Published online January 12, 2005

272–279 *Nucleic Acids Research, 2005, Vol. 33, No. 1*
doi:10.1093/nar/gki167

Inferring combinatorial regulation of transcription *in silico*

Nils Blüthgen*, Szymon M. Kiełbasa and Hanspeter Herzel

Humboldt University, Institute for Theoretical Biology, Invalidenstrasse 43, 10115 Berlin, Germany

Received September 5, 2004; Revised November 3, 2004; Accepted December 15, 2004

ABSTRACT

In this paper, we propose a functional view on the *in silico* prediction of transcriptional regulation. We present a method to predict biological functions regulated by a combinatorial interaction of transcription factors. Using a rigorous statistic, this approach intersects the presence of transcription factor binding sites in gene upstream sequences with Gene Ontology terms associated with these genes. We demonstrate that for the well-studied set of skeletal muscle-related transcription factors Myf-2, Mef and TEF, the correct functions are predicted. Furthermore, starting from the well-characterized promoter of a gene expressed upon lipopolysaccharide stimulation, we predict functional targets of this stimulus. These results are in excellent agreement with microarray data.

INTRODUCTION

The regulation of transcription is a major mechanism controlling the spatial and temporal activity of genes, thereby governing the organization of biological processes in eukaryotic organisms. A complex signaling machinery transduces external and internal stimuli to the activities of transcription factors which are the major means of transcriptional regulation. Through this, eukaryotic cells are equipped to adapt adequately to the environment and to orchestrate events like proliferation and differentiation. In contrast to prokaryotes, where transcriptional regulation can be understood in terms of induction by single factors, the regulation in eukaryotes is mainly carried out by sophisticated interactions of multiple transcription factors. Additionally, the regulatory sites are distributed over large regions of the genome including intronic sequences (1,2). It is rare that individual binding sites are strongly conserved, only the combinatorial action gives rise to a specific control. Therefore, understanding complex gene regulatory networks in higher organisms is an extremely difficult task.

Considering the importance of transcriptional regulation and the vast amount of genomic data available, automated inference of the gene regulatory network is a major challenge in the post-genomic era. However, a straight forward search for transcription factor binding sites represented by consensus sequences or weight matrices leads to the curse of false positives. Wasserman and Sandelin (3) estimate that a simple search for binding sites results in only one functional site per 1000 predictions. Consequently, other available biological properties of gene regulation have to be exploited to improve computational predictions. For instance, groups of co-regulated genes from expression profiling (4,5), phylogenetically conserved regions (6,7) and the clustering of binding sites (8) are studied. Nevertheless, the specificity of binding site prediction is still unsatisfactory (3) and expensive experimental studies such as ChIP on chip experiments (2,9,10) are necessary.

In this paper, we propose a functional view on the gene regulatory network by utilizing the growing systematic representation of expert knowledge compiled in the Gene Ontology (11). We use public software to extract upstream regions of genes (12) and to predict clusters of binding sites (8). Then the genes with a common cluster in their upstream regions are searched for statistical association with annotations from the Gene Ontology. For this purpose, a novel program GOSSIP (Gene Ontology Significance Statistical Interpretation Program; N. Blüthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel and D. Beule, submitted for publication) that takes precisely multiple sample correction into account is used. This approach allows prediction of biological functions controlled by combinatorial action of transcription factors. In a preceding paper, we have demonstrated that this approach works for predicting the functional regulation of both a single factor and a combination of two well-characterized factors (13). Here, we address the inference of more complex, combinatorial regulation by several transcription factors. First, we test and verify our algorithm using a well-studied set of transcription factors that co-regulate skeletal muscle gene expression (14). It turns

*To whom correspondence should be addressed. Tel: +49 30 2093 9112; Fax: +49 30 2093 8801; Email: nils.bluethgen@itb.biologie.hu-berlin.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

out that without using a priori knowledge about the functional targets of these factors, we can predict their function correctly. Subsequently, we show that our approach can also bridge the gap between detailed studies of the regulation of single genes and genome-wide analysis. Fessele *et al*. (15) have unveiled the transcription factors that differentially regulate the expression of the chemokine RANTES upon lipopolysaccharide (LPS) stimulation in monocytes. We apply our framework to this set of transcription factors and compare the predicted functions with profiles of publicly available microarray data. The results show a remarkable similarity, although the microarray data have been generated in a different organism and after a long stimulation, allowing also indirect regulation.

MATERIALS AND METHODS

Algorithm

The analysis presented in this paper combines two algorithms (see Figure 1). First, we perform a genome-wide search of the genes that are potentially regulated by the factors under consideration. For this purpose, the Cluster-Buster program (8) is applied to predict clusters of transcription factor binding sites in upstream regions of the genes. To avoid problems arising from subjective parameter tuning, we use the default parameters of the program.

