Briefings in Bioinformatics Advance Access published December 10, 2008

doi:10.1093/bib/bbn038

Building a knowledge base for systems pathology

Holger Michael, Jennifer Hogan, Alexander Kel, Olga Kel-Margoulis, Frank Schacherer, Nico Voss and Edgar Wingender

Submitted: 3rd June 2008; Received (in revised form): 14th August 2008

Abstract

Translating the exponentially growing amount of omics data into knowledge usable for a personalized medicine approach poses a formidable challenge. In this article—taking diabetes as a use case—we present strategies for developing data repositories into computer-accessible knowledge sources that can be used for a systemic view on the molecular causes of diseases, thus laying the foundation for systems pathology.

Keywords: *databases; content integration; systems pathology; personalized medicine*

INTRODUCTION

Recent advances in molecular biology and genetic research have produced a huge amount of genomic and proteomic data. It is envisaged that these data and the knowledge that can be harvested from them will help to develop personalized and tailored drugs, and their administration thus precisely targeting the specific molecular defects of a given patient [1, 2].

In order to cope with this deluge and to translate data into knowledge that can be harnessed for the development of a personalized medicine, it is a mandatory first step to integrate them in a sensible way. This can also be considered as a first step in the development of a systemic view on the molecular processes that underlie diseases, a 'systems pathology'. Integration in this sense means not only to join data from diverse sources, or to make them at least interoperable, but also to fit them to various customized workflows, and to place them at the disposal of algorithm-based software tools, which can interpret experimental data and formulate predictions.

In this review, we describe several approaches we have used to develop our databases beyond knowledge repositories into platforms that will serve as indispensable tools for explaining biology and molecular medicine. We will present a case study—getting to the molecular causes of diabetes—that highlights the challenges of data integration, follow with a description of our approaches to content as well as workflow integration, and return to the diabetes example to look at how our strategies have worked out in this particular case.

Corresponding author. Edgar Wingender, Department of Bioinformatics, Goldschmidtstr. 1, D-37077 Göttingen, Germany. Tel: +49 (0)551 39 14912; Fax: +49 (0)551 39 14914; E-mail: e.wingender@med.uni-goettingen.de

Holger Michael is a biologist with experiences in yeast genetics and ontology development. He is currently working as CSO assistant at BIOBASE GmbH, Germany.

Jennifer Hogan is a biologist who has worked in the field of developmental signaling. She currently holds the position of Vice President Product Development at BIOBASE Corporation.

Alexander Kel currently holds a position of Senior Vice President R&D of BIOBASE. He has extensive experience in various fields of bioinformatics and systems biology, including promoter analysis and modeling of gene regulatory networks.

Olga Kel-Margoulis currently holds the position of VP Database Curation at BIOBASE GmbH. Her main responsibility is organizing the annotation process and coordination of BIOBASEs databases, and selection of scientific topics for progressive updating of databases.

Frank Schacherer practices biological data integration for over 10 years. He has developed microarray and pathway analysis systems and databases and worked as an industry consultant on data integration projects.

Nico Voss is an informatician with experiences in software development, databases and graph-algorithms. He is developing algorithms and pathway visualization software for analysis systems and participates in database development and regular database build processes.

Edgar Wingender is a 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.

THE CHALLENGES OF DATA INTEGRATION: THE DIABETES CASE

To exemplify how we can significantly benefit when successfully integrating the relevant data, we are sketching here how our databases are employed to give a comprehensive view on a diabetes type 2 disease (MODY, Maturity Onset Diabetes of the Young). Type 2 diabetes results from a reduced ability of the pancreatic β cells to secrete enough insulin, in a timely manner, to stimulate glucose utilization by peripheral tissues. Initially, this causes impaired glucose tolerance, i.e. a reduced capacity to clear glucose from the blood following a glucose load. As the secretory capacity of pancreatic β cells further deteriorates, there is a progressive increase in fasting glucose levels, until the patient develops hyperglycemia in both the fast and fed state. The present inability to propose efficient and convincing strategies for prevention and treatment of MODY is due to the poor understanding of the key pathophysiological mechanisms.

As part of the HumanPSD database, which provides information about the complete proteomes of human, mouse and rat [3], HumanPSD Disease View presents information about diseases including links of genes and proteins to disease terms along with general disease characteristics, links to MESH terms, and synonyms. The links between diseases and genes are given in an ontological way, and genes or proteins are associated with disease terms as biomarkers, therapeutic targets, mouse knockout models, genes involved in molecular mechanisms of diseases, molecular alterations of genes and mRNAs or proteins in diseases. Here, each individual link of a gene or protein to a disease term or to an expression term is collected from a published paper, and correspondingly is accompanied by details of evidence and a PMID link. Systematic collection of these pieces of information over the years resulted in a possibility to make one simple query for all human genes and proteins associated with given a disease, and get a comprehensive list.

Integration of disease-related data with information on transcription regulation and signaling and metabolic networks, collected in the databases TRANSFAC [4] and TRANSPATH [5], respectively, allows to generate new knowledge while constructing queries through the integrated BKL (BIOBASE Knowledge Library). A schematic of the integrated core of BKL is given in Figure 1. For each gene or protein associated with a disease, BKL provides signaling and metabolic pathways where this molecule is a part of gene transcription regulation and TFs involved, as well as interactions of proteins with endogenous metabolites.

Transcription regulation of genes associated with diseases and transcription regulatory networks

As for the example to be described here, transcription factors (TFs) HNF1A, HNF3B, HNF4A, HNF6 and IPF1 are among the proteins that are critically involved in MODY, and details are documented in HumanPSD Disease View. Integration of this data with TRANSFAC enables us to retrieve a transcription regulatory network where all mentioned TFs are involved (Figure 2A). In this figure, each individual link from a TF to a gene refers to the fact of transcription regulation published in a particular paper; each link shown in the figure is represented by corresponding 'Site entry' in TRANSFAC, including details about the site sequence and the genomic

