BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 4. 326–332
Advance Access publication April 24, 2008
doi:10.1093/bib/bbn016

The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation

Edgar Wingender

Submitted: 21st January 2008; Received (in revised form): 11th March 2008

Abstract

Since its beginning as a data collection more than 20 years ago, the TRANSFAC project underwent an evolution to become the basis for a complex platform for the description and analysis of gene regulatory events and networks. In the following, I describe what the original concepts were, what their present status is and how they may be expected to contribute to future system biology approaches.

Keywords: *TRANSFAC database; transcription regulation; gene expression analysis; pathway analysis*

INTRODUCTION

More than 20 years ago, it became obvious that gene regulation at transcriptional level is one of the most crucial steps in controlling how genetically stored information is processed to determine biological reality. At that time, specifically DNA-binding transcription factors (or *trans*-acting factors; TFs) were known mainly from the prokaryotic world, whereas in the eukaryotic kingdoms (here mainly animals, plants and fungi), knowledge about general transcription factors was just emerging, and very few factors that were able to specifically recognize certain DNA-sequence motifs were characterized, nearly all of them exhibiting a zinc finger domain such as TFIIIA and estrogen receptor (ER). Thus, at that time, there was the speculation that maybe, prokaryotic and eukaryotic transcriptional regulation fundamentally differ already in the blueprints of the involved protein molecules, since the helix-turn-helix motif of prokaryotic regulators was already well characterized.

On the side of the genes, methods had been developed to identify the specific binding sites (or *cis*-regulatory elements) which are targeted by these transcription factors, such as DNAse I footprinting, gel retardation, methylation interference or protection, just to name the most popular ones. A quickly growing number of such transcription factor binding sites (TFBSs) were characterized and mapped to eukaryotic promoters or enhancers, usually starting with the knowledge about the gene under the respective control. Thus, the idea emerged to collect information about these sites and to generate a regulatory map for the genes and, down the road, whole genomes [1], for the corresponding database the term TRANSFAC was coined soon after [2]. This idea is still valid, and has been picked up with a broadened scope by the much more recently established Encode project [3], now with much more suitable approaches at hand to tackle such a task systematically and on a large scale.

Twenty years ago, it was already quite obvious that with the methods usually applied at that time, only anecdotal research on TFBS could be done. However, even from the sporadic information gathered, we could learn a lot and attempted to generalize our observations, to identify the underlying principles and to apply them for proper predictions, as it is good scientific practice. What became clear very soon as well was that the number of data points needed for such generalizations was on an unprecedented scale, and that all the individual

Corresponding author. Edgar Wingender, Department of Bioinformatics, Medical School, University of Göttingen, Goldschmidtstr. 1, D-37077 Göttingen, Germany. E-mail: e.wingender@med.uni-goettingen.de

Edgar Wingender is Professor and Director of the Department of Bioinformatics of the Medical School of the Georg-August University, Göttingen, as well as President and CSO of BIOBASE GmbH.

observations made had to be collected computationally, i.e. in an appropriate database.

Finally, another idea behind systematically collecting information about TFs, their binding sites and their DNA-binding specificity was to approach deciphering the DNA recognition code of DNA-binding domains. Successful decoding of how the sequence specificity of DNA-binding proteins is achieved, at least within individual TF classes, could enable us to predict potential binding sites for all thousands of TFs where we do not have any information yet on where they bind and what they may regulate.

Except of the last point, the TRANSFAC database has successfully adopted these issues and actually became the conceptual framework or a paradigm of a number of similar, usually more specialized, projects (e.g., refs [4–11]).

TRANSFAC 'CLASSIC'

From the very beginning, it was the idea of TRANSFAC to generate a database that provides information about mapping gene regulatory sites, representing their sequences and interacting proteins (TFs), classifying these proteins according to their DNA-binding domains (DBDs), and summarizing the information about binding specificities into predictive models [1].

