Integrating Pathway Data for Systems Pathology

Edgar Wingender\textsuperscript{a,}, Jennifer Hogan\textsuperscript{a}, Frank Schacherer\textsuperscript{a}, Anatolij P. Potapov\textsuperscript{b} and Olga Kel-Margoulis\textsuperscript{a}

\textsuperscript{a}BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany
\textsuperscript{b}Department of Bioinformatics, Medical School, Georg August University of Göttingen, Goldschmidtstr. 1, D-37077 Göttingen, Germany

Edited by I. Rojas and U. Wittig (guest editors); received and accepted 21 March, 2007; published 30 March, 2007

\textbf{ABSTRACT:} The HumanPSD database on the complete proteomes of human, mouse and rat has been integrated with the databases TRANSFAC on gene regulation and TRANSPATH on signal transduction to provide a comprehensive systems biological platform for these organisms. As a next step, integration with PathoDB and PathoSign on pathologically relevant mutations is planned together with an extension beyond the limits of the individual cell, towards intercellular networks, by integrating the database EndoNet on hormonal networks as well. The overall aim is to come up with a platform that is suitable to provide knowledge for systems pathology, i.e. a system-wide modeling of pathological states and their development.

\textbf{INTRODUCTION}

Database integration has been recognized as an extremely important issue for the progress of molecular biological sciences since many years, and the holistic view which modern systems biology is aiming at is putting additional emphasis on this need. The last 20 years have witnessed numerous attempts to integrate public domain databases [e.g., Ritter et al., 1994; Crass et al., 2004; Pruess et al., 2004; Kersey et al., 2005]. In spite of the excellent work done in all these projects, the practical results were mostly limited use. This is due to the instability of the underlying individual databases, both in structure as well as their mere existence which is frequently confined to a certain funding period, as well as to severe semantic incompatibilities since the individual databases were originally designed for quite different purposes. Moreover, the links between them are not stable and need particular care, or are not provided at all, which causes the required integration efforts to geometrically increase with the number of participating databases.

In a relatively small scale, we had to solve related problems when integrating our own data resources for objects (genes and proteins) and mechanisms (signal transduction, transcriptional regulation) with each other as well as with additional databases on pathologically relevant mutations in the involved genes and proteins. Since each of these databases was originally developed to fulfill the specific needs of a certain (in many cases: externally funded) project and sometimes by independent entities, their interoperability

\textsuperscript{*}Corresponding author. E-mail: edgar.wingender@biobase-international.com.

Electronic publication can be found in \textit{In Silico Biol.} 7, S1/03 <http://www.bioinfo.de/issb/2007/07/S1/03/>, 30 March 2007.

1386-6338/07/$17.00 © 2006 – IOS Press and Bioinformation Systems e.V. and the authors. All rights reserved
was not an *a priori* part of their design. Moreover, the ways how their contents are retrieved from original publications differ considerably, which also contributes to the heterogeneity among their structures.

In this contribution, we describe our strategy how the data for the individual databases are retrieved from the original publications, and how their integration has been achieved.

**METHODS**

**Databases**

As starting point to establish an integrated resource for pathologically relevant quantitative data, the following databases were used:

- **Proteome databases:** This is a family of databases aiming at the annotation of complete proteomes of (a) human, mouse and rat (Human Proteome Survey Database; HPSD\textsuperscript{TM}), (b) G-Protein Coupled Receptor Proteome Database (GPCR PD\textsuperscript{TM}); (c) *Saccharomyces cerevisiae* (Yeast Proteome Database, YPD\textsuperscript{TM}), (d) *Schizosaccharomyces pombe* (PombePD\textsuperscript{TM}), (e) pathological fungi (MycoPathPD\textsuperscript{TM}), and (f) *Caenorhabditis elegans* (WormPD\textsuperscript{TM}) [Costanzo *et al.*, 2001; Csank *et al.*, 2002; Hodges *et al.*, 2002; Johnson *et al.*, 2005]. For the integration described in the following, only HPSD and GPCR PD were used so far.