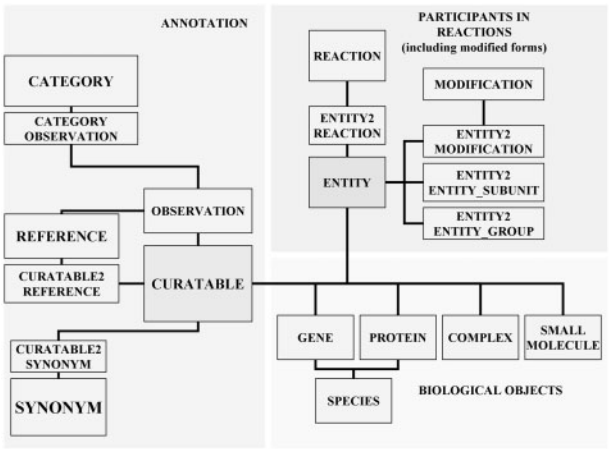

Figure 1: Simplified block layout of the database schema with key tables. The central table is Curatable, which stands for anything that can have curation attached, either as free text annotations or as structured vocabulary from ontologies, here shown as category table. Such curated data are extracted from literature, thus Curatable is linked to a reference table. Synonym handling for all kinds of objects is also tracked at this level. Curatable is a supertype of the various classes of biological objects—genes, proteins, small molecules and so forth, which have additional, type-specific fields. Such entities participate in interactions and chemical reactions, which are tracked in the Reaction table. As biomolecules often interact in modified form, the Entity table allows to attach information about their state such as phosphorylation.

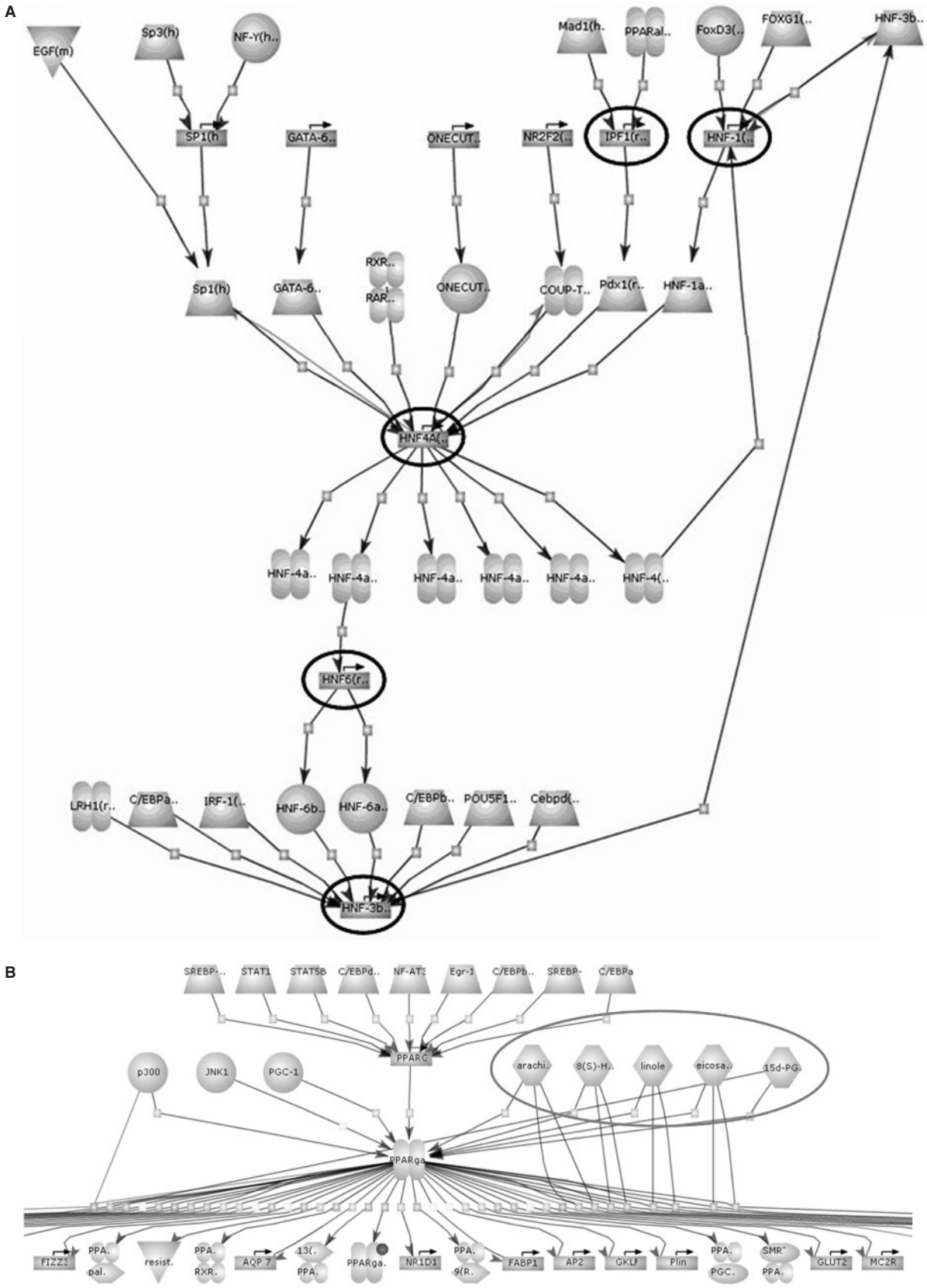

Figure 2: (**A**) Transcription regulatory network of TFs critical for MODY, HNF1A, HNF3B, HNF4A, HNF6 and IPF1. (**B**) Fragment of unified signal transduction and transcription regulatory network upstream and downstream of PPARgamma. Activation of PPARgamma by endogenous metabolites is highlighted by an oval on the right. (**C**) Legend to Figure 1A and B.

location, experimental evidence for binding of TF to a promoter/enhancer and a PMID link. Each protein shown in the figure refers to the 'Protein page' in HumanPSD where one can immediately find more details about the protein molecule, including primary sequence, ontology links to GO terms, to Disease terms, BLAST comparison of the protein sequence with proteins from other organisms and other details. Transcription regulatory networks involving TFs critical in MODY, shown in Figure 1, resulted from data integration in BKL.

Signal transduction networks and regulation by small molecules

PPARγ is another TF which is crucially involved in the development of the MODY state [6, 7]. Integration of transcription regulatory information from TRANSFAC with TRANSPATH reveals how PPARγ is regulated via signal transduction pathways, for example through the insulin pathway (Figure 3A). Similar to other nuclear receptors, PPARγ requires ligand binding for activation. Regulation of PPARγ by ligands, a certain group of small molecules, is also covered (Figure 2B). The TRANSFAC-derived information about how the PPARγ gene itself is transcriptionally regulated by several TFs has been integrated into the same network.

