Integrative content-driven concepts for bioinformatics "beyond the cell"

Edgar Wingender[1,2,*] Torsten Crass[2], Jennifer D Hogan[3], Alexander E Kel[1],
Olga V Kel-Margoulis[1], and Anatolij P Potapov[2]

[1]BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany

[2]Department of Bioinformatics, UKG/University of Göttingen, Goldschmidtstr. 1, D-37077 Göttingen, Germany

[3]BIOBASE Corp., 100 Cummings Center, Beverly, MA 01915, USA

*Corresponding author (Fax, 49-551-39 14914; Email, e.wingender@med.uni-goettingen.de)

Bioinformatics has delivered great contributions to genome and genomics research, without which the world-wide success of this and other global ('omics') approaches would not have been possible. More recently, it has developed further towards the analysis of different kinds of networks thus laying the foundation for comprehensive description, analysis and manipulation of whole living systems in modern "systems biology". The next step which is necessary for developing a systems biology that deals with systemic phenomena is to expand the existing and develop new methodologies that are appropriate to characterize intercellular processes and interactions without omitting the causal underlying molecular mechanisms. Modelling the processes on the different levels of complexity involved requires a comprehensive integration of information on gene regulatory events, signal transduction pathways, protein interaction and metabolic networks as well as cellular functions in the respective tissues / organs.

1. Different kinds of networks

For any attempt to describe, analyse and simulate living systems, we have to acquire a comprehensive knowledge about their components and, even more importantly, their relations and interactions. If we are successful in doing so, we may be able to understand the underlying rules how the corresponding systems are organized. There were deserving early attempts to develop systematic approaches, for instance by applying the principles of cybernetics and systems theory onto biological systems, thus laying first theoretical foundations, but real success of these efforts had to await the advent of genomics and other omics mass data. These data reflect different aspects of biological systems organized at different levels of complexity. For instance, genes as one-dimensional strings of nucleotides (or their symbols) may constitute a conceptionally simple network of functional interactions. More sophisticated are the interaction or reaction networks at the molecular level, since the number of components involved is 1–2 orders of magnitude higher ($10^5 - 10^6$ proteins, including splice variants, posttranslational modifications and complexes, versus 10^4 genes). As a consequence of the complex three-dimensional structures and highly versatile functions of proteins, the way how their interaction networks are organized also includes much more variable patterns in forming numerous branching and cyclic constructs (e.g. feed-back, feed-forward, bi-fan, etc.), which altogether form rather complex circuits. And even higher is the potential complexity of cell-cell interactions through all kinds of intercellular communication mechanisms.

1.1 Intercellular networks

In a multicellular organism, the function of distinct organs, tissues and cell types have to be precisely coordinated to ensure proper functioning of the whole system (organism)

Keywords. Bioinformatics; gene regulation networks; intercellular networks; signal transduction pathways; system biology
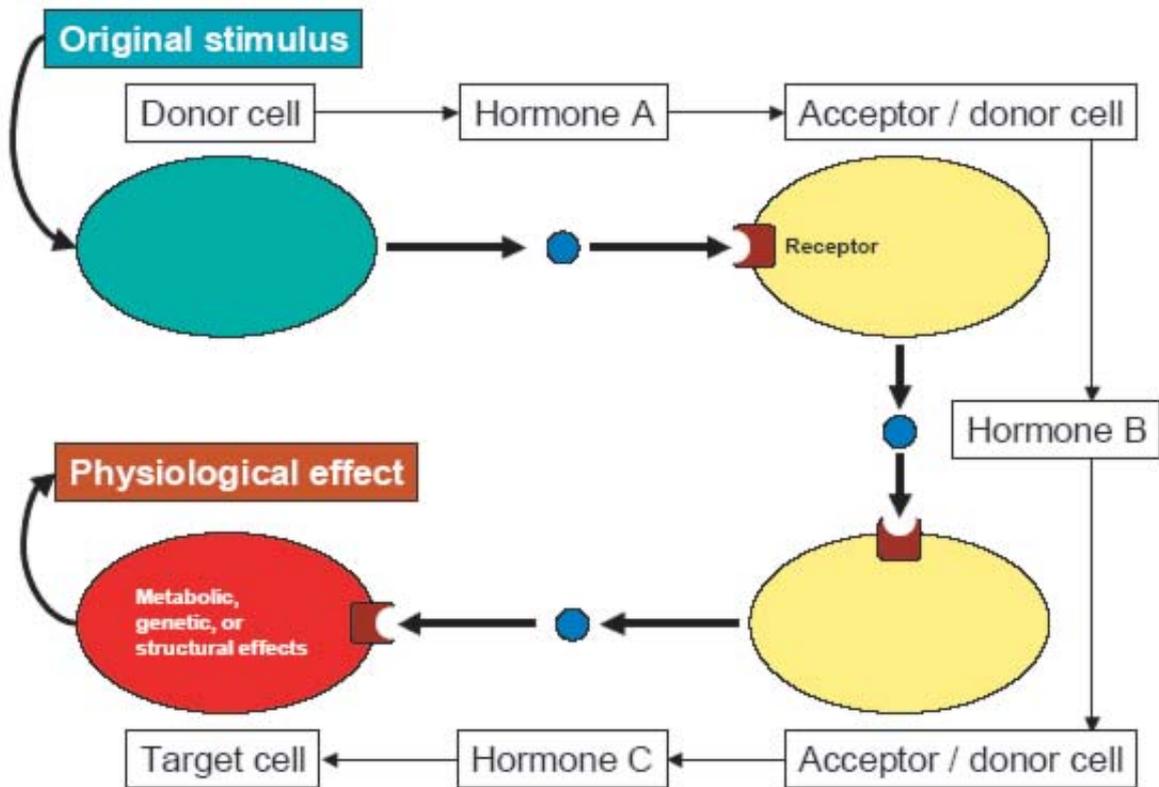
Figure 1. Foundation scheme of a hormonal as example for any intercellular network (Potapov *et al* 2006). An original stimulus triggers the primary donor cell to secrete a hormone A which stimulates an acceptor cell to release hormone B, thereby turning into a donor cell itself. At the end of the cascade, final acceptor or target cells undergo some metabolic, genetic or structural changes to exert the required physiological effect.

and its proper responses to external stimuli (see, for instance, Nussey and Whitehead 2001, for a general introduction). Any trigger that requires concerted action of more than one cell type, be it a physical, chemical or biological one, is then forwarded through a cascade of signalling events involving different cells. These cascades may be organized in a way to diversify and amplify the signal or to constitute parallel signal transmission channels which may render the whole system robust against errors. Starting with the original cell (type) which acted as sensor of the original trigger, a series of molecular communication steps constitutes a complex pathway which we can imagine as a bipartite directed graph consisting of molecular messengers and the involved cells as node classes (figure 1). The entry and the internal cell nodes of this pathway (or network) act as donors emitting (secreting) the molecules (hormones, cytokines, growth factors, or other messengers) required for further processing. The internal cell nodes initially act as acceptors by receiving a signal, before turning into donors themselves. The exit points of each pathway are final acceptor or target cells in which defined processes are initiated which give rise to the specific physiological response to the initial signal.