- **TRANSPATH\textsuperscript{®}:** This is a database on regulatory network information, mostly in human, mouse and rat cells, originally focusing on signal transduction pathways that aim at gene regulatory molecules (transcription factors), but increasingly expanding to metabolic enzymes and structural proteins as targets [Schacherer *et al.*, 2001; Choi *et al.*, 2004; Krull *et al.*, 2006]. Data about intracellular regulatory processes are collected, annotated and stored at different levels of abstraction making use of both a molecular and a reaction hierarchy [Wingender *et al.*, 2007].

- **TRANSFAC\textsuperscript{®}:** A database on transcriptional regulation in eukaryotes, with documentation of transcription factor properties, their genomic and artificial binding sequences, and DNA-binding profiles [Matys *et al.*, 2006].

- **PathoDB\textsuperscript{®}:** is a database on pathologically relevant mutated forms of mammalian transcription factors and their binding sites [Wingender *et al.*, 2001].

- **PathoSign** comprises pathologically relevant mutations of signaling components of mammalian cells [Krull *et al.*, 2006].

- **EndoNet** is a database on intercellular signaling paths as they are exerted by, e.g., hormones to constitute the endocrine network of the human body [Potapov *et al.*, 2006]. All data are linked to CYTOMER\textsuperscript{®} which is a relational database on organs/tissues, cell types, physiological systems and developmental stages, presently focusing on the human system. From this database, an ontology for anatomical, morphological structures for the human organism, including all embryonal stages, and the cell types constituting these structures is derived [Michael *et al.*, 2005].

**Database annotation**

The annotation of the above-mentioned databases is done manually throughout, with gradual differences with regard to involving support by text mining tools to reduce tedious activities of the process, while leaving decisions up to the trained expert. Thus, the proteome databases make use of a sophisticated text mining system, which is based on a long-trained statistical model for the recognition of relevant
articles (since 2000), vocabulary flagging of keywords and words from ontologies like MeSH, and on a system for biological entity recognition dubbed "nametagging".

Three data pipelines have been constructed to retrieve and maintain sequences, references and vocabularies. Consequently, annotation of genes and gene products is done by making use of knowledge in public domain databases such as GenBank, Entrez, PubMed, Pfam, Gene Ontology (GO) and MeSH.

The sequence pipeline populates the Proteome databases with recent information about genes from Entrez Gene (GenBank for model organism species), assigning as much as possible of verified or hypothetical information by domain analysis and homology searches. This is tracked by a refined status control system, involving various checkpoints of quality control by human experts.

In our reference pipeline, out of more than 20,000 references appearing every week, our filters for subjects, journals, and species filter out all but approximately 1,300 articles which are added to the internal literature database. Subsequently, the references are subject to nametagging of the mentioned genes and molecules, and are ranked by journal impact and keyword analysis. At the end, references are fed into workflow tracking systems for manual curation of information about GO assignments, expression patterns, mutant phenotypes, diseases, etc.

For this, the mentioned nametagging plays a central role. It organizes related proteins, e.g. human, mouse and rat proteins into groups of orthologs, called "cognate groups", which have to fulfill criteria for sequence similarity and share common gene symbols, and it connects proteins to the correct references. This process consists of two steps: first, the construction of a manually curated database of symbolic names, synonyms, and disambiguating contextual words. A primary name is given for each gene/protein, extended lists of synonyms are attached, both for short symbols and fully spelt out names. Especially important, "anchors", "antianchors" and "masks" are defined, which are words that must or must not appear in an abstract together with one of the identifying names, respectively, or mask proper entity names that contain the name in question as a substring (e.g., "EGF receptor" is masked for EGF recognition). Through ongoing curation, the nametagging search engine is subject to constant refinement and optimization. In a second step, this database is then used to identify biological entity names in abstract texts, and associate them with the abstracts.

Keyword tagging is used to classify and prioritize references as well as for prompting manual curation. This is also continuously refined by analyzing word occurrences in "good" and "bad" references.