In the beginning, the classification of TFs according to their DBD was restricted to zinc finger domains and the number of finger motifs they contained, plus giving hints on presumptive helix-turn-helix motifs (basically speculated only for yeast GCN4, which later on turned out to rather have a bZIP-type DBD with a basic DNA-contacting and a leucine zipper dimerization domain) [1]. Within the years after the original compilation, there was an explosion of data about genes encoding TFs, their amino acid sequences and structures. A comprehensive classification scheme of TFs mainly based on their DBDs was developed and continuously developed further into a scheme of five superclasses with six hierarchical levels underneath and 5658 instances assigned [12, 13]. More recently, a library of hidden Markov models (HMMs) was established for a more systematic and automatized classification of newly appearing TFs [14]. The TRANSFAC classification of DBDs has, for instance, been proven useful in establishing Bayesian network models for the structural features of the corresponding recognition sites [15].

The documentation of TFBSs had to include information about the gene they are linked to and its organism, the position and the sequence of the binding site. Since these features strongly depend on the method applied, this had to be given as well. Also, the set of occupied sites in a given gene obviously varies with the cellular context (cell type, external stimuli applied, cell cycle status, etc.), hence, these parameters were also included in the database.

The more recently developed high-throughput methods such as the chromatin immunoprecipitation (ChIP)-chip approach are promising in providing us with information about TF-target gene relations on a large scale. However, their output is, in most cases, more about binding 'areas' than 'sites'. Nevertheless, this kind of data fits into the concept behind TRANSFAC and has thus been taken up in a separate table of the TRANSFAC database [16]. Usually, it is not trivial to convert the raw data provided in the supplements of the original publications into unambiguous sequence information, and also the reliability of the resulting data may have some limitations. At least, it cannot be expected to find perfect consensus matches for the TF in question in all the precipitated genomic fragments, since it is known that (i) physiologically relevant true binding sites sometimes are extremely degenerate, (ii) some factors may be forced to bind to 'their' cognate sites because of synergistic effects of another factor sitting nearby (see subsequently about composite elements) and (iii) some principal DNA-binding factors may also hook piggyback on top of other, primarily DNA-bound TFs, just anchored there through protein–protein interactions. Furthermore, not all real binding sites necessarily have regulatory relevance; it is well conceivable that some pseudo-sites are recognized by a factor, making it binding there, but without regulating any nearby gene. Other pseudo-sites, that may be hidden by the chromatin structure and, thus, are never exposed to the cognate factor, will not appear in ChIP-chip studies. Both types of pseudo-sites, however, may acquire functionality when artificially placed in a favorable context. This caveat should be borne in mind when dealing with *in vitro* results of transcriptional regulation studies.

Another feature that was more recently included into the TRANSFAC concept were microRNAs (since release 10.4, December 2006). These data are required when expanding the scope from mere transcriptional networks to comprehensive gene regulation networks, since it became clear during

the last years that the post-transcriptional regulation of gene expression plays an important role, for instance, in tumorigenesis [17, 18]. Since the basic logics of miRNA targeting a sequence element within the 3′-UTR of an mRNA is similar to a TF binding to a TFBS in a promoter or enhancer, miRNA molecules were taken up as 'factors' and their target sequences as 'sites'.

It became one of the fundamental aims of TRANSFAC and other, associated databases: not just to provide mere encyclopedic information, but rather making it amenable for problem-solving tasks, here: predicting TFBS and analyzing promoter structures, leading to a layered structure of the whole system as depicted in Figure 1. TFBS information can be used to detect potential additional sites for a factor; therefore, soon after the establishment of the TRANSFAC database, it was equipped with tools for the identification of putative TFBS by consensus and matrix searches. In the simplest case, this can be done by sequence comparison. Sets of binding sites can be summarized into an International Union of Pure and Applied Chemistry (IUPAC) consensus, which can be used for the prediction of TFBSs (PatSearch/Patch [16, 19]), or a simple statistical model such as a weight matrix can be derived and used for this purpose (MatInspector, Match [20, 21]). These methods have frequently been used for the detection of potential TFBS within individual promoters or common to defined sets of promoters, e.g. [22], but also for comparison of newly emerging patterns with known ones, e.g. [23, 24].

