DATABASES ON EUKARYOTIC GENE EXPRESSION: DATA COLLECTION, PROCESSING, AND APPLICATION FOR ANALYSIS.

O. Kel-Margoulis, V. Matys, C. Choi, E. Goessling, N. Voss, I. Reuter, A. Kel, and E. Wingender.*
BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuettel, Germany.
*To whom correspondence should be addressed, oke@biobase.de.

Introduction. 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.

TRANSFAC® is a widely known database on gene transcription regulation. It contains information about structure and function of gene regulatory regions such as promoters and enhancers; it describes individual DNA binding sites for transcription factors. Functional and structural properties are given in detail for more than 4700 eukaryotic transcription factors, among them about 2400 mammalian factors. Interactions of transcription factors with other proteins, such as co-activators and co-repressors, basal factors, histone modifying enzymes, are also part of 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 of transcription factors, a comprehensive classification has been developed. Further treatment of binding sites for individual transcription factors resulted in the unique collection of weight matrices. Having collected functional properties of transcription factors, we constructed several distinct sets of matrices for application to specific genes, for instance, liver-specific or immune-specific genes.

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). CEs contain two or three closely situated binding sites for distinct transcription factors, and actually are minimal functional units providing combinatorial transcriptional regulation. Both specific factor-DNA and factor-factor interactions contribute to the CE's function. 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, about 330 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. 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 three individual binding sites by corresponding weight matrices and the distance between them.

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.). It comprises data about the participating molecules and the reactions they undergo, thus spanning a complex network of interconnected signaling components. At least one mechanism for cross-coupling of signaling pathways may be provided by cooperative function of transcription factors within composite regulatory elements.

Molecules involved in signaling pathways are classified according to their structural-functional organization. An incorporated tool, PathwayBuilder™, allows to construct all potential pathways based on collected individual pairwise 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.

Applications. The databases can be used for learning purposes as a great facts book on gene expression regulation. For scientists performing experiments, the databases are indispensable at all steps of their work: from planning of the experiment up to interpretation of the results.

The databases are widely used by bioinformaticians. TRANSFAC® and the accompanying software Match™ and Patch™ are used for the identification of potential regulatory signals in genomic sequences, for generating potential gene expression patterns of interesting genes, for interpreting experimentally observed expression patterns and profiles as they may come from gene chip assays. The information provided by TRANSCompel™ is used by sequence analysis tools such as Catch™ that analyze genomic sequences for potential composite elements. 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. PathwayBuilder™ and ArrayAnalyzer™, tools accompanying TRANSPATH® database, can display any part of the overall network, for instance starting with a given molecule of interest, or starting from a list of molecules which have been identified in a microarray experiment.