The system has for example been adapted to retrieve protein-protein interactions from the abstracts of original literature, by identifying abstracts that mention various gene names along with keywords like "bind" or "interact". The statistical filter is then used to rank papers that are probable to give information about direct interactions in the full text, a task that is not possible by simple identification of keywords like "direct" in the abstract. An analysis of significant words in the model showed that mention of certain experimental techniques is a much better indicator for texts containing such direct evidence.

Compared to that, both TRANSPATH and TRANSFAC have a more specific focus. Thus, the relevant literature here is selected by appropriate PubMed search strategies according to the respective curation focus (i.e., the pathway or the transcription factor family which were selected for updating). A sophisticated data input tool has been developed for making optimal use of a rich set of controlled vocabularies, with a huge number of scripts that allow efficient annotation and combination of data from various sources.

PathoDB and PathoSign are subject to different annotation mechanisms. Since they focus on pathologically relevant mutations of molecules which are usually already documented in either TRANSFAC or TRANSPATH, the contents of the latter provide a kind of "anchor" for the maintenance of these two databases. Since they share the scope of providing information about the genotype, the corresponding
molecular phenotype as well as the resulting macroscopic (clinical) phenotype, they have been merged into one production system.

EndoNet was initially populated with textbook knowledge, to provide a proof of concept for its newly developed structure. A specific annotation client was developed to make proper links between TRANSPATH entries for extracellular ligands (firstly, hormone) and their corresponding receptors on the one side and entities of the CYTOMER database and ontology for (human) organs, tissues and cell types on the other [Michael et al., 2005].

As mentioned above for the individual resources, ontologies are used throughout for unifying annotation, for example GO, MeSH or CYTOMER. In areas where annotation from different sources was to be integrated, ontologies were leveraged wherever possible. However, for integrating annotation of gene structure and gene product (usually: protein) functions, these ontologies are of little use, since none of them helps to clearly differentiate between these objects. This consistently leads to problems in the annotation of splice or protein processing variants with completely different functions. The hierarchical data structure originally implemented for the TRANSPATH database has now been adopted for the integrated knowledge base to resolve this problem.

RESULTS AND DISCUSSION

Integration of Data on Genes, Proteins and Reactions

Because of the distinct history of all the mentioned databases, which caused their optimization for very specific project needs and adoption of distinct concepts, the integration of these resources involved quite some conceptual and developmental efforts. While there were reasons to keep the production databases for curation of the contents still intact and separate, successful integration of the database into one comprehensive relational database system on the user side was achieved. One of the results is that a HPSD entry displaying features of a signaling component gains information about the pathway involvement of this molecule, including its visualization. Entries in HPSD which refer to transcription factors (TFs) are enriched by specific contents provided by the TRANSFAC database, including genomic binding sites of this factor. If the activity of such a TF is regulated in response to a signaling cascade, the corresponding signal transduction pathway represented in TRANSPATH is connected as well.

Consistent merging of the corresponding objects described so far, i.e. of protein molecules and genes, was done by tracing all of them back to the corresponding gene as fundamental unit. In spite of all definitional problems about what a gene finally is, gene symbols are the most commonly accepted denominators. However, when proceeding to gene products, there is no 1:1 relation between a gene and the encoded product(s). In addition, many facts reported in literature cannot be traced back to genes or proteins of a particular biological species, and sometimes, it is even more useful to work on a more abstract level summarizing features of orthologous objects of, e.g., mammalian species. In some cases, even higher abstraction such as summarizing related genes or proteins into families is the only adequate way to represent published knowledge. These concepts were developed earlier for the TRANSPATH database [Choi et al., 2004] and adopted now to the other databases reported here. This kind of molecule hierarchy enabled us to assign features and reactions with the appropriate granularity, as predefined by the original literature or database source.

On the other side, when modeling signal transduction pathways it is also necessary to incorporate non-proteinaceous molecules such as DAG, IP3 or Ca2+ ions, which are part of TRANSPATH, but obviously haven’t been represented in HPSD. Molecules that are not directly genome-encoded are particularly
important for intercellular hormonal networks, e.g. steroids. They form an essential part of the EndoNet database, are an integral part of TRANSPATH as signaling molecules, and will be part of the steroid synthesizing metabolic pathways as soon as TRANSPATH is enriched with this type of data.