Second, we test whether the genes with a predicted cluster are associated with biological processes. This is performed by GOSSIP (N. Blüthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel and D. Beule, submitted for publication) using Gene Ontology annotations (11). This algorithm tests each term in the Gene Ontology for enrichment in the annotations of genes from a test group compared to those from a reference group. Here, the test group contains the genes for which Cluster-Buster reports a cluster of binding sites. The reference group is composed of all genes under study. Using the one-sided Fisher's exact test, which is based on the hypergeometric distribution, GOSSIP calculates a *P*-value for the null hypothesis that the annotations for the test group are sampled randomly from the reference group. Since this test is performed on all terms, problems arising from multiple testing have to be taken into account. Therefore, we decided not to use single-test *P*-values but to take the false discovery rate (FDR) as an adequate measure of significance. The FDR(α) quantifies the expected number of false discoveries <NFD(α)> in relation to the total number of positives NP(α) at a single test *P*-value threshold α:

$$\mathrm{FDR}(\alpha) = \frac{\langle \mathrm{NFD}(\alpha) \rangle}{\mathrm{NP}(\alpha)}, \qquad \mathbf{1}$$

Using the hypergeometric distribution, the expected number of false discoveries $\langle \mathrm{NFD}(\alpha) \rangle$ can be calculated for each *P*-value threshold α by

$$\langle \mathrm{NFD}(\alpha) \rangle = \sum_i \sum_j^{p_f(j,Z_i,T,N)<\alpha} h(j,Z_i,T,N), \qquad \mathbf{2}$$

where the first sum runs over all terms i from the Gene Ontology, and in the second sum the probability of j genes being annotated with term i is summed up as long as the one-sided Fisher's exact test $p_f(j,Z_i,T,N)$ does not exceed α. Z_i denotes the number of genes annotated with the term i in the reference group. T and N denote the number of genes in the test group and in the reference group, respectively. $h(j,Z_i,T,N)$ represents the hypergeometric distribution:

$$h(j,Z_i,T,N) = \frac{Z_i!T!(N-Z_i)!(N-T)!}{N!j!(Z_i-j)!(T-j)!(N-Z_i-T+j)!}. \qquad \mathbf{3}$$

Within this paper, we set the threshold α such that the FDR is kept below 5%. Further details and the GOSSIP software program are available at the website http://itb.biologie.hu-berlin.de/~nils/gossip/.

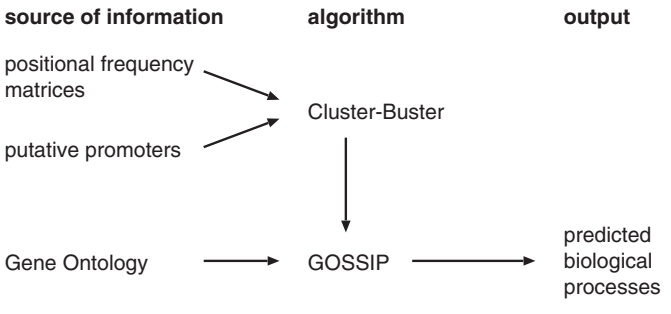

Figure 1. Data flow in our method: Using the Cluster-Buster algorithm, we search for clusters of binding sites in putative promoter regions. The list of genes having a cluster in their promoter are then passed to GOSSIP, which detects association with biological processes using the Gene Ontology. The significantly associated processes are reported.

Data preparation

For 16 032 human UniGene clusters, we extracted sequences upstream of the transcription start sites reported by Ensembl (16). We found 15 362 unique upstream regions since several UniGene clusters pointed to the same genes in Ensembl. We treated the duplicates as single genes and joined their Gene Ontology annotations. We tested sequences of lengths 250, 500, 750, 1000, 1250, 1500, 2000 upstream of the TSS and found that using 1000 bp showed terms with the lowest *P*-values (see Supplementary Figure S2), and other lengths yielded no additional terms. This is in agreement with the estimate by Dieterich *et al*. (17) that the majority of promoters should overlap with these regions.

The Gene Ontology defines a hierarchical controlled vocabulary to annotate genes. It contains three branches: biological process, molecular function, cellular location. We limited our analysis to the branch describing biological processes. The annotations from the Gene Ontology were assigned to the genes using HomGL (12). Each annotation implies a series of more general annotations upward in the hierarchy of the Gene Ontology, which we also take into account.