These methods were continuously developed further. In particular, methods were developed and continuously optimized that provide suitable thresholds for each individual matrix [16, 25, 26].

During these attempts it became increasingly clear that the search for individual TFBS, even if optimized according to the most advanced state of the art, e.g. by involving more sophisticated mathematical models such as HMMs [27] or variable-order Bayesian networks (VOBNs) [28], will always be limited by the fact that sometimes, functional TFBSs in the genome are extremely degenerate, and therefore nearly impossible to recognize, whereas perfect matches with any model do not seem to

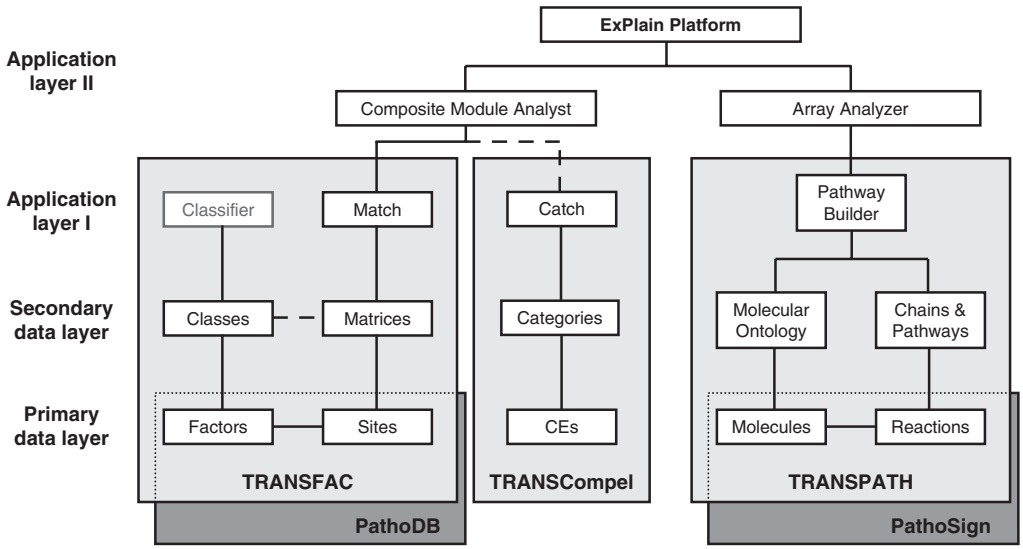

Figure 1: Overall organization of the TRANSFAC system and its integration into a platform for expression data analysis. The primary data layer contains those contents which have been extracted from the primary literature, such as information about TFs and their binding sites for TRANSFAC, information about composite elements (CEs) for TRANSCompel and information about signaling molecules and their reactions for TRANSPATH. The secondary data layer provides information that has been derived from the primary data, as found in the literature or compiled by the database curation staff, such as the classification of TF DBDs and nucleotide distribution matrices, functional categorization of CEs, a signaling molecule ontology or edited reaction chains and pathways. On top of these, there are two application layers with software that works with the contents of the layer underneath (application layer I; note that the DBD classifier has not yet been implemented as part of the TRANSFAC system), or with the results thus obtained (layer II). Also indicated are the existing (straight lines) or planned relations (dashed lines) between these building blocks. Behind the TRANSFAC and TRANSPATH are databases about corresponding pathologically relevant mutations (PathoDB and PathoSign, respectively), which have not been further described here.

represent a functional site; maybe, they are even interacting with the corresponding TF *in vivo*, but even then without functional consequences. One of the most important factors that may make the difference is the context of other TFBS which either provides cooperative binding and synergistic function or, when lacking, precludes an apparently perfect site from becoming functional. Before working on corresponding tools, an appropriate database had to be generated for such contexts as well.

TRANSCompel: THE CONCEPT OF COMBINATORIAL TRANSCRIPTIONAL REGULATION

The importance of the physical and, thus, functional interaction of certain TFBS combinations has been discovered many years ago, and the information about such 'composite elements' was collected and stored in an appropriate database structure, COMPEL. Also this database was continuously developed further to the TRANSCompel resource, and associated tools for the detection of composite elements were added soon as well [29–31]. Proof of principle, however, was already given early in a series of papers of our own group and others, e.g. [32–35].