Another new class of objects has been introduced recently and is now being incorporated into a consistent and comprehensive model of regulatory networks: MicroRNAs. We have summarized all the available information about miRNAs by making use of public databases such as miRBase [Griffiths-Jones et al., 2006], but mostly going back to the original literature. In addition to the information about the miRNAs themselves, their precursors and mature forms (size, sequence), we were particularly interested in the genes coding for these RNAs and their regulation. Thus, the expression and the target genes of miRNAs constituted new classes of regulatory interactions in the TRANSPATH database. MicroRNAs as regulators, in analogy to transcription factors, and their binding to complementary sequences in 3’-UTRs of their target mRNAs analogously to the TF binding to genomic cis-elements, fitted nicely to the existing structure of the TRANSFAC database where they have been incorporated now as well.

This way, the broad coverage of the complete human, mouse and rat proteomes provided by HPSD is seamlessly integrated with regulatory network information provided by TRANSPATH and TRANSFAC. Among the different conceivable end points of signaling pathways: effects on gene regulation (TFs), metabolism (metabolic enzymes), cellular structure (e.g., cytoskeletal proteins), or secretory events (proteins of the secretory apparatus), the first has been well covered now by the described integration. Presently, strong efforts are made to enrich the TRANSPATH database by information on metabolic pathways. Of particular interest, and not yet systematically covered by any other database that deals with properties of metabolic enzymes, are enzyme modifications that influence their activity in response to signal transduction pathways. Such information will be part of the TRANSPATH database after the next release. Nevertheless, the intrinsic differences between signaling and metabolic pathways require different concepts for their representation, which makes it also necessary to join these concepts to a more comprehensive one [Choi et al., 2004; Wingender et al., 2007]. First attempts have been done to incorporate effects on structural proteins as well, e.g. during spindle formation. Regulation of secretion will be of particular importance when we proceed with the integration of EndoNet contents, since secretion of many hormones is subject to tight control. The final core schema of the integrated database is shown in Fig. 1.

The annotations are stored in a relational database with an \( n : m \) mapping towards biological objects. Cross-referencing was implemented where possible by using links to external public resources (mostly accession numbers like GenBank entries), in other areas it was done manually. For all the cross-referencing, mapping tables were constructed, which list the corresponding identifiers/accession numbers in the source databases. The update plan to cope with evolution of the different resources consists in re-applying the mapping in cases where it is automatically inferred based on external identifiers, and revising the mapping in the case of manual correspondence. In the long term, the source databases (as well as they are under our own control) will be adapted so that they all use the same identifier structures, and everything can be integrated automatically.

It should be noted that the system described here has some overlap with a number of public domain efforts where, however, it is necessary to follow numerous links between independently maintained databases (such as, e.g., functional annotation of gene products (polypeptides) by Swiss-Prot/UniProt [Wu et al., 2006] assigned to genes and their structure reported in Ensembl [Hubbard et al., 2007], just to mention two of the most important and excellently maintained public resources). However, there is no public resource which provides biological contents in a comparably consistent and comprehensive way. Thus, many of the contents described above are represented in an unprecedented way (e.g.,
Fig. 1. The core schema of the integrated database consists of three major building blocks: The biological objects, where the genes, proteins (sensu polypeptides), complexes and small molecules are represented; reaction participants, where the reactions between entities (either of the biological objects) are stored; annotation, all the observed facts reported in the original literature are deposited.

transcriptional regulation information including high throughput data and site models applicable for predictions). Some of these concepts which were largely accepted by the community have been adopted later on by some public domain initiatives, usually as transient efforts bound to limited funding. Data exchange with further resources, being in the public or private domain, is most desirable. Therefore, efforts are under way to generate a SBML representation of the pathway data contained in the resources described here; as soon as the BioPAX consortium has delivered the level 3 definitions for signal transduction data exchange formats, these will be adopted as well.