Integration of signal transduction and metabolic pathways

Natural ligands for PPARγa are endogenous metabolites, and these are part of the metabolic pathways collected in TRANSPATH. For example, palmitic acid, a natural ligand for PPARα, -β and -γ, is synthesized from acetyl-CoA. The enzyme for the rate-limiting step of the metabolic pathway, acetyl-coenzyme A carboxylase, is modified via adiponectin signaling (Figure 3B). The adiponectin receptors are thought to play a crucial role in obesity-linked insulin resistance [8].

Thus, integrating disease-related gene and protein information (HumanPSD Disease View) with gene regulatory, signaling and metabolic information (TRANSFAC and TRANSPATH) provides an unprecedented comprehensive view on the integrated network representing causes and effects of a particular disease. Putting together different types of information in BKL allows to elucidate particular aspects and molecular mechanisms of diseases with the help of new knowledge, which clearly resulted from database integration and was not published as such before. The question is how we can systematically retrieve and integrate all relevant information of the principally different networks mentioned above.

CONTENT INTEGRATION

The field of content integration comprises a considerable number of different strategies with diverging focal points, which often makes a comparison quite difficult [9]. Among the multitude of methods for integrating content from diverse sources, e.g. [10, 11], we pursue two complementary strategies. First, the retrieval of data from diverse sources,

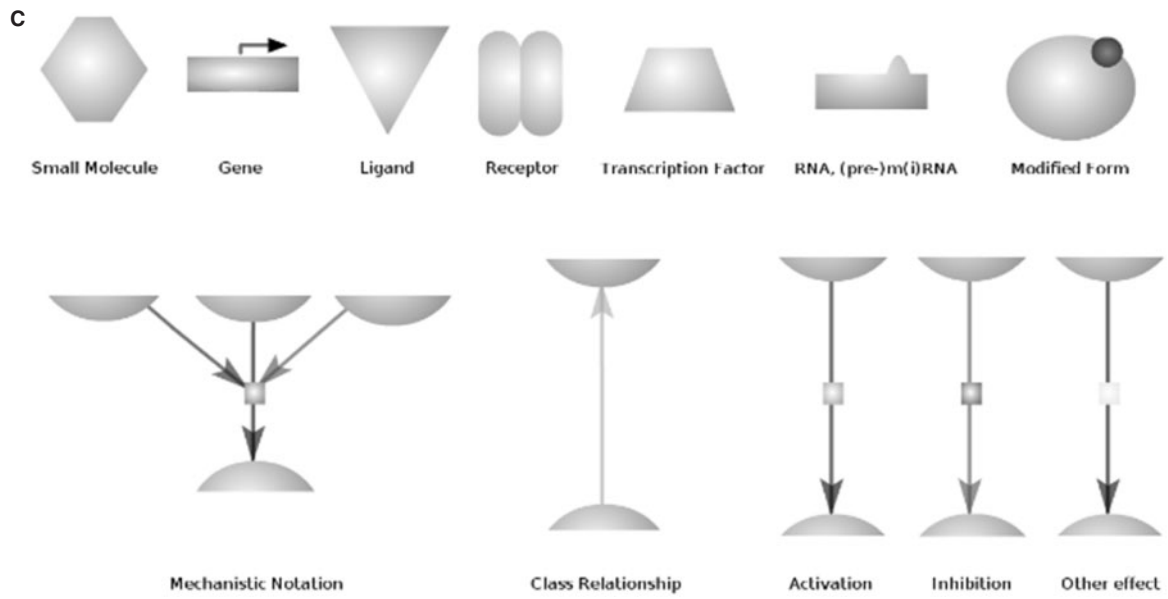

Figure 2: Continued.