There are different kinds of messenger molecules, genome-encoded as well as non-genome encoded ones. The first group, peptide and protein hormones as well as all cytokines, growth factors and many others, are encoded by individual or multiple genes, in case they are multi-subunit proteins. Interestingly, by far most of them require more or less extensive posttranslational processing, at least by leader peptide cleavage, or, in many cases, by clipping of a longer precursor polypeptide into a series of biologically active oligopeptides.

Acceptor cells are primarily defined by expression of the respective receptor molecules. Inside all acceptor cells, specific signal transduction (ST) pathways are triggered by the incoming signal. They may aim at any of the known targets of ST pathways, i.e. at: (i) transcription factors (TFs) in order to change the genetic program of this cell; (ii) metabolic enzymes to adapt the metabolism of the cell; (iii) the secretion machinery in order to control the release of the next wave messenger molecules; (iv) structural components of the cell to change the overall cell morphology.

In those acceptor cells which turn into donor cells (i.e. internal nodes), process (i) is obligatory if the next-wave

messenger molecule is genome-encoded, e.g. a peptide hormone. If the induced messenger is not genome-encoded, e. g. a steroid hormone, either process (i) or (ii) has to be activated, for stimulating the expression of metabolic enzyme genes and, thus, to its *de novo* synthesis, or for activating already existing enzyme molecules, respectively.

Each of these subsequent events is subject to a different kind of network, with specific distinguishing features, some of which shall be described in the following.

1.2 *Signal transduction pathways*

In highly compartmentalized eukaryotic cells, any external signal that is to provoke specific cellular effects has to be transduced to its target compartment. Only very few of them are mediated by molecules that are allowed to enter the cell themselves and target their docking molecules, and even these (some small hydrophobic molecules) may require some cellular carriers for being transported to, e.g. the nucleus where they may bind a nuclear receptor (see, e.g. Siegenthaler 1996; Li and Norris 1996; Nettles and Greene 2005).

The majority of extracellular ligands, in particular polypeptides and proteins, but also some smaller molecules,

do not enter a target cell but require a membrane receptor to bind to and initiate a more or less complex signal transduction cascade. These cascades may require a series of (reversible) binding and *de facto* irreversible reactions, usually enzymatic modifications of target molecules. Generally, each reaction generates the 'active' component of the following step, usually the active catalyst of the subsequent reaction (figure 2A). This way, the signal is transmitted ('transduced') through a number of steps, each having the potential to amplify the signal in an enzymatically catalyzed reaction, to diversify the signal due to low substrate specificity of the catalyst, to branch off the pathway by leading to more than one active products, to enable other pathways to cross-regulate ("cross-talk" to) it, and to specify the transmitted signal by requiring additional ones to come in and combine with it.

ST pathways and networks are well represented by directed graphs. As shown in figure 2A and described previously, we have proposed to represent them as bipartite directed graphs, with molecules and reactions as the two node classes (Schacherer *et al* 2001). It is important to differentiate between ST and protein-protein interaction networks, the latter being undirected and comprising any kind of interaction, independent of the functional purpose.
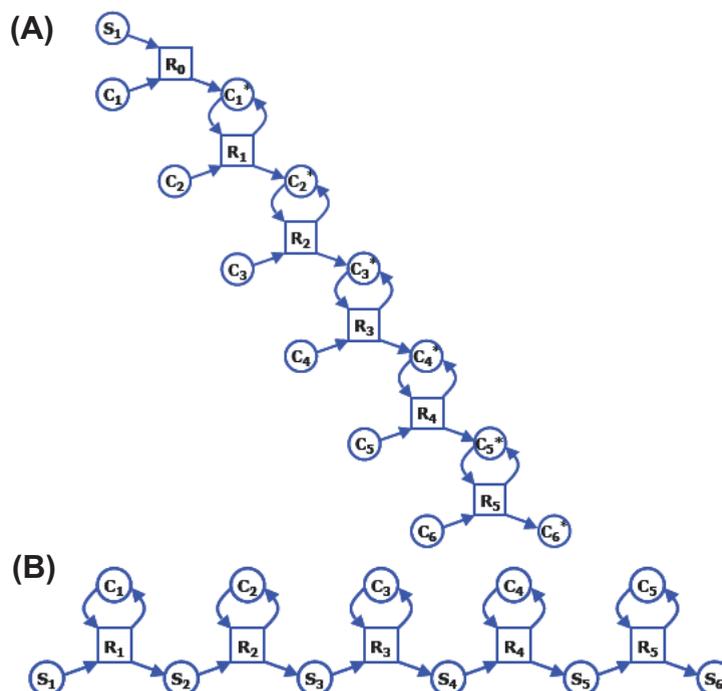

Figure 2. Typical generalized topology of a signal transduction path **(A)** and a metabolic path **(B)**. In the former, an incoming signal (S_1) binds to a receptor, thereby generating the 'active' component (usually, the catalyst) of the next step, and so on. In contrast, in a metabolic path **(B)**, an incoming substrate S_1 is converted by a number of largely pre-existing catalysts into a final product S_6. While in the signalling path, the initiating molecule S_1 and the resulting molecule C_6^* in general do not have anything (material) in common, S_1 and S_6 in the metabolic path share at least some of their atoms.

Moreover, ST pathways have to comprise also non-proteinaceaous components such as second messengers like inositol-1,4,5-trisphosphate (IP3), diacylgylcerol (DAG) or calcium ions.

1.3 *Metabolic networks*

Like ST pathways, metabolic pathways or networks consist of sequences of catalyzed reactions. However, unlike ST pathways, the active catalysts are usually preexisting and hand over the product of the previous step towards the next one. The final product has at least some atoms in common with the input substrate (figure 2B) which also clearly differentiates metabolic from ST pathways. Thus, while metabolic pathways represent generally the transformation of matter, ST pathways transfer signals, i. e. information.

Catabolic as well as anabolic pathways comprise the breakdown of complex and the synthesis of more complex cellular substances. However, in today's general language, metabolic pathways are most commonly understood as the network of those reactions that convert small molecules into each other.