In this paper, we analyze two sets of transcription factors. The first set consists of transcription factors involved in skeletal muscle-specific gene expression. It is represented by positional frequency matrices constructed from *in vitro* measurements (for Mef-2, Myf and SRF) and from genes that are not muscle specific (for Sp-1 and TEF) (14). In the second part, we analyze combinations of transcription factors that regulate the RANTES/CCL5 promoter in monocytes upon

LPS stimulation (15,18). This set consists of the transcription factors AP1, CEBP, CREB, ETS, NF-κB (p50 and p65) and Sp-1. It is represented by the Transfac matrices (19) with accession numbers: V$AP1_Q6_01, V$CEBP_Q2, V$CREB_Q4, V$ETS_Q4, V$NFKAPPAB50_01, V$NFKAPPAB65_01 and V$SP1_Q6_01. Additionally, we evaluate the biological validity of the result for the second set by analyzing a microarray data set for LPS-stimulated monocytes generated by the Alliance for Cellular Signaling, available at the signaling gateway microarray data center (http://www.signaling-gateway.org/data/micro/cgi-bin/micro.cgi?expt=operon), with accession numbers: MAE040216Z53, MAE040217Z53, MAE040218Z53, MAE040216Z63, MAE040217Z63 and MAE040218Z63. These are expression profiles of mouse monocytes 4 h after LPS-treatment, including dye-swap and three replications. With GOSSIP we obtain biological profiles for each microarray. In this analysis, the background set consists of the significantly expressed genes ('isWellAboveBG' for both channels) and the set of regulated genes contains genes whose *P*-value was smaller than 0.01 ('logRatioPValue' < 0.01). Both sets are mapped to UniGene clusters with HomGL to avoid multiple entries per gene in the lists, since this could bias the analysis. The profiles for all six microarrays show identical terms associated with the up-regulated genes and none with the down-regulated genes.

RESULTS

Processes regulated by muscle transcription factors

Wasserman and Fickett (14) have studied transcription factor families associated with skeletal muscle-specific gene expression. Having identified transcription factors regulating skeletal muscle-specific genes, the authors have constructed a set containing five positional frequency matrices based on binding sites (Mef-2, Myf, SRF) selected *in vitro* or on promoters that do not play any role in the muscle-specific expression (Sp-1, TEF). The promoters of the skeletal muscle-specific genes were intentionally not used to generate the matrices. Applying our method to all five matrices yields significantly associated biological processes: muscle development (FDR = 0.003), muscle contraction (FDR = 0.008) and B-cell activation (FDR = 0.017). Since the Sp-1 factor is involved in the regulation of many other functions, we performed a detailed study with all 26 matrix subsets containing at least two matrices out of the original five matrices (results are shown in Supplementary Table S1). In 12 cases, no Gene Ontology term was found significantly overrepresented. The set consisting of Mef-2, Myf and TEF matrices was the most specific for muscle contraction. Also the term striated muscle contraction was reported with the highest significance here. The results for these matrices are shown in Figure 2. Interestingly, smooth muscle contraction was not significant in any of the analyzed sets. This finding independently confirms that the selected transcription factors may contribute to skeletal muscle-specific expression (14). Most subsets containing the Mef-2 matrix resulted in terms related to B-cell activation. Transcription factors of the Mef-2 family are differentially expressed in B-cells and these cells have Mef-2C-containing, Mef-2-specific DNA binding complexes, suggesting a possible role for Mef-2C activity in B-cells (20).

From the response of an individual gene upon LPS induction to a functional profile

The chemokine RANTES/CCL5 plays diverse roles in the pathology of inflammatory diseases (15). It is a chemoattractant for T-cells and monocytes, rapidly produced in monocytes after stimulation with LPS. LPS is a cell-wall component of Gram-negative bacteria. Fessele *et al*. (15) have investigated the regulation of RANTES/CCL5 expression upon LPS stimulation in human monocytes. They have found that CREB, CEBP, p50/p65, Sp-1, ETS and AP1 transcription factors bind the RANTES/CCL5 promoter in monocytes differently in untreated and LPS-stimulated cells. Additionally, the authors have built *in silico* promoter models, and have found four genes matching their model (21).

We assume that more LPS-induced genes are regulated by the same factors in monocytes. If this hypothesis holds, we could find functions that are regulated upon LPS stimulation in monocytes by applying our framework to this set of transcription factors. The resulting profile of this set of transcription factors is shown in Figure 3a, statistical details can be found in Supplementary Table S2. We find several biological processes to be significant, including response to stress, and response to biotic stimulus as well as chemotaxis and cell communication. All of them can play a role in the response of monocytes after being exposed to bacterial LPS. The terms response to stress and response to biotic stimulus are more general terms upward of inflammatory response in the hierarchy of the Gene Ontology. Cell communication includes the secretion of chemokines and e.g. chemokine RANTES is up-regulated upon LPS stimulation. Also, a regulation of genes involved in chemotaxis seems plausible, since the macrophages will move towards the bacteria and secrete chemokines to attract other macrophages.