One particularly successful approach, the Composite Module Analyst (CMA) was recently implemented and published [36]. The underlying genetic algorithm could be demonstrated to work well in a number of studies [36–40]. The basis, however, is still the identification of putative TFBSs by a matrix-based scanning. CMA became the core piece of a recently introduced platform, the ExPlain system, which will be described in a bit more detail subsequently.

PHYLOGENETIC FOOTPRINTING

Given the fuzziness of TFBS recognition patterns, it would be helpful to combine their identification with, e.g. positional weight matrices with an independent criterion. As such, conservation of potential regulatory elements among related genomes became an increasingly popular criterion, the approach usually known as phylogenetic footprinting. No doubt, in cases when a predicted element is 100% conserved, maybe even among more than just a pair of genomes (e.g. human and mouse), and is so in a significantly less conserved environment, everybody would immediately agree that this piece of genomic sequence must have functional relevance. As usual, reality is not that clear-cut, and instead one has to decide which genomes to compare and which thresholds to apply for accepting an element as conserved. Defining statistically reasonable cut-off values in a standard human–rodent comparison, a number of studies concluded that ~70% of all experimentally known TFBSs are conserved [41, 42].

An even more important concern may be the question whether pure sequence-conservation is really the most important criterion, or whether pattern conservation as proven by a still high matrix match score might be much more relevant. Both sets of sites, sequence- and pattern-conserved ones, are largely overlapping (~58% of all sites are conserved according to both criteria), but not identical; thus, sequence-only conserved sites are about 14%, pattern-only conserved ones 11% of all sites documented in the TRANSFAC database. However, these numbers largely vary among the sites for different TFs [42].

During these studies it was interesting to observe that frequently, even positions in a TFBS pattern that are quite variable among instances from different genes may be highly conserved among orthologous genes. It seems that the DNA-binding domains of at least some TFs can interact with a considerable variety of sequences, but that each of these variable sequences may imprint a specifically tuned DNA–protein complex structure that exactly fits to the requirements of the individual 'enhanceosome' (Sauer *et al.*, manuscript in preparation). Moreover, we have observed that many of the ~30% non-conserved TFBS can still be found in the orthologous promoter, although at another position, and therefore are not to be considered conserved. This is consistent with recent reports about turnover of TFBSs and TSSs [43, 44]. Interesting enough, however, is that in particular, *cis*-regulatory elements that are part of composite elements are to a much higher percentage 100% conserved than single sites.

CONNECTION WITH PATHWAY ANALYSIS

The activity of many transcription factors is regulated by modifications, usually phosphorylation, which is brought about in response to the activation of specific signal transduction pathways by all kinds of environmental stimuli. Therefore, we started to establish a database on signal transduction pathways (STPs) the end-points of which are TFs; this resource was termed TRANSPATH [12, 45]. Here as well

we added a tool which combines information about individual reactions to whole pathways and networks ('PathwayBuilder' [46]).

Although this is not the place to explain the features of TRANSPATH in all detail, it should be pointed out that two properties have already proven extremely useful:

(i) to identify all end-points of STPs, which are of gene regulatory relevance and, together with the TRANSFAC resources, all known and potential target genes of a certain stimulus and
(ii) vice versa, to infer regulatory pathways from the internal structure of (a set of) promoters.

In particular, the latter function has been implemented in our ExPlain platform as 'upstream analysis' of gene sets that were revealed in expression (e.g. microarray) experiments. From a given set of IDs of major chip manufacturers, the system automatically retrieves the corresponding promoter sequences, analyzes what their internal structures have in common and, thus, which TFs may be involved in the analysis of this set of genes. Following the STPs which are known to activate these TFs further upstream, it can be frequently observed that these pathways have convergence points, that are individual molecules that might be good candidates to keep specific control of the regulation of the gene set of interest.