Different representations of these metabolic (or 'biochemical') pathways may exhibit the complete chemical reaction equations, or just the sequence of most important biochemical substances (i.e. usually omitting ubiquitous molecules like CO_2, H_2O, or ATP). The catalysts are usually assigned to the reaction equations, frequently as enzyme code numbers according to the international enzyme catalogue (EC numbers) (as done, for instance, by the typical KEGG maps; Kanehisa *et al* 2006).

An alternative would be a more enzyme-centric view which displays the series of enzymes as they appear in a metabolic path, connected by product-educt relations of the corresponding reactions. The resulting pathways can be transformed into a genetic network if the enzymes are replaced by the genes which encode them. This view can also be obtained with the aid of KEGG or with the enzyme database BRENDA (Schomburg *et al* 2004).

1.4 *Gene regulation networks*

Genetic networks represent abstract functional relations between genes. Nothing is said about the kind of functional interaction between the genes linked in such a network. This is different in a gene regulatory, or transcriptional, network. In such a transcription network (*sensu strictu*), the nodes represent genes encoding transcription factors (TFs), and the edges stand for the transcriptional regulation of these genes by the encoded transcription factors.

Unfortunately, the available experimental information about verified transcription factor binding sites (TFBS)

or other data about TF-target gene relations is still scarce when we consider mammalian systems. A rough estimate suggests that we have knowledge of about only 1% of all TFBS that we may expect to exist in the human genome: Assuming a number of ~30,000 genes in the human genome with just 6 TFBS per gene (unpublished conservative rough estimate from TRANSFAC annotations of promoters that have been studied in-depth), we would expect a lower limit of 240,000 TFBS in the whole genome. This is an extremely conservative assumption since it leaves aside the existence of multiple promoters and enhancers for many genes, and the variability of TFBS arrays under different spatio-temporal cellular conditions. The TRANSFAC database (see below for details) documents about 2500 human TFBS, or about 1% of what we could expect.

The picture does not change significantly if we include more recent high-throughput data obtained by chromatin immuno-precipitation of bound genome regions and subsequent microarray hybridization (ChIP-chip; Ren *et al* 2000). These data give no precise information about exact location and sequence of the TFBS, but just indicate a region of several hundred base pairs in which a TFBS is located. Moreover, in many cases the factor under study seems not to bind directly to this region but rather in an indirect manner, mediated through protein-protein interactions with another TF – a mechanism which provides one possible explanation for the lack of a clear consensus match in many of the isolated sequences. Nevertheless, these studies provide useful additional TF-target relations, though not to an extent that would significantly change the numbers estimated above.

Thus, for a comprehensive transcriptional network analysis, we depend on reliable methods for predicting TFBS and, thus, TF-target gene relations. We will refer to these methods in greater detail further.

2. Corresponding database sources

Careful analysis of the intrinsic features of the individual networks described above led to the development of a number of databases which will be described in the following

2.1 *EndoNet: An information resource about endocrine networks*

We have recently described a new database on endocrine networks (EndoNet) which provides information about hormone and hormone receptor molecules (Potapov *et al* 2006). For the previous, the location (organ / tissue / cell type) is documented by making use of the Cytomer ontology (see below) for defining hormone donor cells. For the latter,
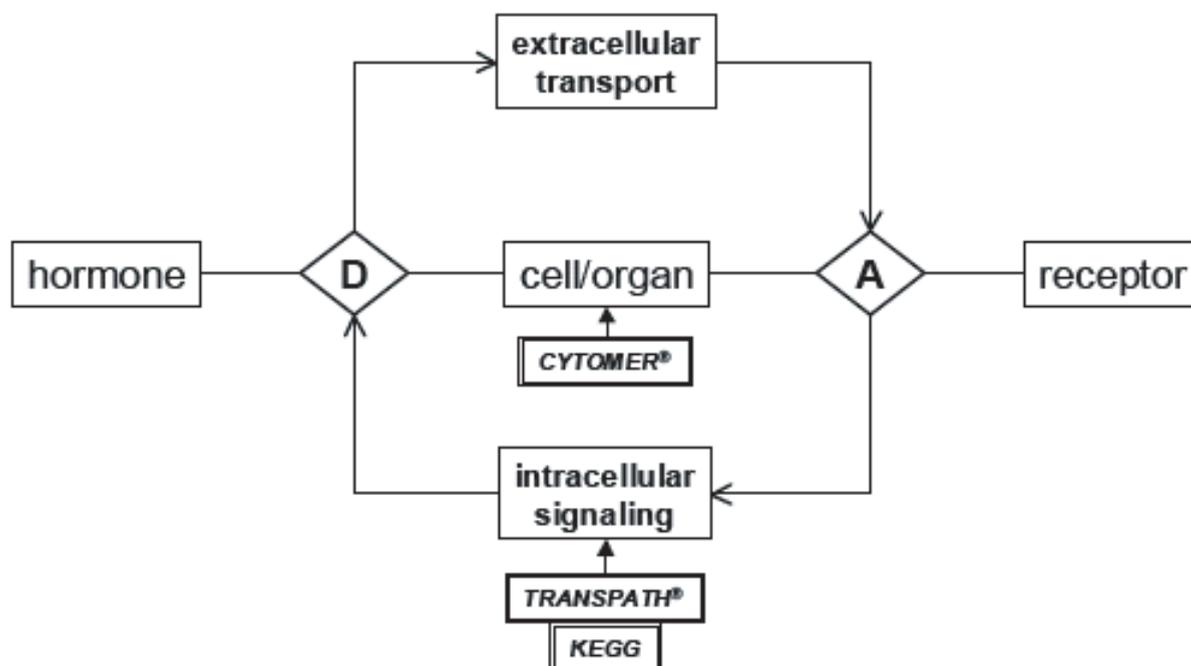
Figure 3. Conceptual schema of the EndoNet database (Potapov *et al* 2006). Donor cells D act on acceptor cells A through molecules (here: hormones) that are transported via a certain medium, usually blood. Both cells are assigned to the Cytomer hierarchy of tissues and cells. In intracellular pathways through which an acceptor cell may be turned into a donor cell for another kind of hormone is modelled in other databases, e.g. TRANSPATH or KEGG.

the tissue in which they are expressed is given as well, since this defines hormonal acceptor cells. The conceptual scheme of the database is shown in figure 3.