Next, we compare our predictions with microarray data obtained by the Alliance for Cellular Signaling (AfCS) from mouse monocytes 4 h after LPS stimulation. Profiling the list of up-regulated genes with GOSSIP yields a similar pattern of enriched Gene Ontology terms (see Figure 3b) for all six microarrays analyzed. Importantly, we find all of our predicted target functions also in the microarray data except cell communication and cell adhesion. It is not surprising that we find additional functions in the microarray data since we started our analysis from the regulation of only one gene (RANTES). It is likely that there are also other pathways and transcription factors involved in the response after LPS treatment than those which lead to the expression of RANTES. The expression profiles are expected to include the response governed by other pathways leading to more terms.

Next, investigations were made into whether the genes predicted to be regulated by the selected transcription factors are indeed up-regulated in the microarray experiment. Also, we investigated whether filtering with the significant Gene Ontology terms improves the prediction of target genes. To address this, we defined three sets of genes: genes which are present on the microarray, those which have a predicted cluster of binding sites in their upstream regions, and those which are additionally involved in one of the predicted biological processes. For these three sets of genes, we compute the distribution of fold-changes in the microarray experiment, as shown in Figure 4. On the entire microarray, 216 of the 11 617 genes (1.8%) have

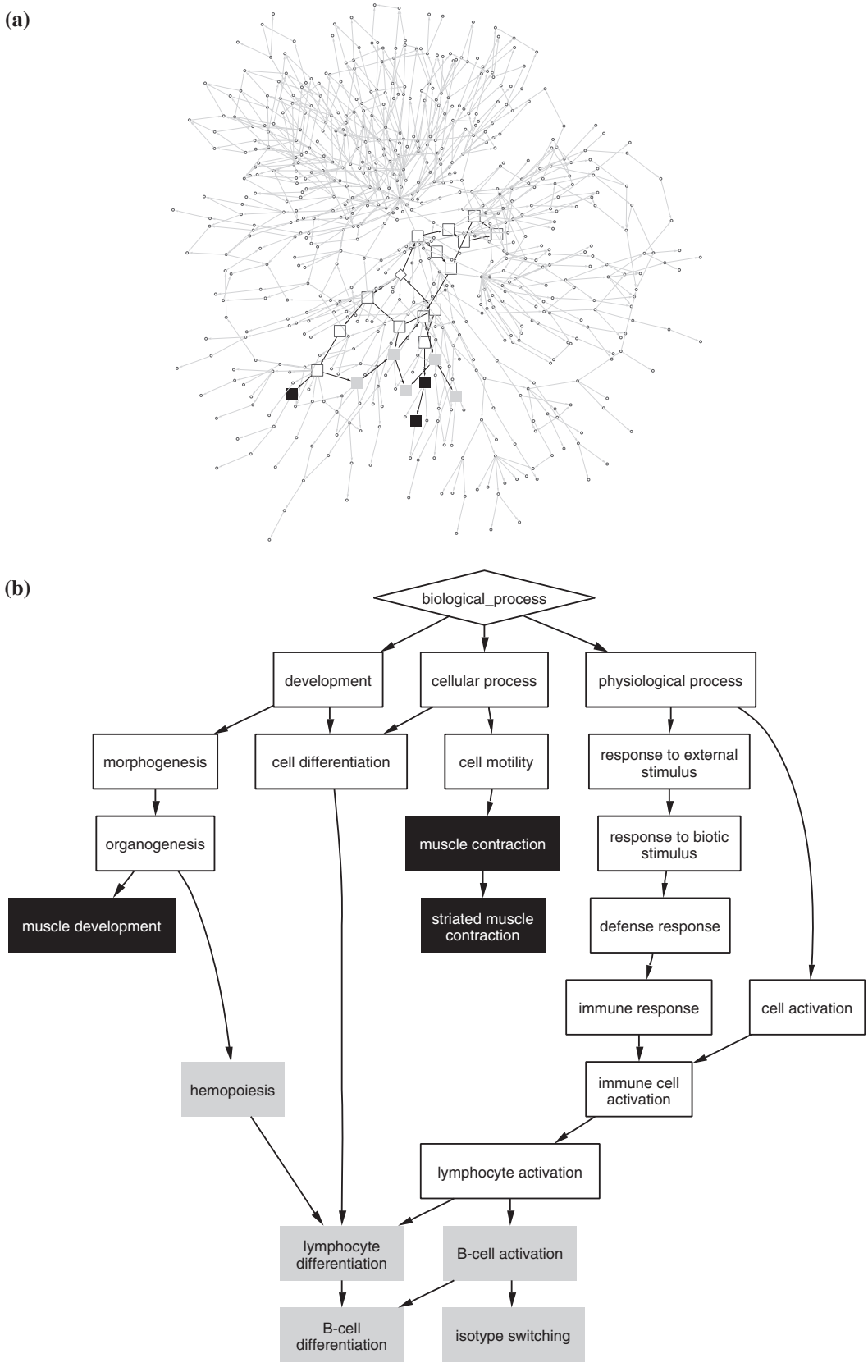

Figure 2. Results for the set of transcription factors Mef-2, Myf and TEF, known to regulate the expression of muscle-specific genes (14). The black and gray boxes correspond to significantly overrepresented biological processes of the Gene Ontology within the predicted target genes [thresholds of FDR $\leqslant$ 0.01 and FDR $\leqslant$ 0.05, respectively]. The diamond shows the root node for biological processes. (**a**) An illustration of the complexity of the analysis: overrepresented terms drawn in the context of all 655 terms assigned to the genes with predicted clusters of binding sites. (**b**) A fragment emphasizing all significantly overrepresented terms.

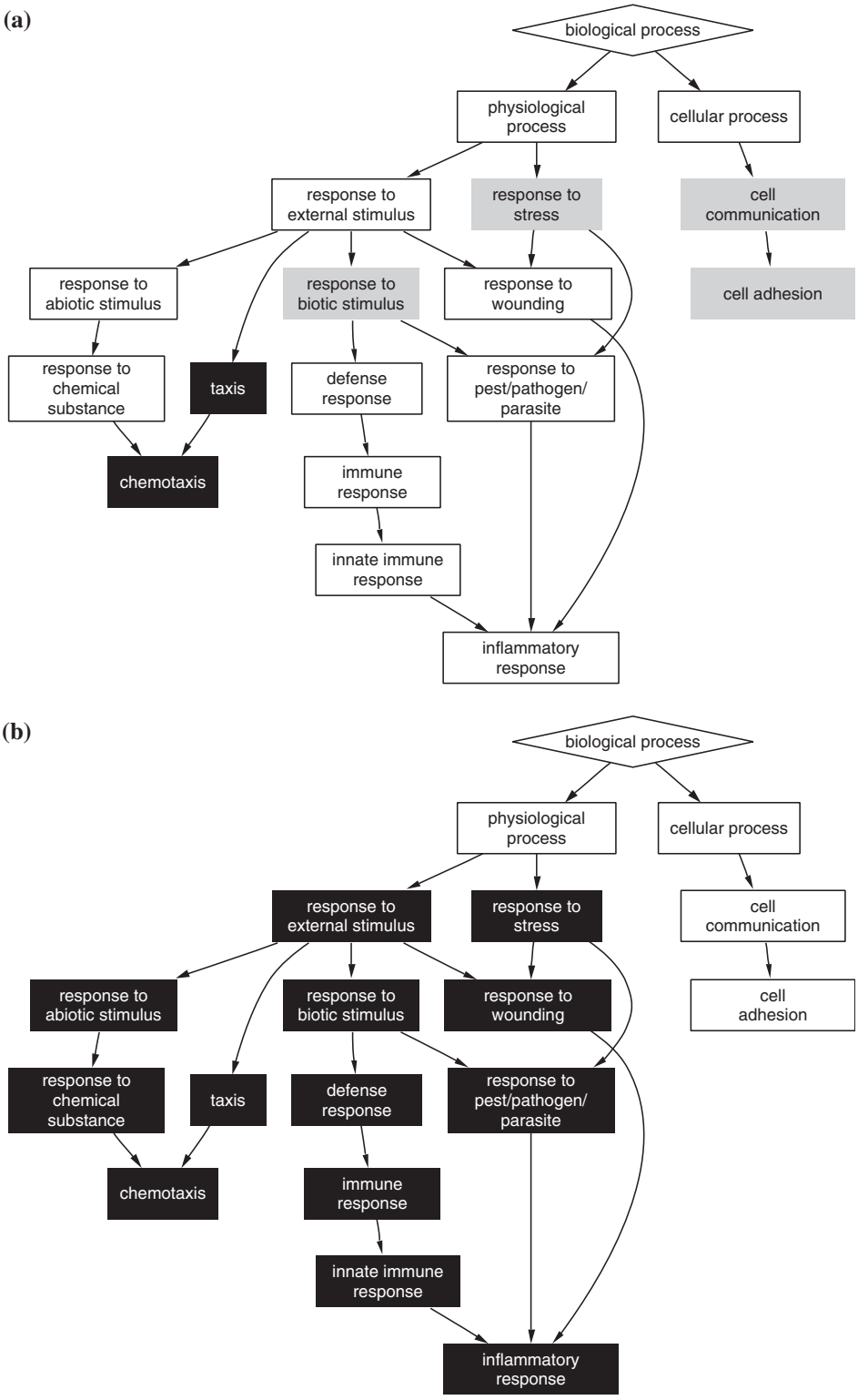

Figure 3. Black and gray boxes indicate the significantly associated biological processes [FDR $\leqslant$ 0.01 and FDR $\leqslant$ 0.05, respectively] with (**a**) the set of transcription factors CREB, CEBP, p50/p65, Sp-1, ETS and AP1 as predicted by our method; (**b**) up-regulated genes upon LPS stimulation in monocytes (microarray data from the alliance for cellular signaling).