The ExPlain platform can combine this specific 'upstream' analysis with different kinds of functional annotation such as Gene Ontology (GO) annotation [47] or disease annotation from the HumanPSD database [48, 49], or with a 'downstream' pathway analysis of the start gene set by mapping them directly onto the STPs stored in TRANSPATH. It was shown recently in a proteomics study, that such a combined 'upstream' and 'downstream' analysis was giving very robust results when characterizing a certain established disease state where it is not easily possible to trace back the causes [50].

ANALYSIS OF TRANSCRIPTIONAL NETWORKS

Systems biology aims at gaining holistic views on complete biological systems such as cells, organs or whole organisms. Still, these views may consider different aspects of the systems under study, such as providing a comprehensive view on the whole metabolism, all the signaling cascades, or the genetic networks, underpinned by quantitative data and aiming at dynamic simulations. However, a successful smooth integration of these different views is still outstanding.

Over the last few years, we have analyzed different kinds of networks for their local and global topological properties. There are first hints that metabolic, signal transduction and transcriptional networks differ in some of these properties. In the course of these studies, TRANSFAC provided the basis for the first comprehensive topological analysis of the mammalian transcription network and a subnetwork (centered around p53). A particularly useful parameter turned out to be the betweenness centrality of the individual nodes, identifying TF genes of special regulatory importance [51]. Suffering from significant incompleteness, however, these investigations will be redone by adding TF–TF gene relations that are predicted by state-of-the art analyses of the TF gene promoters.

NEXT STEPS

Since its beginning about 20 years ago, the TRANSFAC database became a kind of industry standard in its way to represent transcription regulatory data. We, therefore, offer this database as a platform for related data that have emerged from systematic screening studies. Thus, in 2005, data about *Arabidopsis* TFs, which are stored in the DATF database about known and predicted TF sequences, their annotation and classification and expression data [52] were integrated by courtesy of the authors. Similarly, contents of the *Drosophila* DNAse I footprint database about TFBS in this fly were included as well [53]. Both databases were and still are publicly available, independent resources, but there may be users who appreciate having access to these data in the full context of TRANSFAC contents. We propose this as a model for other related contents as well. These resources may have a more specific view, as in the two cases mentioned, and may provide additional information that is important for the specific project in the context of which these data were collected. Allowing integration of those parts which match the TRANSFAC structure provides an additional access point to these valuable data sets.

As mentioned earlier, miRNA data also fit well into the TRANSFAC structure and have been included accordingly. The relations of miRNAs with their (known) upstream control factors and with their downstream mRNA targets have also been integrated with the pathway data in the TRANSPATH database (since release 7.4). Their

integration into a comprehensive network analysis protocol is forthcoming. Another topic under consideration is the uptake of data about chromatin structure, in addition to information about scaffold/matrix attachment sites (S/MARs) that are already collected in the S/MARt DB [54] and that are going to be integrated into TRANSFAC in near future.

We have already addressed the fact that the TRANSFAC system has included ChIP-chip data some time ago. This will certainly be another growth area, in spite of the different pitfalls this type of data inherently has, but it is one of the most promising methodologies to approach the original aim of TRANSFAC, i.e. to provide a genome-wide map of regulatory protein interaction sites, in particular, when it is combined with functional high-throughput assays [55, 56]. Thus, although the structure is already there, standardizing this information still requires considerable manual work and maybe facilitated by proper interaction with the community in the future. Also, data from other technologies such as ChIP-PET (paired-end ditag) results will be included in the future as well.

These high-throughput data may help us to finally define the DNA-binding specificity of many of those TFs where we do not yet have any information about this key property. However, it appears doubtful whether it will ever be realistic to have the DNA-binding specificities of all TFs of all organisms experimentally characterized. Thus, having a clue about the underlying rules for the recognition of specific DNA sequences by certain types and instances of DBDs could yield to predictions of the DNA-binding specificity of yet uncharacterized TFs as well as of hypothesizing which TF (class) may interact with a certain element [57]. We are confident that TRANSFAC as a general platform can further contribute to these aims as well.