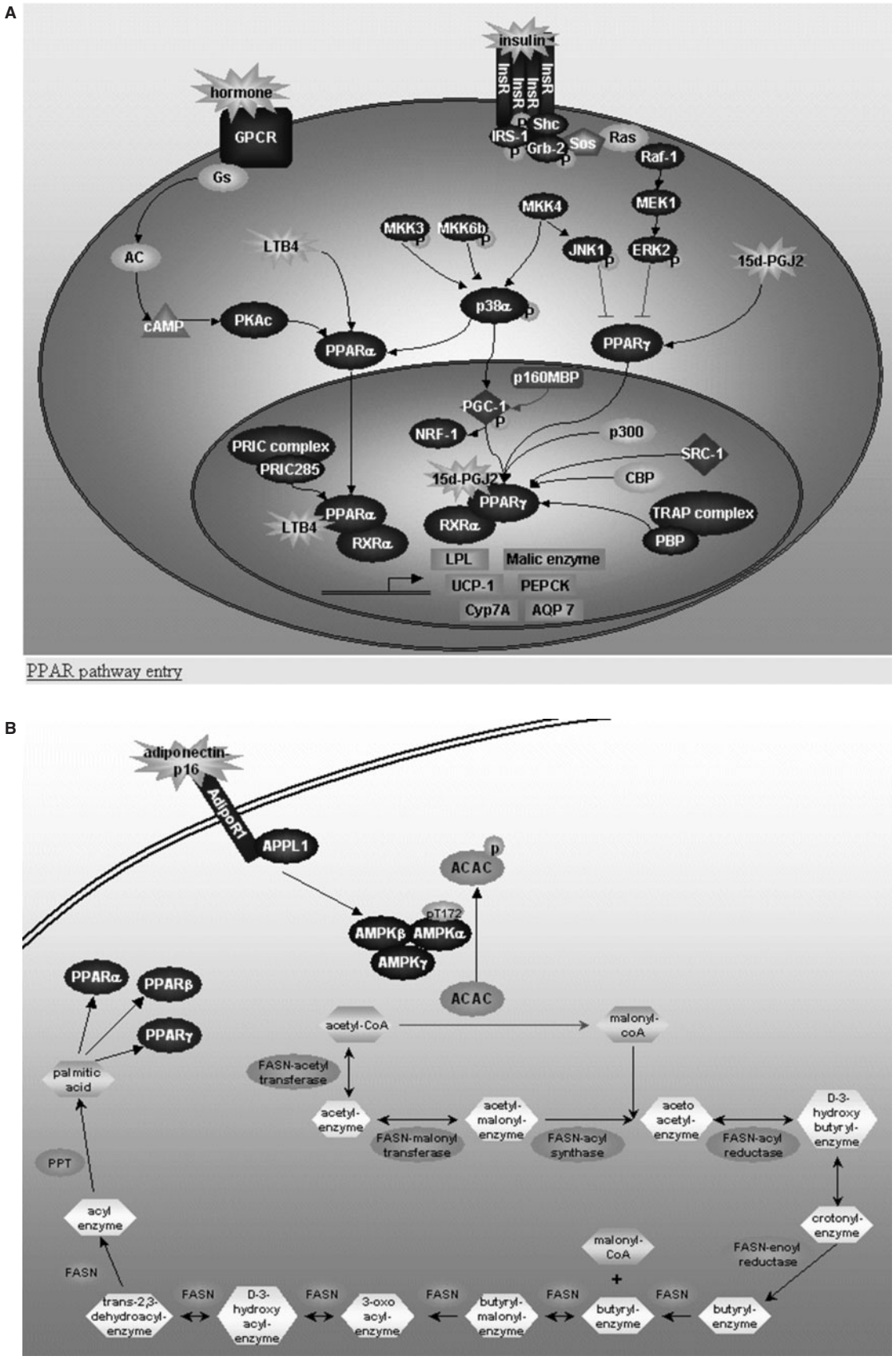

Figure 3: (**A**) Map of PPAR regulation via signaling pathways. (**B**) Integration of signal transduction and metabolic pathways. PPAR regulation via adiponectin signaling involves a metabolic pathway, the synthesis of palmitic acid from acetyl-CoA.

annotation, is organized in such a way as to yield information amenable to further integration. Second, we use controlled vocabularies and ontologies for the annotation of data as well as for integrating and harmonizing data from diverse and disparate sources.

It is almost a truism that ontologies form an indispensable component of any effort to structure, integrate and analyze the huge amounts of biological data resulting from different kinds of high-throughput approaches. A considerable number of ontologies for diverse biomedical purposes have been developed, and efforts are underway to (re-)design them in such a way that they become more interoperable and logically better formed [12]. In parallel, the ontologies existing are put to use in the ongoing data integration projects.

One of the most venerable ontologies in the biomedical field is the Gene Ontology (GO), a controlled vocabulary which describes the attributes of genes and gene products in any organism [13]. GO provides this vocabulary in three sets of explicitly defined, structured vocabularies that describe biological processes, molecular functions and cellular components of gene products in both a computer- and human-readable manner. GO is employed in a wide variety of tools and applications. For example, all protein reports in our HumanPSD database have been assigned to more than 7000 unique GO terms. These annotations have been shown to be quite useful for the analysis of microarray data using the HumanPSD database (included in the Proteome BioKnowledge Library) [14]. The novel gene expression analysis platform ExPlain™ [15], which applies a novel knowledge-driven approach for analyzing of complexes of coexpressed genes, falls back on GO terms to provide the user with functional classifications for the coexpressed genes in a given microarray experiment.

A second ontology with important uses in our projects is CYTOMER. This was originally a relational database that comprised human anatomical structures, tissues, cell types, physiological systems and developmental stages. To achieve more flexibility for the further development of CYTOMER, an OWL-based ontology was derived [16]. The ontology is being actively maintained and developed with regard both to structure and content. A visualization tool based upon the prefuse toolkit, which will be accessible over the web, has recently been devised (manuscript in preparation). CYTOMER is intended to provide a reference for mapping sources of biological activity (e.g. gene expression, hormone production and hormone activity) in the human body. In that capacity, it has been embedded into the Endonet information resource about intercellular regulatory communication [17]. This database comprises information about hormones, their receptors, as well as the locations where they are synthesized and where the receptors are expressed. These diverse locations are mapped to CYTOMER, providing the user with reference points within the anatomical hierarchy.

For the disease module, we developed a special format and vocabulary referred to as 'BioKnowledge LinGO' for capturing the most salient information regarding the association between genes or proteins and diseases in a structured approach that promotes easy data integration, and sophisticated querying, while preserving the human readable format of free text annotations. BioKnowledge LinGO utilizes a combination of public domain ontologies, specifically GO and Medical Subject Headings (MeSH), proprietary controlled vocabularies to mimic the basic subject–verb–object structure of an English sentence. Figure 4 illustrates how BioKnowledge LinGO works to capture the link between a gene or protein (subject) and a disease (object).

A generalized disease statement would take this form:

> Increased [molecular function] of Protein X correlates with decreased [symptom] of Disease Y

Disease annotations are constructed by associating a subject phrase (gene or protein data) to an object phrase (disease data) via the linking term (represented by the horizontal 'verb' bar). The subject phrase discusses a change in a biological parameter of the gene or protein and includes data relating to one of the following concepts: the protein's

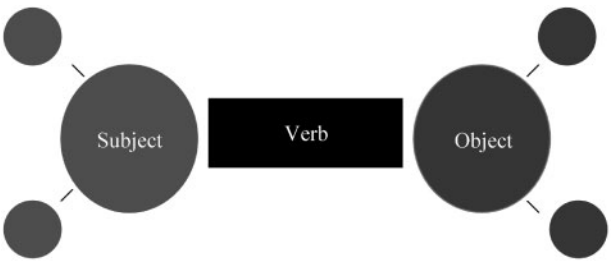

Figure 4: BioKnowledge LinGO schema for capturing the link between a gene or protein (subject) and a disease (object).

molecular function, the protein's cellular localization, the protein's expression in a specific type of cell, the protein's expression in a specific tissue, a specific domain of the protein, specific mutations in the gene or protein, microbial exploitation of the protein during infectious diseases, any specific protein modification (e.g. phosphorylation, etc.) or types of antibodies to the protein. Object phrases discuss a change in a parameter of the disease and includes data relating to one of the following concepts: a subtype or form of the disease, a specific type of tumor development associated with a disease, a defect in organ development associated with a disease, a defect in organ function associated with a disease, a defect in cellular function associated with a disease, a symptom or phenotype of a disease, a defect in cell proliferation associated with a disease, a defect in cell differentiation associated with a disease, a defect in a cell process associated with a disease or a defect in a biological process associated with a disease. The subject phrase is then linked to an object phrase using a term that defines either a cause-and-effect relationship, a correlative relationship or a preventive relationship. Each subject phrase and object phrase can be further modified to indicate changes in the above concepts.