a fold change of 2 or higher. Among the 1144 genes from the microarray that have a predicted cluster of binding sites in their upstream regions, 35 genes (3%) are up-regulated (significant enrichment, $P < 0.005$ using χ^2-test). After considering these genes which are additionally annotated with the predicted Gene Ontology terms, the specificity increases further ($P < 0.02$). Out of the 360 genes which match this category, 18 genes (5%) have a fold-change higher than 2. The 784 genes

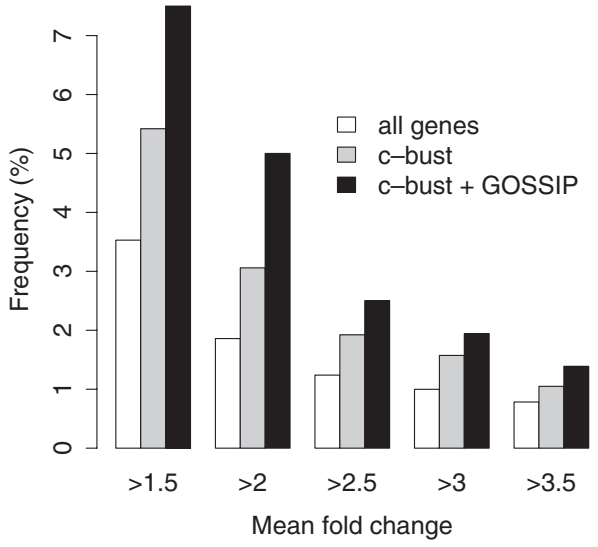

Figure 4. Normalized cumulative histograms of mean fold changes in the microarray data set for all genes (open bars), genes where Cluster-Buster detected a cluster of binding sites (gray bars) and after additional filtering with Gene Ontology (closed bars).

that do not pass the last filtering step do not differ significantly from the overall distribution ($P \approx 0.6$). These results show that a portion of the predicted target genes is up-regulated, and that the intersection with the Gene Ontology significantly improves this ratio. This confirms our initial assumption that many genes respond to LPS through regulation by the same set of transcription factors. Additionally, it shows that the usage of functional annotations can improve the specificity of the genome-wide identification of transcription factor targets. It is particularly interesting, since our prediction has been performed on human sequences and the experiment has been done in mouse monocytes, which reflects a high degree of evolutionary conservation.

Specificity and sensitivity

To assess the specificity of our results, we compared them to results obtained from random sets of positional frequency matrices. These random sets were constructed by permuting the positions of the matrices, thereby preserving their information content and GC-content. The analysis of 1200 sets of permuted muscle-related matrices Mef-2, Myf and TEF yielded 76 sets (6.3%) with one or more significant terms. Interestingly, none of the permuted data sets yielded results related to muscle development or B-cell activation. On average, we found 0.46 false discoveries with $P < 0.0008$ (this is the P-value of the least significant term lymphocyte differentiation). Considering the eight significant terms in the original data set, this corresponds to an FDR of 5.8%.

For the transcription factors mediating the response of the LPS-stimulus, we performed a similar analysis: 545 (27%) out of 2000 sets of permuted matrices were associated with at least one significant term and a total of 2231 terms were significant. However, care must be taken in interpreting these results, since the set contains factors recognizing sites with high-GC-content (Sp-1, NF-κB p50). Although Cluster-Buster uses a background model that takes GC-variation into account, such factors prefer sites in GC-rich upstream regions. This affects the analysis since these regions themselves are associated with certain processes. The 10% of genes having the highest GC-content in their upstream regions are significantly associated with 35 terms describing processes like development/neurogenesis, regulation of transcription, ion transport, phosphorylation and signal transduction (see Supplementary Figure S1). From the significant terms reported for the permuted matrices, 1725 (77%) were identical to the terms associated with GC-rich upstream regions, on average each of the terms occurred in 54 permuted sets. Interestingly, the terms which were not associated with GC-rich upstream regions were reported in only 1.8 sets on average. Therefore, for a correct interpretation, the composition of the positional frequency matrices must be taken into account, as GC-rich matrices can induce more false positives. In such cases, an analysis of permuted matrices can be used to find the expected number of false discoveries. However, the significant terms for the original set of LPS-associated matrices have no significant overlap with those terms that are associated with GC-rich promoters. Furthermore, they are rarely reported in the permutation analysis: in four sets (cell communication), in two sets (taxis, chemotaxis), once (response to biotic stimulus), and in no set (response to stress, cell adhesion).