Key Points

- The TRANSFAC database provides information about eukaryotic transcription factors, their DNA-binding sites and DNA-binding profiles.
- TRANSFAC is also used as repository of related data collections and high-throughput data such as ChIP-chip data of which more than 141 000 sequence fragments are presently stored.
- Computational knowledge-based tools have been devised to analyze promoter structures.
- Predicted promoter structures can be incorporated into pathway analyses to identify candidate master regulators of a specific biological event.

References

1. Wingender E. Compilation of transcription regulating proteins. *Nucleic Acids Res* 1988;**16**:1879–902.
2. Wingender E, Heinemeyer T, Lincoln D. Regulatory DNA sequences: predictability of their function. In: Collins J, Driesel AJ, (eds). *Genome Analysis - From Sequence to Function; BioTechForum- Advances in Molecular Genetics* 1991;**4**:95–108.
3. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004; **306**:636–40.
4. Ghosh D. TFD: the transcription factors database. *Nucleic Acids Res* 1992;**20**(Suppl):2091–3.
5. Huerta AM, Salgado H, Thieffry D, *et al*. *Nucleic Acids Res* 1998;**26**:55–9.
6. Higo K, Ugawa Y, Iwamoto M, *et al*. PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res* 1998;**26**:358–9.
7. Lescot M, Déhais P, Thijs G, *et al*. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res* 2002;**30**:325–7.
8. Davuluri RV, Sun H, Palaniswamy SK, *et al*. AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 2003;**4**:25.
9. Münch R, Hiller K, Barg H, *et al*. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res* 2003;**31**:266–9.
10. Steffens NO, Galuschka C, Schindler M, *et al*. AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in Arabidopsis thaliana. *Nucleic Acids Res* 2005;**33**:W397–402.
11. Zhao F, Xuan Z, Liu L, *et al*. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res* 2005;**33**:D103–7.
12. Heinemeyer T, Wingender E, Reuter I, *et al*. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res* 1998;**26**:362–7.
13. Heinemeyer T, Chen X, Karas H, *et al*. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res* 1999;**27**:318–22.
14. Stegmaier P, Kel AE, Wingender E. Systematic DNA-binding domain classification of transcription factors. *Genome Inform* 2004;**15**:276–86.
15. Xing EP, Karp RM. MotifPrototyper: a Bayesian profile model for motif families. *Proc Natl Acad Sci USA* 2004;**101**: 10523–8.
16. Matys V, Kel-Margoulis OV, Fricke E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;**34**:D108–10.
17. Mendell JT. MicroRNAs: critical regulators of development, cellular physiology and malignancy. *Cell Cycle* 2005;**4**: 1179–84.
18. Kumar MS, Lu J, Mercer KL, *et al*. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet* 2007;**39**:673–7.
19. Wingender E, Karas H, Knüppel R. TRANSFAC database as a bridge between sequence data libraries and biological function. *Pac Symp Biocomput* 1997;**2**:477–85.