The major aim of providing information about molecules (hormones and their receptors) is not to duplicate the work done by, e.g. Swiss-Prot or HumanPSD, though both these databases are extensively cross-linked by EndoNet entries (see below for further information about HumanPSD). Instead, EndoNet is devoted to the task describing the molecular entities which are the active components in the hormonal network, rather than to refer to any particular gene product(s). In other words, for oligo- or polypeptide hormones that have been produced from larger precursors by extensive processing, those mature peptides are considered that have been proven to act as extracellular signal molecules, here: hormones. Similarly, for those hormones and receptors that consist of several subunits, the corresponding EndoNet entry refers to the whole active complex, but explains its subunit structure and links to the corresponding protein and gene entries in other databases. The present content of EndoNet is summarized in table 1. Though presently confined to hormonal networks, the basic concept of EndoNet allows to model any kind of molecular cell-cell-interaction and will be expanded correspondingly in future.

The cellular origin of hormones and location of receptors is described by referring to Cytomer. This is a relational database on human and mouse anatomical structures, tissues, cell types, physiological systems and developmental stages from which an OWL-based ontology has recently been derived (Michael *et al* 2004).

Although EndoNet's view of the participating cells is allowing them to play the role of both acceptor and donor simultaneously, the internal processes which interlink a cell's input and output are not part of its contents. Rather, the signal transduction cascades that forward the hormonal signal to the proper intracellular targets are part of our signal transduction database, TRANSPATH. As for relevant metabolic pathways, we plan to give references to the KEGG database in near future.

2.2 *TRANSPATH: An information resource for signalling pathways and their pathological aberrations*

In TRANSPATH, signalling pathways and networks are organized as bipartite directed graphs (Schacherer *et al* 2001). Both classes of nodes, molecules as well as reactions, are subject to complex hierarchies (Choi *et al* 2004; Krull

Table 1.   Summary of database contents (status June 2006).

Entity	Number of entries	Number of human entries
EndoNet		
Hormone molecules	369	369
Receptor molecules	548	548
Tissues/cells	224	224
TRANSPATH		
Molecules	54,340	
Reactions	95,150	
TRANSFAC		
Factors	8406	1660
Sites	17,905	2500
Factor-site interactions	22,447	
Genes	14,582	10,811
Matrices	811	
TRANSCompel		
Composite elements	421	
Genes	274	123
Interactions	1898	
TRANSPro		
Promoter sequences	155,989	55,633
PathoDB		
Mutated factors	12,178	12,168
Mutated sites	75	75
HumanPSD		
Proteins	50,692	19,456
GO assignments	305,578	116,782

et al 2006). Polypeptide molecules are given as individual splice variants, at the level of finest granularity, summarized by all molecules encoded by one gene (called 'isogroup'), forming 'families' at a number of levels of increasing abstraction. These molecular hierarchies are implemented for each species represented in the database (mainly human, mouse and rat), as well as for a species-independent representation of ortholog groups. These hierarchies are orthogonal to a systematic and, again, hierarchical representation of posttranslational modifications (PTMs) of polypeptide entries on each level. Besides, non-genome encoded molecules to which species-assignment obviously does not apply (e.g. second messengers), are also included in the database.

Reactions are given on three principally distinct levels, called "evidence level", "pathway level" and "semantic projection" (figure 4). The first two provide individual pathway steps as chemical reaction equations. The difference is that the "evidence level" exactly refers to what has been shown in the cited publication, including species-specificity of the participating molecules (in many cases of heterogeneous origin) and some redundancies reflecting different aspects of the same pathway step such as physical interaction of an enzyme with its substrate and the subsequent modification reaction, if both have been evidenced separately. These findings have been summarized and abstracted from the species origin of the individual molecules when proceeding to the pathway level. However, they are still given in a 'mechanistic' view, i.e. as chemical reaction equations. This changes when the corresponding reactions are projected to the 'semantic' level. Here, we focus on the important key components that actively process the signal, thereby omitting abundant compounds such as H_2O or CO_2 as well as the exact PTM and complexing status of the key components.

Presently, TRANSPATH has 54,340 molecule and 95,150 reaction entries. Many of the 'interaction' and 'binding' reactions have been incorporated from HumanPSD (Human Proteome Survey Database).

The Proteome databases are a collection of databases which represent the complete Proteomes of a number of organisms. The first was the Yeast Proteome Database (YPD) for *Saccharomyces cerevisiae* with comprehensive protein reports for all known yeast proteins (Costanzo *et al* 2001). It was complemented later on by PombePD (*Schizosaccharomyces pombe* proteins), WormPD (on *C. elegans* proteins), and MycoPathPD (pathological fungi such as *Candida albicans*). Again later on, HumanPSD and GPCR PD were added for covering the whole space of human, mouse and rat proteins and, with some special features, on their G-protein coupled receptors (Hodges *et al* 2002). One of the most prominent features of HumanPSD is its extensive annotation of expression patterns and disease association of these proteins. We come back to this when describing some of our recent tool developments below. Detailed annotation of protein functions in HumanPSD resulted in a significant contribution to GO development (Harris *et al* 2004) and has enabled improved analysis of differentially regulated genes revealed by microarray experiments (Johnson *et al* 2005). Moreover, protein-protein interactions (not necessarily physical ones) are extensively documented in HumanPSD and linked to an interactive visualizer.

At the bottom end of most signalling cascades are modifications of transcription factors changing their activities to activate or repress transcription of specific sets of genes. These TF-target gene relations have been mostly incorporated from the TRANSFAC database, but detailed information about the TF binding sites involved were omitted.
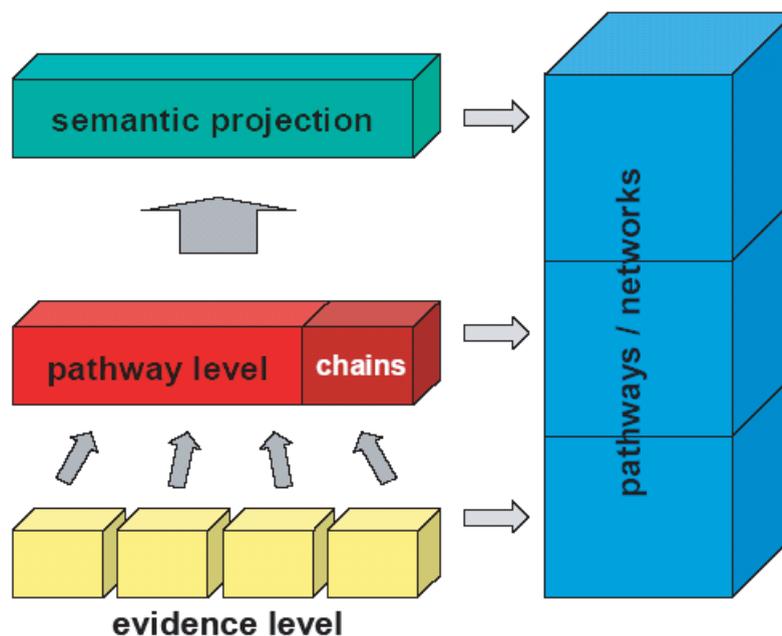
Figure 4. Three-level model for signalling reactions in the TRANSPATH database. The "evidence level" represents as many experimental details obtained from the original publications as possible, while the "pathway level" summarizes this information and 'chains' reactions that have been proven to sequentially occur in a certain cellular environment. Reactions on either level are mechanistically represented, as biochemical reaction equations, whereas the "semantic projection" focuses on those signalling components that are actively involved in transmitting the signal. Expanded pathways and networks can be computationally reconstructed on either level.