Structure and example of a disease annotation using controlled vocabularies and BioKnowledge LinGO:

Adverb1–Subject Verb–PROTEIN–Linking Term–Adverb2–Object Verb–DISEASE

example:

decreased expression of OCLN in colon/large intestine correlates with increased cellular extravasation associated with Ulcerative Colitis

Adverb1 = 'decreased'
Subject Verb = 'expression in colon/large intestine'
PROTEIN = OCLN
Linking Term = 'correlates with'
Adverb2 = 'increased'
Object Verb = 'cellular extravasation'
DISEASE = Ulcerative Colitis

Finally, to promote sorting and retrieval by general relationship type, each annotation is assigned to one of four categories: 'biomarker' (correlates the presence of the protein to a disease, does not necessarily involve the protein in the pathogenesis), 'molecular mechanism' (associates a role for the protein in causing or maintaining the pathogenesis of the disease), 'therapeutic target '(associates a potential therapeutic role for the protein in preventing or ameliorating symptoms of the disease) or 'negative correlation' (excludes a protein from causing or maintaining the pathogenesis of the disease).

The approach described above allows us to retrieve quite complex disease-related information. The next task is to relate molecular causes to physiological outcomes, and consequently to make predictions across different levels of abstraction. Systems modeling requires a concept for capturing information about biological entities on all these different levels. Only if molecular findings are combined with systemic ones, can they be used to predict systemic outcomes.

Such models can support different applications. One is to overlay or train them with high-throughput data, to build predictors. Another, which we discuss here in more detail, is to answer questions a human investigator might have about biomedical knowledge.

Information on molecules and their relationships is classically provided in various bioinformatics databases on genes, proteins and metabolites. These high-quality resources have usually been curated from scientific publications. There are also domain-specific data resources, both databases and ontologies with information on interactions, pathways, disease states, cellular conditions and so on. A hard problem with these databases is that data is siloed, and integrating it is an ongoing and sometimes frustrating challenge. There are two starting points for automated integration at present: direct links and shared objects.

In some cases, databases and ontologies have direct links to objects from other sources and levels, for example signaling molecules to pathways or domains to proteins. Each of these relationships is tailor-made, which leads to complex data structures, and makes it hard to develop general algorithms against the data that would allow smart searches across databases. Also, with the number of data types growing, explicit modeling of their specific relation falls short. And, even if all these many one-to-one relationships between data sources were available, it still would be difficult to reflect more complex relationships of three or more types.

A second integration strategy is shared elements. To see which diseases are correlated with a pathway, one can do so by comparing which genes are shared between disease and pathway, and guess that there may be a relation. While this is powerful and elegant,

there are no validated, direct statements. All is conjecture. Is this so because we just do not know about these level-spanning relationships, that is: does this state of affairs reflect the knowledge about biology?

Rich, complex, system spanning statements are in fact quite common in scientific publications. Such information must therefore be valuable to understand the biology. It would be powerful to have these in a form amenable to computation.

Anypath is a concept we introduce here for making these combinations explicit in a general manner. It is related to research on semantic networks, topic maps and entity relationship modeling, but differs in that it makes heavy use of a concept known as 'reification', the use of relationships as inputs or targets of other relationships.

Words in sentences are first tagged either as entities, which can be recognized as belonging to types for which type hierarchies like ontologies can exist, or as relationships, or not at all, if they are 'filler words' that do not contribute materially to the main statement.

Relationships can be directed, and most importantly can have other relationships as their input or output. This feature allows representation of how relationships influence other relationships and thus can capture the complex statements common in literature in a computable graph structure. Each statement is thus turned into a small graph, the statement graph.

Many graphs share entities, and the overall graph constructed for the whole body of statements by merging all the statement graphs is a knowledge network. This overall relationship graph can be used for making inferences that are not supported explicitly by a single document or statement.

Relationships in this model also can have conditions and qualifiers. Conditions are a parsimony feature and modify an entity, for example phosphorylated forms of a protein to that protein. In their absence, either explicit relationships between the entity and modification would have to be constructed, which is tedious, or each modified entity would have to be an object in it's own right, which would lose the information that they share the same 'base' entity. Qualifiers are statements that give meta-information about the relationship statement, for example the degree of confidence that the author had in the relationship, or whether the relationship is generally or partially applicable.

While manual assignment will provide the highest quality result, especially in the case of complex sentences, for simple sentences it could conceivably also be done by an automated text mining system. Figure 5 shows an example statement and a graph extracted from it.

Any information system is only as good as the information you can get out of it. As a baseline comparison, consider a simple indexing of all words in a document, so that the document or sentences can be retrieved based on keyword search over all the abstracts. This is the Pubmed or Google approach, and anything that generates extra effort needs to be able to beat it in usefulness.

Tagging words with types or mapping them to ontologies allows the use of more general terms to match specific terms in a search. This feature can be considered an improvement as it allows aggregation and drill down in queries.

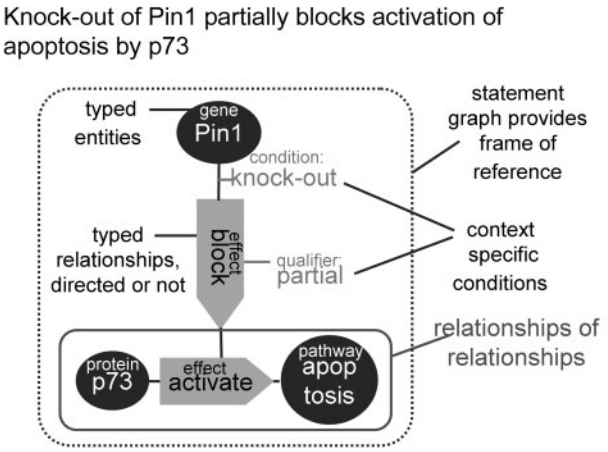

Figure 5: A graph of biological entities and their relationships, similar to many other pathway databases. Distinguishing features are: (i) Not limited to molecules. Entities are typed, allowing all kinds of them: genes, proteins, pathways, diseases and so on. These can be mapped to ontologies. (ii) Relationships between entities can be treated as entities in their own right, similar to natural language where clauses can act as subjects or objects. This feature is called 'objectification' or 'reification'. (iii) Conditions like knock-out, time, dosage, phosphorylation which modify the entities can be attached to the link between entity and relationship, to express that the entity does not always exhibit that behavior, but only under those conditions. This allows aggregation to common entities, like many different relationships known for a protein, while accurately depicting that the protein will only do something when phosphorylated. Multiple conditions can be combined. (A different approach to the same is to create lots of variants of the protein, and subsume them under a generalized version.) (iv) The source abstract or sentence automatically provides a limiting context.