20. Quandt K, Frech K, Karas H, *et al*. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 1995;**23**:4878–84.
21. Kel AE, Gössling E, Reuter I, *et al*. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 2003;**31**:3576–9.
22. Davies SR, Chang LW, Patra D, *et al*. Computational identification and functional validation of regulatory motifs in cartilage-expressed genes. *Genome Res* 2007;**17**:1438–47.
23. Siu FK, Lee LT, Chow BK. Southwestern blotting in investigating transcriptional regulation. *Nat Protoc* 2008;**3**:51–8.
24. Wang X, Gu J, Zhang MQ, *et al*. Identification of phylogenetically conserved microRNA cis-regulatory elements across 12 Drosophila species. *Bioinformatics* 2008;**24**: 165–71.
25. Pickert L, Reuter I, Klawonn F, *et al*. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics* 1998;**14**:244–51.
26. Chekmenev DS, Haid C, Kel AE. P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res* 2005;**33**:W432–7.
27. Marinescu VD, Kohane IS, Riva A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics* 2005;**6**:79.
28. Ben-Gal I, Shani A, Gohr A, *et al*. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* 2005;**21**:2657–66.
29. Kel OV, Romaschenko AG, Kel AE, *et al*. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res* 1995;**23**:4097–103.
30. Kel-Margoulis OV, Kel AE, Reuter I, *et al*. TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res* 2002;**30**:332–4.
31. Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, *et al*. COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res* 2000;**28**:311–5.
32. Frech K, Brack-Werner R, Werner T. Common modular structure of lentivirus LTRs. *Virology* 1996;**224**:256–67.
33. Brazma A, Vilo J, Ukkonen E, *et al*. Data mining for regulatory elements in yeast genome. *Proc Int Conf Intell Syst Mol Biol* 1997;**5**:65–74.
34. Frech K, Quandt K, Werner T. Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol* 1998;**1**:29–38.
35. Kel A, Kel-Margoulis O, Babenko V, *et al*. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol* 1999;**288**: 353–76.
36. Kel A, Konovalova T, Waleev T, *et al*. Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics* 2006;**22**: 1190–7.
37. Kel A, Reymann S, Matys V, *et al*. A novel computational approach for the prediction of networked transcription factors of aryl hydrocarbon-receptor-regulated genes. *Mol Pharmacol* 2004;**66**:1557–72.
38. Moehle C, Ackermann N, Langmann T, *et al*. Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *J Mol Med* 2006;**84**:1055–66.
39. Waleev T, Shtokalo D, Konovalova T, *et al*. Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res* 2006;**34**:W541–5.
40. Kel A, Voss N, Jauregui R, *et al*. Beyond microarrays: finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics* 2006; **7**(Suppl 2):S13.
41. Levy S, Hannenhalli S. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* 2002;**13**:510–4.
42. Sauer T, Shelest E, Wingender E. Evaluating phylogenetic footprinting for human-rodent comparisons. *Bioinformatics* 2006;**22**:430–7.
43. Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. Functional evolution of a cis-regulatory module. *PLoS Biol* 2005;**3**:e93.
44. Frith MC, Ponjavic J, Fredman D, *et al*. Evolutionary turnover of mammalian transcription start sites. *Genome Res* 2006;**16**:713–22.
45. Schacherer F, Choi C, Götze U, *et al*. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* 2001;**17**:1053–7.
46. Krull M, Voss N, Choi C, *et al*. TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res* 2003;**31**:97–100.
47. Harris MA, Clark J, Ireland A, *et al*. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**:D258–61.
48. Wingender E, Hogan J, Schacherer F, *et al*. Integrating pathway data for systems pathology. *In Silico Biol* 2007;**7**: S17–25.
49. Wingender E, Crass T, Hogan JD, *et al*. Integrative content-driven concepts for bioinformatics ''beyond the cell''. *J Biosci* 2007;**32**:169–80.
50. Zubarev RA, Nielsen ML, Fung EM, *et al*. Identification of dominant signaling pathways from proteomics expression data. *J Proteom* 2008;doi:10.1016/j.jprot.2008.01.004.
51. Potapov AP, Voss N, Sasse N, Wingender E. Topology of mammalian transcription networks. *Genome Inform* 2005;**16**: 270–8.
52. Guo A, He K, Liu D, *et al*. DATF: a database of Arabidopsis transcription factors. *Bioinformatics* 2005;**21**:2568–9.
53. Bergman CM, Carlson JW, Celniker SE. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly. *Drosophila melanogaster*. *Bioinformatics* 2005;**21**:1747–9.
54. Liebich I, Bode J, Frisch M, Wingender E. S/MARt DB: a database on scaffold/matrix attached regions. *Nucleic Acids Res* 2002;**30**:372–4.
55. Bulyk ML. DNA microarray technologies for measuring protein-DNA interactions. *Curr Opin Biotechnol* 2006;**17**: 422–30.
56. Hudson ME, Snyder M. High-throughput methods of regulatory element discovery. *Biotechniques* 2006;**41**:673, 675, 677.
57. Narlikar L, Gordân R, Ohler U, *et al*. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics* 2006;**22**:e384–92.