2.3 *TRANSFAC: An information resource about transcriptional regulation*

Since its very beginning as a mere tabulated collection of published data in 1988 (Wingender 1988), TRANSFAC has been a compilation of eukaryotic TFs, their genomic binding sites (TFBS) and DNA-binding profiles. In addition, a classification scheme of transcription factors according to their DNA-binding domains was developed (Wingender 1997) and refined over the years (Stegmaier *et al* 2004). More recently, a new 'table' comprising ChIP-chip data has been added (Matys *et al* 2006). It provides presently information about 3841 DNA-fragments interacting with, so far, 8 different TFs.

Since TRANSFAC and its structure became a *de facto* standard in the field, it offers itself as a platform to host also contents of third-party data if they conform to the quality standards of TRANSFAC. Thus, information about 1440 *Arabidopsis thaliana* transcription factors have been taken over from the DATF database (Guo *et al* 2005). These data have been generated by an *Arabidopsis* TF proteome project at Peking University (Gong *et al* 2004). Similarly, 899 genomic site entries have been imported from the *Drosophila* DNase I footprint database (Bergman *et al* 2005).

There are a couple of smaller, specialized databases in the periphery of TRANSFAC which provide data on particular aspects of gene regulation. Among them is the TRANSCompel database on composite elements (CEs), i.e. combinations of single TFBS which together provide particular function on the regulation of the associated gene. The usually synergistically, sometimes antagonistically acting TFBS that constitute CEs have generally been very carefully characterized with regard to their function. The exact experimental evidence for each CE is documented in the database.

In addition, the database module PathoDB provides information about pathologically relevant mutations that have been found in TF-encoding genes or in TFBSs, together with proper genotype and phenotype information (Matys *et al* 2006). We are presently working on an integrated view of these data together with those that similarly describe mutations in genes of signalling components (PathoSign) (Krull *et al* 2006) and with those of the Human Gene Mutation Database, created and maintained by Cardiff University (Stenson *et al* 2003).

The fourth component of the TRANSFAC suite of databases is the promoter database TRANSPro which collects sequences around documented transcription start sites (TSSs), making use of Eukaryotic Promoter Database (EPD) (Schmid *et al* 2006), DBTSS (Suzuki *et al* 2004; Yamashita *et al* 2006) and Ensembl (Hubbard *et al* 2005) annotation of TSSs (Chen *et al* 2006). The latest addition

also considers TSSs reported by the Fantom consortium (Maeda *et al* 2006). The user may define the range of interest for his/her research within a –10,000 and +1000 nucleotides around these TSSs of many human (55,633 promoters), mouse (69,446), rat (3595) and *Arabidopsis thaliana* (27,315) genes. TRANSPro has recently been combined with information to tissue-specificities of the expression of the genes which they control. This novel tissue-specific promoter database (TiProD) allows to select sets of promoters according to their tissue-specificity, but also with regard to any gene ontology (GO) category (Chen *et al* 2006).

Finally, S/MARt DB is a database on scaffold / matrix attached regions of eukaryotic genomes which will not be in the further focus of this review (Liebich *et al* 2002).

3. Tools for predictions

One of the goals of bioinformatics is certainly to contribute to reveal the rules behind the huge amount of facts which we have collected in our databases. These rules qualify an exact science if they prove to have predictive power, i. e. to predict features and/or behaviour of a certain system. If so, it is of general interest to have them implemented in software, preferably through a Web interface, so that the community can apply it for further scientific progress.

3.1 *Analysis of transcription regulatory regions*

On the different levels of biological processes described here, those defined on the lowest complexity level (the genomic DNA sequence) can be most reliably predicted so far, though this is still far from perfection. For instance, many approaches for identifying individual transcription factor binding sites have been developed over the last 15 years, which to review is not the appropriate place here. One of them has been developed by ourselves and implemented into the program Match which comes along with the TRANSFAC database (Kel *et al* 2003). It uses the positional frequency matrix collection of the database and works with different sets of thresholds that have been optimized individually for each matrix and either minimize the number of false positive or of false negative predictions, or the error sum of both of them. The user can override and re-define these default values and compose own profiles, comprising only those matrices that are of the user's special interest and their individually defined thresholds. In addition, a number of pre-defined profiles for, e.g. tissue-specific TF sets and their matrices are available as well.

Compared with the outdated approach of simple string matching using known binding sequences for the identification of potential TFBS, the matrix approach introduces much more flexibility and generally provides higher sensitivity at the same specificity. However, its general assumption is that the individual positions are independent from each other which may be true in most, but certainly not in all cases. To accommodate these concerns, variable order Markov chains and Bayesian networks have been developed and successfully applied to prokaryotic sigma-70 binding sites in *Escherichia coli* (Ben-Gal *et al* 2005); a corresponding Web tool is available (*http://pdw-24. ipk-gatersleben.de:8080/VOMBAT/*). Our own approach was to combine matrix with string matching, making use of the variability a positional frequency matrix provides as well as of the knowledge about patterns that are working in a real genomic environment (P-Match) (Chekmenev *et al* 2005). The approach has been proven advantageous for predicting binding sites of many TF, though certainly not all, in particular in the range of high sensitivity (low rate of false negative hits), which is usually accompanied by a drastic increase in the false positive matches (decrease of specificity) and which could be largely avoided by the P-Match algorithm (Chekmenev *et al* 2005).

Another method to increase the reliability of single site prediction which gains increasing popularity is phylogenetic footprinting (Tagle *et al* 1988). In most cases, this is done by primate-rodent (e.g., human-mouse) comparison and it has recently been reconfirmed that the approach has its validity, but on average leads to a false-negative rate of about one third (Sauer *et al* 2006). Interestingly, the binding sites of different transcription factors largely differ in the degree of their sequence-conservativity, indicating that this method should be applied with caution.