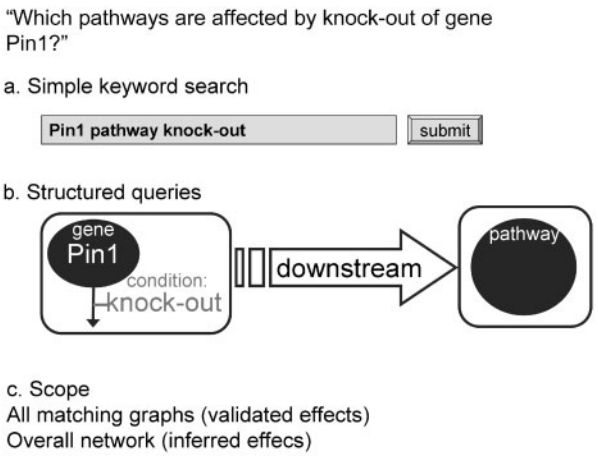

Figure 6: Possible queries. (**A**) The simplest way to query is to look in all entity types and entities for keywords, and return lists of matching statement graphs or subnets of the overall graph. (**B**) To make use of the structural relationships, queries can be formulated that describe also the relation between the search types and entities (for example any connection within a certain distance, upstream, downstream). (**C**) The statement or document scope and shared entities between statement graphs allow for a choice between traceable author statements and searching for inferred connections across documents.

Adding explicit relationships with direction, both extend this to searches that constrain how the entities are related to each other and to searches against the overall body of knowledge that span several source documents. The user can decide to either only look for validated statements from a single source, or for statements and relationships that span across sources.

The result from such searches can be a simple list of documents, or an extraction of the relationship graph. Figure 6 shows an example question and several ways to ask it.

WORKFLOW INTEGRATION

The approach we use and present here is based on custom integration of existing ontologies or databases to construct a dedicated database and interface. This yields high-quality results, but is labor-intensive. Therefore, it is hard to replicate for groups where data integration is a necessary evil, not the focus of their work. One way to make this approach work is to provide integration centrally for a large user base. In fact, public sites like InterPro [18], which integrates domain information, or Galaxy, an environment for genomic analysis (http://galaxy.psu.edu/), provide exactly this service for their area of focus.

As comprehensive semantic integration of data turns out to be an ever-elusive challenge, practical solutions that settle for simpler, more modest goals have been employed profitably.

One traditional approach in this direction is to trade off semantical homogeneity for larger scope of integration. It is employed by integration systems such as SRS [19] and BioRS [20] that use shared identifiers to automatically relate content from different data sources and allow queries across data sources.

A second approach is to trade off data atom level integration against compilation of data blocks. This is employed by web services like the Distributed Annotation System [21]. Data is provided online in a machine readable, published format. It thus can be requested from another application and included in the results of that application on the fly. This has the advantage that there is no need to develop and maintain importers for data updates, data is automatically always current and at the same time data can be presented in an aggregated way to the end user. Because the exchange formats are published and usually employ extensible schemata like XML, there is also an incentive to maintain backward compatibility and honor the exchange format. This is a boon because it introduces a layer of insulation between the data provider and the consumer. Integration has traditionally been plagued by providers changing their internal structures, names, tables and formats in their quest to improve their offering and thus breaking downstream code that consumers wrote to integrate it with other data.

A third approach is to trade off width of integration for more depth in regard to a particular analysis task. This is employed by workflow systems such as the InforSense platform [22], or Taverna workbench [23] which act as a glue layer between various data sources and analysis packages. The task-driven approach employs customization to set up a workflow for each task. There is a need to develop adapters for the various tools to be integrated, which has beneficial side effects: it forces tool and database providers to define clear interfaces for adapter development, and the resulting adaptors are reusable components that can be distributed with the workflow platform and used to construct other workflows. Thus, this approach contributes to the integration challenge by fostering creation of

standard components that work together while creating usable individual solutions.

All three approaches are being employed in today's bioinformatics landscape. They are successful because of their modesty and because they embrace heterogeneity, distributed authority and federation.

APPLYING INTEGRATION STRATEGIES: DIABETES REVISITED

Task-oriented integration strategies can help to infer novel information on disease mechanisms. We can exemplify this when analyzing the available information on type 2 diabetes mellitus (MODY) in the BKL database and applying an integrated promoter and pathway analysis onto these data.

As a first step in our attempts to understand molecular mechanisms of such a complex disease as type 2 diabetes, we had collected all available information about known genes and proteins associated with this disease as described above. Now, a query to the BKL database using the keyword 'diabetes' results in 10 entries on different levels of abstraction. One of them provides detailed information on type 2 diabetes mellitus, which includes a listing of proteins known to be associated with this disease as biomarkers, therapeutic targets or as part of the molecular mechanism. This list of proteins is one important entry point for further data mining of BKL information. A flexible mechanism implemented for data mining and navigation through the system leads us, in a first path, to a list of 265 human genes encoding diabetes-relevant proteins (see table in Appendix 1).

One of the most interesting questions on the molecular mechanisms of diseases is how the genes linked to the disease are transcriptionally regulated. Because of the integration with TRANSFAC contents, we can extract 658 known transcription factor binding sites (TFBS) in the promoters and enhancers of these genes.

The next step in the analysis of diabetes-related genes involves the computational platform ExPlain™. It allows to perform statistical estimations