Integration of Disease Information

The Proteome databases comprise a rich collection of disease data: Connections of the documented genes / proteins to known diseases, categorized by MeSH terms, are given. More specifically, it is denoted for which diseases they may serve as Biomarkers or as therapeutic targets. The molecular effect in connection with the disease is described in a short statement, and negative correlations (i.e. diseases where the respective gene/protein is known not to be involved in) are given as well.

What remains to be done is to integrate the pathologically relevant data that we have collected in PathoDB and PathoSign for proteins of regulatory function. Another resource that will be of enormous help in this is the Human Genome Mutation Database (HGMD), in which more than 67,000 inheritable mutations of the human genome have been described, including their location in certain functional genome regions of the respective genes (e.g., coding, intronic, upstream regions), and what their disease correlation is [Cooper et al., 2005; Khan et al., 2006]. All these resources, when integrated, will enable us to provide a more consistent picture on disease mechanisms by closing the gap between genotype, molecular and clinical phenotype.
The combination of databases with focus on different aspects of biological processes also precipitated an enrichment in object types and reports in the integrated database. Whereas the Proteome databases originally only contained one kind of report for genes and their products, and one for diseases, now additional reports for data types like pathways, DNA binding sites, composite regulatory elements are available, and the function of the genes, gene products, their complexes and modified forms is clearly modularized.

With one of the next steps, the integrated data view will also provide a glance on public domain contents in a specific window, rather than just to give hyperlinks to the corresponding entries in other databases. To make these external data amenable to the same search engine that is used for the internal contents, integrated search indices will be generated.

**Retrieval of Quantitative Data**

Systems Biology aims at exceeding the static view on biological objects such as genes and proteins and their relations, by simulating the dynamics of their interaction networks. However, it can be doubted whether we can realistically expect to gain quantitative knowledge about all reactions in a cell so that we can simulate the complete integrated networks (regulatory and metabolic) in a mathematically exact way, for instance by ordinary differential equations (ODEs). A number of sophisticated methods for more qualitative or semi-quantitative simulations such as Boolean or Petri nets, have been proposed, which again have their own specific limitations. One of the most promising approaches, however, Hybrid Functional Petri Nets (HFPN) goes back one step by allowing inclusion of ODEs for the key steps of a network [Matsuno et al., 2003]. Here, we need again knowledge at least about the most critical and rate-determining steps in a pathway or network.

Much of this knowledge has already been gathered in databases like BRENDA, where kinetic and thermodynamic data for many enzymes from all biological realms are stored [Schomburg et al., 2004; Barthelmes et al., 2007]. While these are mostly referring to substrate interactions of metabolic enzymes, we decided to retrieve additionally and systematically quantitative (i.e., both kinetic and thermodynamic) data about protein-protein and protein-DNA interactions, which are of particular importance for the regulatory processes in a cell.

However, these data are usually not mentioned in the abstract of the original publications, but are hidden in the full text body, or presented in tables. The usual text mining approaches will therefore fail to systematically retrieve this kind of information rendering full-text annotation unavoidable. We have started to extract these data for the COMBIO consortium, to provide quantitative data for the p53 network and for spindle formation reactions, both processes being of crucial importance for cell cycle progression. As a first step towards modeling also the pathological aberrations, for which exact quantitative data will be available only in rare cases, we will include semi-quantitative assessments about “residual activities”, which is usually given in the corresponding research reports.

**Availability**

The integrated version of the mentioned resources will be available from BIOBASE from late spring on upon request. A prototype will be made available on the BIOBASE Web server (http://www.biobase-international.com). Database versions that are freely accessible for users from non-profit organizations can be found on the Gene Regulation Portal (http://www.gene-regulation.de) or on the server of the Department of Bioinformatics, Medical School, University of Göttingen (http://www.bioinf.med.uni-goettingen.de).
ACKNOWLEDGEMENTS

Parts of this work are supported by the European Commission under LSHG-CT-2004-503568 “COM-BIO”.

REFERENCES