Nevertheless, all the existing methods for the prediction of individual TFBS have severe limitations. As an emerging picture, it seems that the required specificity for fine-tuned transcriptional regulation in a higher eukaryotic cell is achieved by the context of a number of TFBS, combining to (usually binary) composite elements and more complex promoter modules. To identify a specific combination of (predicted) TFBS that may be functionally relevant in a given set of promoters means, we have to search for specific TFBS types, their scoring, relative (to each other) and absolute (to the respective TSS) distance correlation and mutual orientation. That requires optimization within a large search space which we recently have suggested to do by using a genetic algorithm which has been implemented in a tool (CMA, composite module analyzer) (Kel *et al* 2006a).

3.2 *Pathway reconstruction*

As described above (§2.2), the TRANSPATH database harbours a huge number of signal transduction reactions on different abstraction levels. Many of them have been pre-compiled and are labelled to belong to certain 'canonical'

pathways, such EGF or HIF-1alpha pathway. However, the individual reactions can be combined to a much larger number of (potential) pathways and networks when we link those that produce a certain state as result with those that start from the same state. This is achieved by the tool PathwayBuilder, which is part of the TRANSPATH database. It can easily end up with gigantic maps of predicted networks which are hard to visualize, and impossible to think out beforehand.

Moreover, many of these pathways may be "false positively" predicted ones, for instance due to the fact that some reactions may have been shown under pure *in vitro* conditions and will hardly occur in a cellular environment. To cope with this, we have developed an algorithm and implemented it in PathwayBuilder which during pathway reconstruction gives priority to those sequences of reactions which have been experimentally shown to follow each other in a certain cellular environment. The user can assign a special preference to these 'chains' of reactions by adapting the cost function of the length of the shortest paths in the reconstructed network (Kel *et al* 2006b).

The Pathway Builder has been re-used also for visualizing transcriptional networks as they can be retrieved from the TRANSFAC database (§2.3).

3.3 *Analysis of expression data*

Attempts to obtain insights into a certain biological process frequently make use of new high-throughput technologies such as microarray-based comparison of the transcriptomes of different cells. Usually, the resulting sets of affected genes, either up- or down-regulated, and their respective products are mapped onto known pathways or networks, e.g. metabolic networks as they are represented in the KEGG database, or signalling networks from TRANSPATH. The information obtained from this kind of approach may give answers to the question, which functions of the cell are affected by the conditions under consideration ("downstream analysis"; figure 5).

However, when considering the causes that have led to the observed changes and thus trying to generate an explanatory model, we have to analyse the regulatory
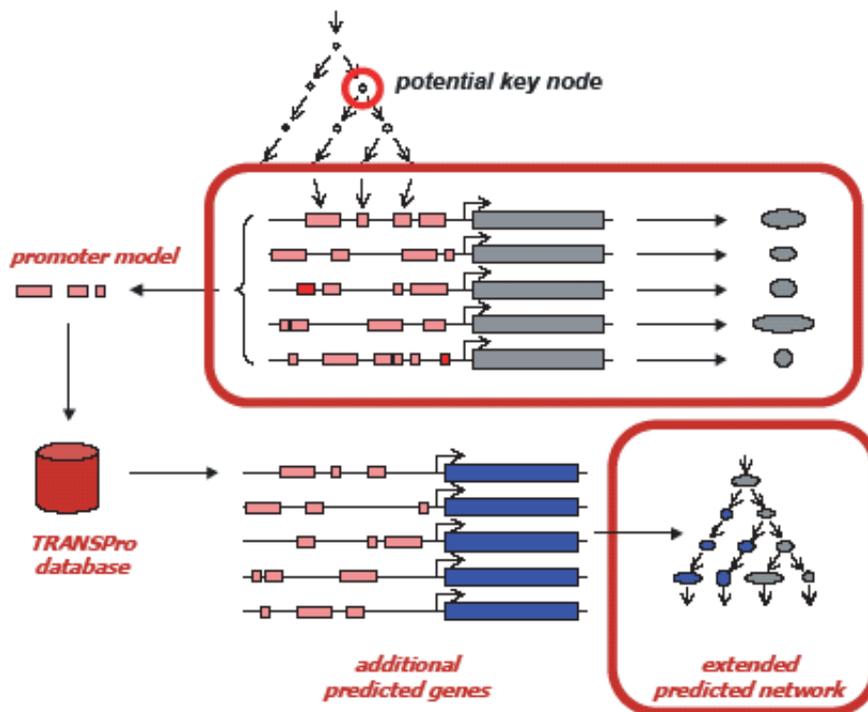

Figure 5. General strategy of gene expression analysis by the ExPlain system. A set of induced genes which may have been obtained from a microarray experiment (central box), is translated into its products and mapped onto, e.g., signal transduction pathways as they are represented in the TRANSPATH database (grey ovals). Independently, the system retrieves the promoter sequences, analyses them for potential transcription factor binding sites (TFBSs, red rectangles), constructs promoter models by comparing these promoters with those of a control set, and predicts the relevant transcription factors. Pathway analysis may reveal potential key nodes that controlling the set of genes under study (top part). In addition, a promoter database such as TRANSPro can be screened with the obtained promoter model for additional genes that belong to this set of co-regulated genes (blue rectangles). Their products (blue ovals) can be mapped on the available pathways as well.

mechanisms of the up- and down-regulated genes. Usually, the available experimental information for this is scarce so that we depend on reliable predictive methods to analyse the corresponding promoters, and to come up with "promoter models", i. e. sets of TFBS with a certain scoring, order, orientation and distance correlation. Once a promoter model has been established, we can hypothesize the transcription factors involved, and can derive the signalling pathways upstream of them controlling their activities. If we are lucky, we may find a node in the network where the different paths which we follow upstream converge, and which therefore may be candidates for mastering the observed set of genes and, thus, the process under consideration (figure 5). Note that the gene of such a key node itself is not necessarily upregulated; sometimes, it is even not a genome-encoded node in the signalling network.

With a promoter model at hand, we may also go back to the promoter database and identify additional genes that might belong to the same cluster of co-regulated genes, although they have not been observed in the experiment. Their induction/repression may have been masked for whatever reason from observation or analysis, or the appropriate probe was missing from the array used.

