Introduction.

The availability of extended DNA sequences and whole genomes opened up a new direction for bioinformatic studies: investigation of genome-level regularities, search for new genes, analysis of regulatory regions, prediction of gene function on the basis of network construction. Reliable search and prediction tools should be based on confirmed examples and models provided by modern experimental techniques. The enormous amount of experimental data lead to the necessity to collect, systematize, and process these data prior to their application for further analysis and predictions. The databases TRANSFAC®, TRANSCompel® and TRANSPATH® provide a worldwide unique knowledge system on the regulation of gene expression.

Databases.

TRANSFAC® [1] contains information about structure and function of gene regulatory regions; it describes individual DNA binding sites for transcription factors. Functional and structural properties are given in detail for more than 5000 eukaryotic transcription factors, among them about 2600 mammalian factors. Interactions of transcription factors with other proteins, such as co-activators and co-repressors, general factors, histone modifying enzymes, are also part of the data collection as they are indispensable for transcriptional regulation. The data collected in the TRANSFAC® undergo to systematization and processing. Thus, based on the collection, a comprehensive classification of transcription factors according to the type of DNA-binding domain has been developed. Further treatment of sequences of individual binding sites resulted in a unique collection of weight matrices almost completely covering all known DNA patterns of TF binding sites. TRANSFAC® and the accompanying software MatchTM and PatchTM are used for the identification of potential regulatory signals in genomic sequences. Having collected functional properties of the corresponding transcription factors, we can propose expression patterns for genes of interest, as well as to interpret experimentally observed expression patterns and profiles as they may come from gene chip assays.

We have constructed several specific profiles that are part of MatchTM [2]. A profile is a set of matrices with optimized cut-offs designed for function-driven searches. Currently, we provide immune cell-, muscle-, liver- and cell cycle-specific profiles.

TRANSCompel® is devoted to a particular aspect of transcriptional regulation: cooperation between transcription factors that are bound to their target sites within composite regulatory elements (CEs) [3]. CEs contain two or three closely situated binding sites for distinct transcription factors, and actually are minimal functional units providing combinatorial transcriptional regulation. There are two main types of CEs: synergistic and antagonistic ones. In synergistic CE’s simultaneous interactions of two or three factors with closely situated target sites result in a non-additively high level of a transcriptional activation. Within an antagonistic CE two factors interfere with each other. In some cases competition for overlapping sites leads to a mutually exclusive binding. Presently, more than 350 individual CE’s have been collected.

Classification of composite elements is an essential part of the database: functional classification based on the combinatorial regulation provided by a CE, and structural classification based on the type of DNA binding domains of factors involved.

TRANSCompel® database is accompanied by Catch® – a pattern-based sequence analysis tool that enables users to find matches similar to known composite elements within genomic sequences. Information about structure of known composite elements and specific regulation provided by them appears to be extremely useful for promoter prediction and for applied gene engineering as well.

Similar CEs, that are CEs in different genes consisting of binding sites for the same factors, are used to construct models. A model includes the description of two or more individual binding sites by corresponding weight matrices, the distance between them and their mutual orientation as well. A software searching for the models in DNA sequences has been developed.

TRANSPATH® presents information about signal transduction pathways that lead to the modifications of transcription factors and thus regulate gene expression in response to extracellular signals (such as hormones,
cytokines etc.) [4]. It comprises data about the participating molecules and the reactions they undergo. At least one mechanism for cross-coupling of signalling pathways may be provided by cooperative function of transcription factors within composite regulatory elements.

An incorporated tool, PathwayBuilder™, allows to construct all potential pathways based on collected individual pair wise reactions. Connection and integration with the TRANSFAC® database allows to outline the whole pathway between extracellular signal molecules and the genes that respond to these triggers.

A new supplementary database, TRANSPRO, includes sequences of gene 5’ regions. Sequences are extracted from the RefSeq database for those genes that are linked to both HGNC and LocusLink. For each of more than 8,600 genes, TRANSPRO contains 10,000 nucleotides upstream and 1000 nucleotides downstream relative to the first nucleotide of the most 5’ exon in a gene, as it is annotated in the RefSeq. Currently, linking of the TRANSFAC® genes, factors, and TRANSPATH® molecules to the TRANSPRO is in progress. TRANSPRO sequences can be directly searched for potential regulatory elements by sequence analysis tools Match™, Patch™, and Catch®.

Application of the TRANSFAC®-TRANSCompel®-TRANSPATH® knowledge system for the analysis of array experiments.

We suggest several steps that could be applied to a list of molecules which have been identified in a microarray experiment. As the first step, analysis of the structure of regulatory regions of corresponding genes can be performed. Match™, Patch™, and Catch® software can be applied to find potential binding sites for transcription factors and potential composite elements. At the next step, one can use the TRANSPATH® database to find which signalling pathways are responsible for activation or repression of the found transcription factors. The ArrayAnalyzer™ tool that is applied at the next step has especially been designed to analyse data coming from microarray experiments. The ArrayAnalyzer™, based on information collected in TRANSPATH®, enables to find the upstream signalling molecules that could be responsible for coordinated regulation of genes coming from a microarray experiment.

We are currently applying this approach on the analysis of LPS-regulated genes in human neutrophils [5]. We have analysed promoters of 100 upregulated and 50 downregulated genes. For that, we have extracted sequences from -500 nucleotides to +100 nucleotides relative to the start of transcription, and applied Match™ with both immune analysed promoters of 100 upregulated and 50 downregulated genes. For that, we have extracted sequences from -500 nucleotides to +100 nucleotides relative to the start of transcription, and applied Match™ with both immune and cell cycle-specific profiles. To reveal combinations of TF binding sites we applied a new, improved version of ClusterScan tool [6] based on an implementation of genetic algorithms. We have found that matches for 8 matrices are correlated with upregulation, and that of 4 other matrices with downregulation. Among the factors implicated in LPS-dependent gene induction are c-Ets, NF-kappaB, E2F, GATA, TCF-1, Oct. Factors of the Ets and NF-kappaB are correlated with upregulation, and that of 4 other matrices with downregulation. Among the factors implicated in LPS-dependent gene induction are c-Ets, NF-kappaB, E2F, GATA, TCF-1, Oct. Factors of the Ets and NF-kappaB families are well known to be involved in the inducible gene expression in response to a variety of stimuli. GATA3, TCF-1, Oct-2 are specific for immune cells, and E2F is a key regulator of cell cycle. Our results suggest that the combination of these transcription factors is responsible for upregulation of a very specific set of genes in response to bacterial infection. On the next step we have applied ArrayAnalyzer™ to study upstream signalling pathways leading to the activation of the revealed combination of TFs. Thus, TRANSFAC®, TRANSCompel®, and TRANSPATH® present a unique knowledge system that provides information collected from the original publications as well as allows analysis of regulatory sequences, gene networks, and data coming from microarray experiments.

References.