The lack of functional data for combinatorial gene regulation in higher eukaryotes makes it difficult to construct a true positive set, and, consequently, to estimate the sensitivity of our analysis. However, there are promoters with clusters of binding sites recognized by the same factor, e.g. clusters of E-boxes found in the promoters of circadian clock genes (22). These clusters can be specific enough to unveil their functional targets. For example, clusters of binding sites for the transcription factor E2F can clearly be associated with the S-phase of the cell-cycle (13). Therefore, profiling of single factors might provide a rough estimate of the sensitivity. Applying our method to the 78 matrices for mammalian factors from Jaspar (23) and human upstream regions, we found 20 significant functional profiles with 142 terms in total. As we do not know the true functions, we estimate the number of false profiles by permutation analysis. Here we found on average 0.9 profiles with 4.3 terms. Given these numbers, we estimate that about 19 factors out of the 78 factors under study can be correctly associated with their functional targets. Since it is not known which fraction of the 78 transcription factors exhibit clusters of binding sites in their target genes, this number cannot be translated directly into a specificity.

Many terms of the Gene Ontology, especially the more specific terms, are annotating a few genes only. For example, three genes are annotated with the term isotype switching in the set of 15 362 unique upstream regions. In the example of the Mef-2/Myf/TEF set of transcription factors Cluster-Buster detects clusters of binding sites in 499 out of the 15 362 upstream regions. All three genes that describe isotype switching are among these 499 genes. The P-value of such a coincidence is 0.00012 (Fisher's exact test), and after multiple testing the FDR is 0.017. This demonstrates that terms that annotate only few genes can be significant. As long as the assumption that the false positive genes predicted by the Cluster-Buster analysis scatter randomly with respect to functional association holds, the multiple-testing correction takes the number of annotations into account.

DISCUSSION

The availability of whole-genome sequences and the growing systematic annotations like the Gene Ontology provide the means for more function-oriented data mining beyond the level of single genes. In this paper, we propose an approach, which allows the inference of biological functions regulated by a combinatorial interaction of transcription factors *in silico*. Contrary to other widespread techniques, our method does not intend to predict which factors control genes of similar expression profiles. Instead our search only requires a set of positional frequency matrices representing transcription factors to predict their biological function *in silico*. First, using Cluster-Buster, we predict a list of potential target genes for a set of transcription factors. Afterwards, a rigorous statistical test for association with biological processes implemented in GOSSIP is applied to all biological processes provided by the Gene Ontology. Therefore, the search is not biased by any prior knowledge related to the factors and gives a chance to detect novel regulatory associations.

The well-studied examples of muscle-related transcription factors presented in this paper illustrate the method's utility. As expected, the clusters of sites for the muscle transcription factors are enriched significantly with terms that are skeletal muscle specific. Our method suggests that Mef-2 plays a major role in the context of B-cell activation, which is in agreement with the literature. Notably, no tuning of parameters was necessary within these studies.

Our approach bridges the gap between detailed studies of single promoters and genome-wide approaches. The combinatorial action of transcription factors found to control the gene RANTES was used to predict the regulated functions upon LPS stimulation. The predicted functions show a remarkable agreement with a profile of differentially expressed genes after LPS stimulation in mouse monocytes.

Additionally, the approach provides a gene list supporting the evidence of the reported enriched processes. This gene list can be understood as the cross-section of the genes regulated by the studied factors and genes annotated with at least one of the overrepresented terms. Owing to the filtering property of the cross-section, the final gene list has less false predictions than the primary list of potentially regulated genes, as we have validated with microarray data.

From analysis of random data sets consisting of permuted matrices, we estimate that about 5% of the terms reported to be associated with the targets of transcription factors are chance predictions. This analysis has also shown that the method is less specific in the case when the GC-content of the positional frequency matrices is high, because there are several terms associated with GC-rich upstream regions. If this is the case, then the results have to be interpretated with care, and a permutation analysis can help to estimate the significance. Owing to the lack of true positive data sets, the estimation of the sensitivity is problematic. From studying clusters of the same binding sites for single mammalian transcription factors, we estimate that our method successfully associates biological processes with clusters of transcription factor binding sites in more than 25% of the cases.