We have now combined most of our previously developed databases and tools to a comprehensive platform for analysing gene expression data in both down- and upstream direction (ExPlain). Raw data about induction/repression of gene expression rates can be uploaded as text or Excel spread sheets, manually divided into positive and negative control sets, and analysed for functional clusters in terms of any GO category, expression profile, disease function (according to HumanPSD) or signalling pathway (from TRANSPATH). Any such defined cluster of genes, or the whole set, may be subjected to an upstream analysis as outlined above. For this, automated retrieval of all relevant promoters from TRANSPro is followed by analysis for single TFBS using TRANSFAC matrices and Match with any pre- or user-defined profile. The sets of predicted TFBS can be analysed further with the CMA, and the potential pathways acting on the suggested sets of transcription factors and potential key nodes are identified with PathwayBuilder and ArrayAnalyzer on the basis of TRANSPATH contents. Altogether, the ExPlain tool provides an interactive workflow implementing the scheme shown in figure 5.

4. Applications

Proof-of-principle has been given for most of the concepts outlined above. For instance, the power of context-sensitive analysis of transcription regulatory regions was proven for E2F sites, which play an important role in cell-cycle regulation, as well as for composite elements comprising NFATp and AP-1 elements which are involved in T-cell

activation-associated gene induction events (Kel *et al* 1999, 2001). The same approach plus promoter module construction led to challenging hypotheses about the involvement of certain TFs in development of aggressive behaviour in mice (D'Souza *et al* 2003), and was successfully proven in the case of aryl hydrocarbon receptor target genes (Kel *et al* 2004). A more stringent knowledge-based approach was used to model the promoter structure of genes that are involved in antibacterial cell response (Shelest and Wingender 2005). Linking promoter modelling with comprehensive pathway analysis led to interesting hypotheses in connection with certain disease studies, some of them have already been verified experimentally and will be published elsewhere (Kel *et al* unpublished results).

Systematic application of network topology analysis has been demonstrated to lead to useful results. Comparative analyses of the whole signalling network represented in the TRANSPATH database ("Reference network) as well as the subnetworks of insulin signalling and the p53 network all revealed clear scale-free and modular properties. When confining this analysis to the transcriptional network, extracted from the databases TRANSFAC and TRANSPATH, where the node represent TF genes and the edges transcriptional regulation by the encoded TFs, these features were confirmed. Most interesting, however, proved a topology parameter which has been rarely investigated in the past: betweenness centrality (BC). A node's BC is defined as the fraction of those shortest paths between all pairs of vertices that pass through this node. Nodes with high BC value do not necessarily have a high connectivity (degree) by themselves. Nevertheless, they are of key importance for the whole network. When we computed this parameter for all TF genes in the transcriptional network and ranked the factors according to their BC values, the top-most five entities were all either tumour suppressor or proto-oncogenes (p53, c-fos, Egr1, c-jun, WT1) (Potapov *et al* 2005). Unfortunately, the overall network size is relatively small (121 nodes, 212 edges) due to the limited number of experimentally proven TFBS in TF gene promoters, for which reason this observation can be taken only as preliminary result. Presently, we make attempts to enrich the network by TF-target gene relations through potential TFBS predicted by a combined matrix-phylogenetic footprinting approach.

5. Discussion

The different kinds of networks shortly depicted above and represented in the databases described in this contribution reflect some of the main organizational levels of biological systems. Although systems biology has usually modelled and investigated the behaviour of cellular systems so far, the advent of increasingly efficient high-throughput

methodologies will enable us to step beyond this borderline. We have to achieve a "vertically integrative" systems biology where we can smoothly zoom between the different complexity levels, from the systemic all-organism layer down to the nuclear (transcriptional) events in individual cells that play a key role in a certain physiological process. Systems biology also includes the development of predictive models for the behaviour of the system under investigation. The complexity of the higher-level (e.g. physiological) systems renders it extremely unlikely that it will be possible to exactly describe all molecular details involved in a certain physiological process by ordinary differential equations (ODEs). Rather, we have to identify the essential steps, those that are subject to thorough molecular or genetic regulation, model these steps with the required precision, and keep the remainder of the network at a much lower granularity. Thus, identifying the appropriate granularity for each system model will be one of the major tasks when proceeding with systems biology approaches to systemic descriptions. A second one will be to provide the appropriate biological semantics that have to be connected to these mathematical models. To do this in a formally strict way, corresponding ontologies have to be developed and greed upon by the community. It is our very hope that the resources described above can contribute to both these important tasks.

6. Availability

Many of the databases and tools described in this article are freely accessible for users from non-profit organizations (TRANSPATH Public, TRANSFAC Public, TRANSCompel Public, PathoDB, Match, P-Match, CMA: *http://www. gene-regulation.de*; Cytomer, EndoNet, PathoSign: *http:// www.bioinf.med.uni-goettingen.de/services/*; TiProD: *http:// tiprod.cbi.pku.edu.cn:8080/index.html*). Cytomer, PathoDB, TRANSFAC, TRANSCompel, TRANSPATH, are registered trademarks, HumanPSD, GPCR PD, PombePD, TRANSPro, WormPD, YPD, ExPlain, Match, and P-Match are trademarks of BIOBASE GmbH, Wolfenbüttel, Germany. HGMD is a registered trademark of Cardiff University, Cardiff, Wales, UK.

References

Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S and Grosse I 2005 Identification of transcription factor binding sites with variable-order Bayesian networks; *Bioinformatics* **21** 2657–2666

Bergman C M, Carlson J W and Celniker S E 2005 *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*; *Bioinformatics* **21** 1747–1749

Chekmenev D S, Haid C and Kel A E 2005 P-Match: transcription factor binding site search by combining patterns and weight matrices; *Nucleic Acids Res.* **33** W432–W437

Chen X, Wu J M, Hornischer K, Kel A and Wingender E 2006 TiProD: The Tissue-specific Promoter Database; *Nucleic Acids Res.* **34** D104–D107

Choi C, Crass T, Kel A, Kel-Margoulis O, Krull M, Pistor S, Potapov A, Voss N and Wingender E 2004 Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database; *Genome Inf. Ser.* **15** 244–254

Costanzo M C, Crawford M E, Hirschman J E, Kranz J E, Olsen P, Robertson L S, Skrzypek M S, Braun B R, Hopkins K L, Kondu P, Lengieza C, Lew-Smith J E, Tillberg M and Garrels J I 2001 YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information; *Nucleic Acids Res.* **29** 75–79

D'Souza U M, Kel A and Sluyter F 2003 From transcriptional regulation to aggressive behavior; *Behav. Genet.* **33** 549–562

Gong W *et al* 2004 Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes; *Plant Physiol.* **135** 773–782