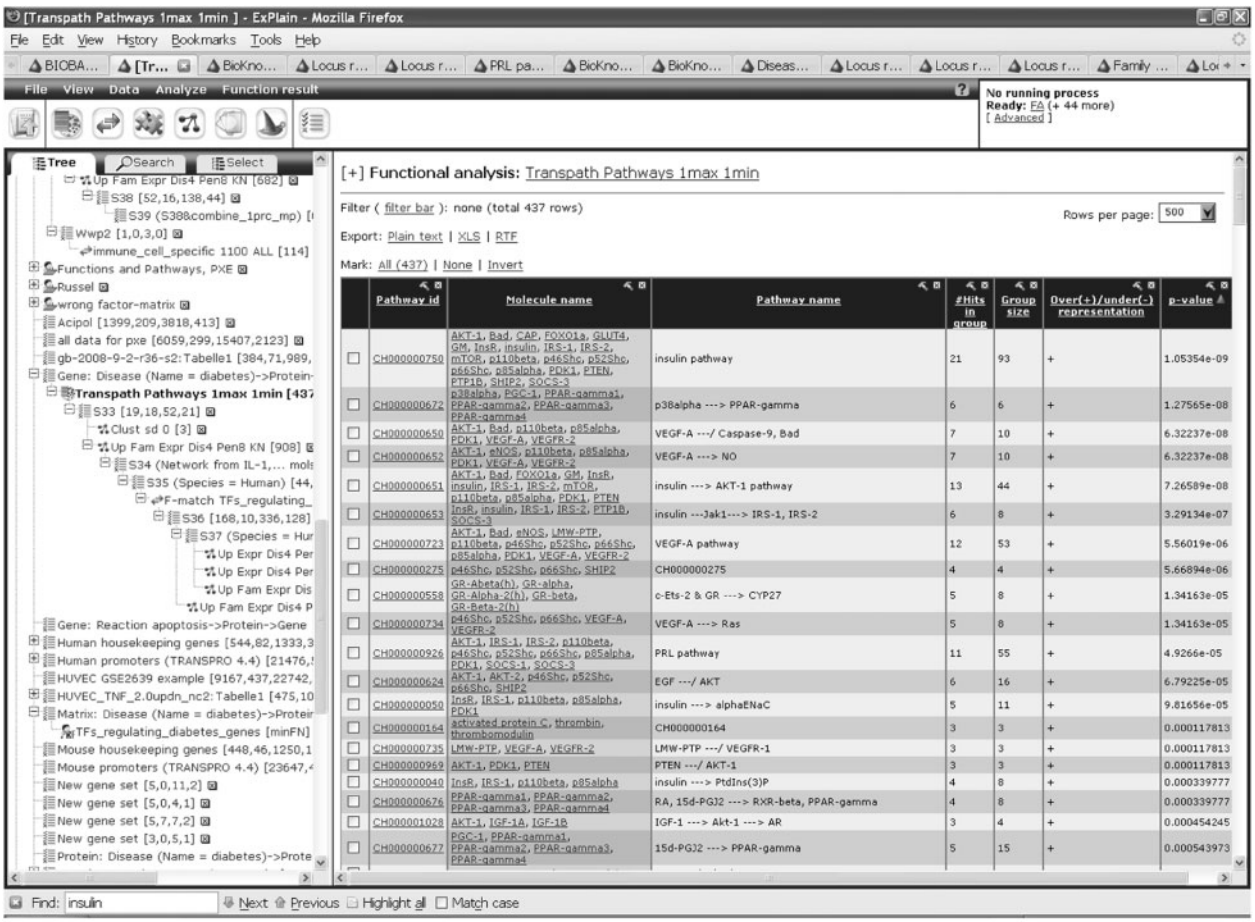

Figure 7: Results of a pathway analysis, illustrating the downstream analysis workflow of ExPlain.

of association of these genes with various signal transduction and metabolic pathways. The analysis of the statistical association is done by direct mapping of the products of the genes on the diagrams of canonical pathways collected in TRANSPATH, and computing a *P*-value using a hyper-geometric distribution (see description in Appendix 2). In Figure 7, the results of such a pathway analysis which we refer as 'downstream analysis workflow' is shown [see table in Appendix 3 for the full list of identified statistically significant pathways ($P < 0.01$)]. It is interesting to see that the insulin pathway is the most significantly enriched with diabetes-related genes (*P*-value $= 10^{-9}$). The VEGF and PRL pathways have been found to be also highly enriched (*P*-values 1.3×10^{-5} and 4.2×10^{-5}, respectively). In order to identify crosstalks between different pathways which might be influenced by the 'disease' genes, we applied a network clustering algorithm which searches through the entire network of signal transduction interactions in the cell starting from the 'disease' genes and their products and identifies groups of closely connected molecules (network clusters). A formal description of this algorithm is given in Appendix 4. In Figure 8, we demonstrate a network cluster, which we identified using this algorithm. This cluster connects 16 diabetes-related proteins of the insulin pathway [shown as the end

