we compared the predicted RIs ($\geq$1600 votes) to known RIs of 74 TFs from 25 TF families of the Transcriptional Regulatory Element Database (TRED). Details on the procedure and all results can be found in Supplementary Table S6. In brief, we tested for association (by means of overrepresentation) of sets of RIs from TFs in TRED (TRED TF-modules) with our predicted RIs for these TFs (our predicted TF-modules). In 85.4% of all tested TFs, the correct TF-family was among the top three TRED TF-modules, and in 73.5% the actual TF was assigned correctly (again considering the top three hits). These results further validated RIP using an independent source of RIs of a significant number of well-studied TFs.

## 4 CONCLUSIONS

We presented a novel machine learning approach (RIP) that predicts gene regulation on a genome-wide scale with considerably high precision. The predictions are based on a broad range of conditions and can be applied to more specific experiments as well. Presumably, only a minor fraction of all RIs has been discovered so far. Given knowledge of 2896 True RIs for 949 human genes, a number of 6073 RIs (at the most stringent cutoff) predicted out of 3 959 907 candidate RIs of 303 TFs and 13 069 genes seems reasonable. Also a number of 73 923 RIs (lower stringency, requiring only 80% of positive votes) yielded useful results when applied to the case study of IFN signaling. We employed descriptors for inferring gene regulation from three different aspects: (i) a meta-analysis was performed to obtain groups of genes with high correlation in different cell and tissue types from different experiments. (ii) TFBSs were predicted *in silico* using PWMs to scan promoters of every investigated gene. With these predictions, known transcription factor–gene regulations could be well distinguished from other transcription factor–gene combinations. (iii) Statistical descriptors were used to exploit the association of co-regulation, correlation and TFBS predictions. This information was integrated by a machine learning approach yielding a powerful tool to infer regulatory networks that can be adjusted for recall and precision at a high level of prediction performance. We applied RIP to infer RIs in human. RIP is intended to be an *in silico* approach to extend lists of known and experimentally derived RIs. It needs PWMs (not necessarily extracted from TRANSFAC) of known TFs with known binding sites for initial learning. Thus, other TFs can also be included into the analysis if their binding motifs have been identified in some target genes and a PWM motif could be generated. With that knowledge, it is possible to extend the gold standard and predict RIs for these TFs. If a comprehensive gold standard of experimentally validated True RIs is given (and PWMs for the corresponding TFs), the method can be readily applied also to other well-studied organisms. The presented RIP classifier offers a wide range of applications for gene expression analyses such as identification of key transcription factors and pathways involved in the pathology and changed function of the investigated cells.

## ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Andrea Califano and his team for their cooperativeness and help with the ARACNE algorithm.

*Funding*: BMBF-FORSYS consortium Viroquant (#0313923), Helmholtz Alliance on Systems Biology (SBCancer), and the Nationales Genom-Forschungs-Netz (NGFN+) for the neuroblastoma project ENGINE (#01GS0898).

*Conflict of Interest*: none declared.

## REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

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

Bonneau,R. (2008) Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.*, **4**, 658–664.

Bonneau,R. *et al.* (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.

Brach,M.A. *et al.* (1993) Transcriptional activation of the macrophage colony-stimulating factor gene by IL-2 is associated with secretion of bioactive macrophage colony-stimulating factor protein by monocytes and involves activation of the transcription factor NF-kappa B. *J. Immunol.*, **150**, 5535–5543.

Chen,H.Z. *et al.* (2009) Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer*, **9**, 785–797.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.

Friedman,A.D. (2007) Transcriptional control of granulocyte and monocyte development. *Oncogene*, **26**, 6816–6828.

Fu,X.Y. *et al.* (1990) ISGF3, the transcriptional activator induced by interferon alpha, consists of multiple interacting polypeptide chains. *Proc. Natl Acad. Sci. USA*, **87**, 8555–8559.

Grenningloh,R. *et al.* (2005) Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses. *J. Exp. Med.*, **201**, 615–626.

Harris,S.L. and Levine,A.J. (2005) The p53 pathway: positive and negative feedback loops. *Oncogene*, **24**, 2899–2908.

Imbriano,C. *et al.* (2005) Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters. *Mol. Cell Biol.*, **25**, 3737–3751.

Joshi,A. *et al.* (2009) Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, **25**, 490–496.

Kaelin,W.G. Jr (2002) Molecular basis of the VHL hereditary cancer syndrome. *Nat. Rev. Cancer*, **2**, 673–682.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Madar,A. *et al.* (2010) DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS ONE*, **5**, e9803.

Maglott,D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.

Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (Suppl. 1), S7.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Segal,E. *et al.* (2003a) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19** (Suppl. 1), i273–i282.

Segal,E. *et al.* (2003b) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Sinha,S. (2006) On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, **22**, e454–e463.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Tassiulas,I. *et al.* (2004) Amplification of IFN-alpha-induced STAT1 activation and inflammatory function by Syk and ITAM-containing adaptors. *Nat. Immunol.*, **5**, 1181–1189.

Taylor,R.C. *et al.* (2008) Network inference algorithms elucidate Nrf2 regulation of mouse lung oxidative stress. *PLoS Comput. Biol.*, **4**, e1000166.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Valen,E. *et al.* (2009) Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput. Biol.*, **5**, e1000562.

Weinberg,R.A. (2006) *The Biology of Cancer*. Garland Science, New York.

Westermann,F. *et al.* (2008) Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol.*, **9**, R150.

Wingender,E. *et al.* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Zhou,X.J. *et al.* (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol.*, **23**, 238–243.

Zhu,W. *et al.* (2004) E2Fs link the control of G1/S and G2/M transcription. *EMBO J.*, **23**, 4615–4626.