Guo A, He K, Liu D, Bai S, Gu X, Wei L and Luo J 2005 DATF: a database of *Arabidopsis* transcription factors; *Bioinformatics* **21** 2568–2569

Harris M A *et al* 2004 The Gene Ontology (GO) database and informatics resource; *Nucleic Acids Res.* **32** D258–D261

Hodges P E, Carrico P M, Hogan J D, O'Neill K E, Owen J J, Mangan M, Davis B P, Brooks J E and Garrels J I 2002 Annotating the human proteome: the Human Proteome Survey Database (HumanPSD™) and an in-depth target database for G protein-coupled receptors (GPCR-PD™) from Incyte Genomics; *Nucleic Acids Res.* **30** 137–141

Hubbard T *et al* 2005 Ensembl 2005; *Nucleic Acids Res.* **33** D447–D453

Johnson R J, Williams J M, Schreiber B M, Elfe C D, Lennon-Hopkins K L, Skrzypek M S and White R D 2005 Analysis of gene ontology features in microarray data using the Proteome BioKnowledge Library; *In Silico Biol.* **5** 389–399

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K F, Itoh M, Kawashima S, Katayama T, Araki M and Hirakawa M 2006 From genomics to chemical genomics: new developments in KEGG; *Nucleic Acids Res.* **34** D354–D357

Kel A, Kel-Margoulis O, Babenko V and Wingender E 1999 Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells; *J. Mol. Biol.* **288** 353–376

Kel A, Konovalova T, Waleev T, Cheremushkin E, Kel-Margoulis O and Wingender E 2006a Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations; *Bioinformatics* **22** 1190–1197

Kel A, Voss N, Jauregui R, Kel-Margoulis O and Wingender E 2006b Beyond microarrays: Find key transcription factors controlling signal transduction pathways; *BMC Bioinformatics* (in press)

Kel A, Reymann S, Matys V, Nettesheim P, Wingender E and Borlak J 2004 A novel computational approach for the prediction of networked transcription factors of aryl hydrocarbon-receptor-regulated genes; *Mol. Pharmacol.* **66** 1557–1572

Kel A E, Kel-Margoulis O V, Farnham P J, Wingender E and Zhang M Q 2001 Computer-assisted identification of cell cycle-related genes – new targets for E2F transcription factors; *J. Mol. Biol.* **309** 99–120

Kel A E, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis O V and Wingender E 2003 MATCH: A tool for searching transcription factor binding sites in DNA sequences; *Nucleic Acids Res.* **31** 3576–3579

Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O and Wingender E 2006 TRANSPATH®: an Information resource for storing and visualizing signaling pathways and their pathological aberrations; *Nucleic Acids Res.* **34** D546–D551

Li E and Norris A W 1996 Structure/function of cytoplasmic vitamin A-binding proteins; *Annu. Rev. Nutr.* **16** 205–234

Liebich I, Bode J, Frisch M and Wingender E 2002 S/MARt DB: a database on scaffold / matrix attached regions; *Nucleic Acids Res.* **30** 372–374

Maeda N *et al* 2006 Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs; *PLoS Genet.* **2** e62

Matys V, Kel-Margoulis O V, Fricke E, Liebich I L S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel A E and Wingender E 2006 TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes; *Nucleic Acids Res.* **34** D108–D110

Michael M, Chen X, Fricke E, Haubrock M, Ricanek R and Wingender E 2004 Deriving an ontology for human gene expression sources from the CYTOMER® database on human organs and cell types; *In Silico Biol.* **5** 7

Nettles K W and Greene G L 2005 Ligand control of coregulator recruitment to nuclear receptors; *Annu. Rev. Physiol.* **67** 309–333

Nussey S S and Whitehead S A 2001 *Endocrinology: An Integrated Approach* (Oxford: BIOS Scientific Publishers)

Potapov A P, Voss N, Sasse N and Wingender E 2005 Topology of mammalian transcription networks; *Genome Inf. Ser.* **16** 270–278

Potapov A, Liebich I, Dönitz J, Schwarzer K, Sasse N, Schoeps T, Crass T and Wingender E 2006 EndoNet: an information resource about endocrine networks; *Nucleic Acids Res.* **34** D540–D545

Ren B, Robert F, Wyrick J J, Aparicio O, Jennings E G, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert T L, Wilson C J, Bell S P and Young R A 2000 Genome-wide location and function of DNA binding proteins; *Science* **290** 2306–2309

Sauer T, Shelest E and Wingender E 2006 Evaluating physogenetic footprinting for human-rodent comparisons; *Bioinformatics* **22** 430–437

Schacherer F, Choi C, Götze U, Krull M, Pistor S and Wingender E 2001 The TRANSPATH signal transduction database: a knowledge base on signal transduction networks; *Bioinformatics* **17** 1053–1057

Schmid C D, Perier R, Praz V and Bucher P 2006 EPD in its twentieth year: towards complete promoter coverage of selected model organisms; *Nucleic Acids Res.* **34** D82–D85

Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G and Schomburg D 2004 BRENDA, the enzyme database: updates and major new developments; *Nucleic Acids Res.* **32** D431–D433

Shelest E and Wingender E 2005 Construction of predictive promoter models on the example of antibacterial response of human epithelial cells; *Theor. Biol. Med. Model.* **2** 2

Siegenthaler G 1996 Extra-and intracellular transport of retinoids: a reappraisal; *Horm. Res.* **45** 122–127

Stegmaier P, Kel A E and Wingender E 2004 Systematic DNA-binding domain classification of transcription factors; *Genome Inf. Ser.* **15** 276–286

Stenson P D, Ball E V, Mort M, Phillips A D, Shiel J A, Thomas N S, Abeysinghe S, Krawczak M and Cooper D N 2003 Human Gene Mutation Database (HGMD): 2003 update; *Hum. Mutat.* **21** 577–581

Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Kel A E, Arakawa T, Carninci P, Kawai J, Hayashizaki Y, Takagi T, Nakai K and Sugano S 2004 Large-scale collection and characterization of promoters of human and mouse genes; *In Silico Biol.* **4** 429–444

Tagle D A, Koop B F, Goodman M, Slightom J L, Hess D L and Jones R T 1988 Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints; *J. Mol. Biol.* **203** 439–455

Wingender E 1988 Compilation of transcription regulating proteins; *Nucleic Acids Res.* **16** 1879–1902

Wingender E 1997 Classification scheme of eukaryotic transcription factors; *Mol. Biol. (Mosk)* **31** 584–600

Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K and Sugano S 2006 DBTSS: DataBase of Human Transcription Start Sites, progress report 2006; *Nucleic Acids Res.* **34** D86–D89