Several sophisticated algorithms have been developed to predict regulatory elements in higher eukaryotes. Although they use additional information like expression profiles and phylogenetic footprinting, the number of false predictions remains high. Our paper is not devoted to predicting the regulation of a single gene, but aims to integrate different sources of genome-wide information. It shows that with advanced prediction programs such as Cluster-Buster and the expert knowledge represented by the Gene Ontology, the tools are now in hand to infer regulatory complexities when a rigorous statistic is applied. This genome-wide approach to transcription regulation allows the prediction of functions regulated by the combinatorial action of transcription factors. Additionally, it can filter the list of potential target genes to reduce the number of false discoveries.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dieter Beule, who was involved in the development of GOSSIP, and Martin Frith and Zhiping Weng for providing us the Cluster-Buster program. We thank Didier Gonze and Susanna Akehurst for reading the manuscript carefully and Christine Sers for discussions on the biological impact. N.B. acknowledges support from DFG (SFB 618), and Sz.M.K. from BMBF. We are very thankful for the microarray data from the Alliance of Cellular Signaling and thank Sangdun Choi's and Robert Hsueh's groups for providing these data. Funding to pay the Open Access publication charges for this article was provided by Deutsche Forschungsgemeinschaft (DFG), Sonderforschungsbereich 618.

REFERENCES

1. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
2. Euskirchen,G., Royce,T.E., Bertone,P., Martone,R., Rinn,J.L., Nelson,F.K., Sayward,F., Luscombe,N.M., Miller,P., Gerstein,M., Weissman,S. and Snyder,M. (2004) CREB binds to multiple loci on human chromosome 22. *Mol. Cell Biol*., **24**, 3804–3814.
3. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet*., **5**, 276–287.
4. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet*., **29**, 153–159.
5. Kielbasa,S.M., Blüthgen,N., Sers,C., Schäfer,R. and Herzel,H. (2004) Prediction of *cis*-regulatory elements of coregulated genes. *Genome Informatics*, **15**, 117–124.
6. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nature Genet*., **26**, 225–228.
7. Dieterich,C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003) CORG: a database for comparative regulatory genomics. *Nucleic Acids Res*., **31**, 55–57.
8. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*., **31**, 3666–3668.
9. Ren,B., Cam,H., Takahashi,Y., Volkert,T., Terragni,J., Young,R.A. and Dynlacht,B.D. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G_2/M checkpoints. *Genes Dev*., **16**, 245–256.
10. Martone,R., Euskirchen,G., Bertone,P., Hartman,S., Royce,T.E., Luscombe,N.M., Rinn,J.L., Nelson,F.K., Miller,P., Gerstein,M.,

Weissman,S. and Snyder,M. (2003) Distribution of NF-κB-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA*, **100**, 12247–12252.
11. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
12. Blüthgen,N., Kielbasa,S.M., Cajavec,B. and Herzel,H. (2004) HOMGL–comparing genelists across species and with different accession numbers. *Bioinformatics*, **20**, 125–126.
13. Kielbasa,S.M., Blüthgen,N. and Herzel,H. (2004) Genome-wide analysis of functions regulated by sets of transcription factors. In Giegerich,R. and Stoye,J. (eds.), *German Conference on Bioinformatics 2004*. Gesellschaft für Informatik, Bonn Vol. P-53, pp. 105–113.
14. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
15. Fessele,S., Boehlk,S., Mojaat,A., Miyamoto,N.G., Werner,T., Nelson,E.L., Schlondorff,D. and Nelson,P.J. (2001) Molecular and *in silico* characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. *FASEB J.*, **15**, 577–579.
16. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
17. Dieterich,C., Cusack,B., Wang,H., Rateitschak,K., Krause,A. and Vingron,M. (2002) Annotating regulatory DNA based on man–mouse genomic comparison. *Bioinformatics*, **18**, S84–S90.
18. Fessele,S., Maier,H., Zischek,C., Nelson,P.J. and Werner,T. (2002) Regulatory context is a crucial part of gene function. *Trends Genet.*, **18**, 60–63.
19. Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.
20. Swanson,B.J., Jack,H.M. and Lyons,G.E. (1998) Characterization of myocyte enhancer factor 2 (MEF2) expression in B and T cells: MEF2C is a B cell-restricted transcription factor in lymphocytes. *Mol. Immunol.*, **35**, 445–458.
21. Werner,T., Fessele,S., Maier,H. and Nelson,P. (2003) Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J.*, **17**, 1228–1237.
22. Gekakis,N., Staknis,D., Nguyen,H., Davis,F., Wilsbacher,L., King,D., Takahashi,J. and Weitz,C. (1998) Role of the clock protein in the mammalian circadian mechanism. *Science*, **280**, 1564–1569.
23. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.