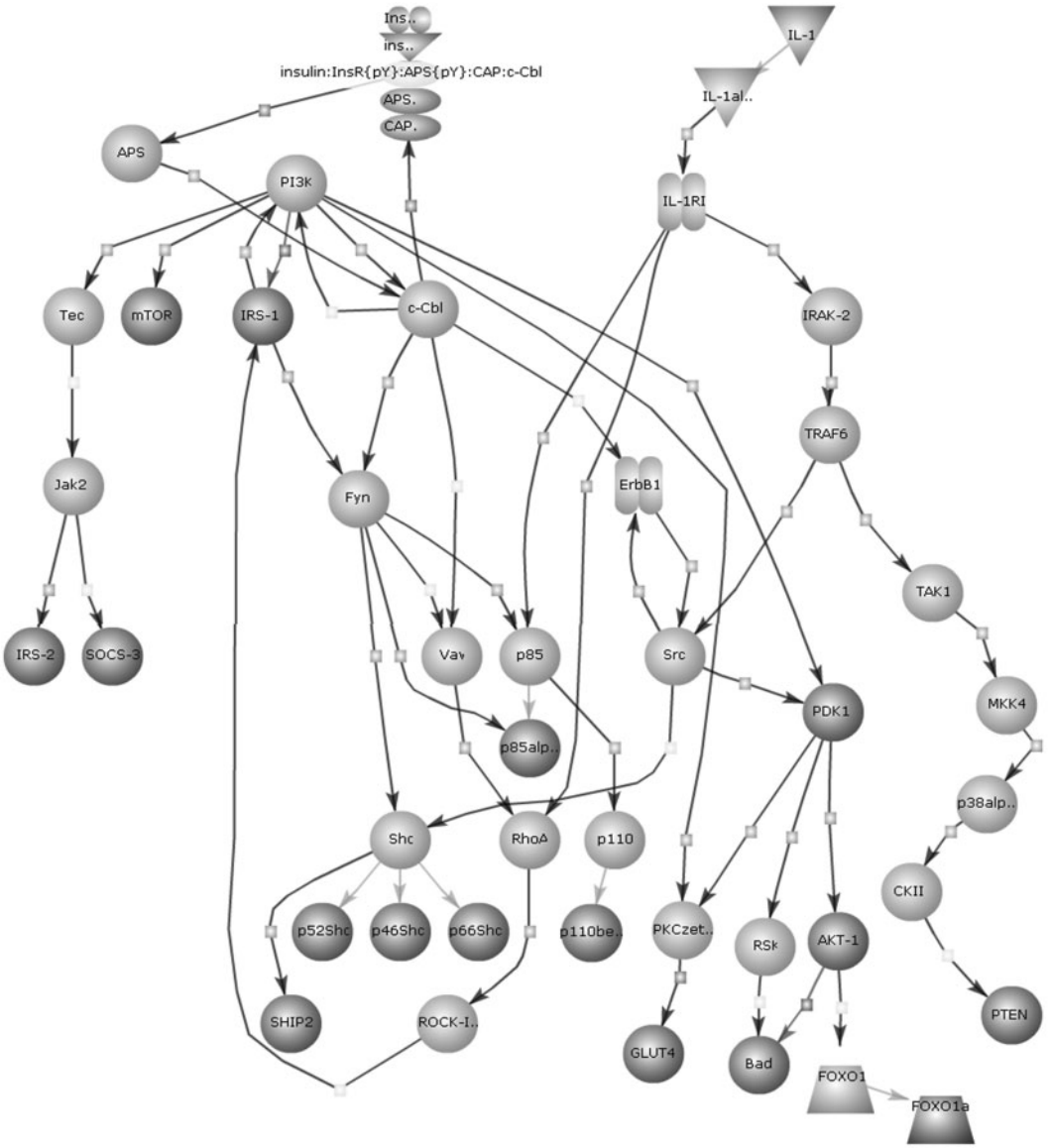

Figure 8: Network cluster connecting 16 diabetes-related proteins of the insulin pathway (dark grey balls along with FOXO1a, lower right corner).

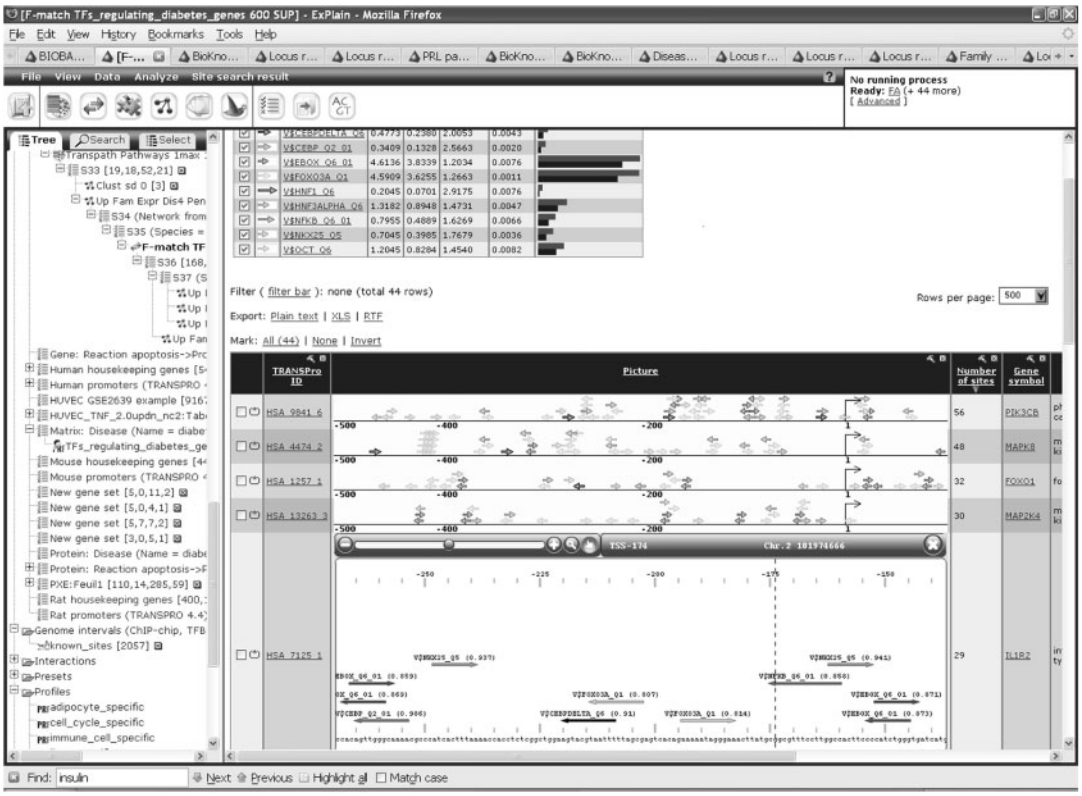

Figure 9: Result of an upstream analysis depicting maps of the predicted TFBSs in the promoters of first five diabetes-related genes (see Appendix 6).

nodes of the graph (blue circle)]. This cluster clearly shows the crosstalk of the signal flow from the insulin-receptor complex and the IL-1 signaling pathway (top nodes in the graph). Indeed, a query to the disease module of BKL shows that there is a growing body of evidence of relations between diabetes mellitus, type 2 and IL-1 signaling (such as correlation of the abnormal expression and secretion of IL-1beta, as well as small nucleotide polymorphisms (SNPs) in IL-1beta gene correlate with increased incidences of type 2 diabetes susceptibility) [24–26]. Therefore, such comprehensive integration of the information in the 'downstream analysis workflow' helps us to derive new information connecting disease and signaling pathways. We were able to identify the most important components of the signal transduction machinery and cellular metabolic mechanisms, which are affected in diabetes due to impaired activity or expression of the disease-related genes. It also helps us to identify genes as novel promising diabetes biomarkers—those genes which can be found in the network neighborhood of known diabetes-related genes.

In order to understand which of these TFs are involved in the regulation of genes in the diabetes pathogenesis, we applied the so-called 'upstream analysis workflow' of ExPlain^{TM}. Its core consists of the Composite Module Analyst (CMA) [27, 28], and it is fully integrated with the contents of the BKL database that it uses for the analysis. In the upstream analysis workflow, we interrogate promoter sequences of the genes under study.

We have applied this upstream analysis workflow to the promoters of the genes whose products are included in the 'diabetes-related insulin/IL-1 network cluster' revealed above. This cluster includes products of 44 human genes [127 alternative promoters of these genes are taken from the TRANSPRO database; see Appendix 5 for a full list of promoters with genomic positions of transcription start site (TSS) considered in the analysis]. Based on information in TRANSFAC, we have built a profile of TFs, which are known to be regulators of the genes under analysis. Among them, we found that sites for such factors as C/EBP, FOXO3, HNF1, HNF3, NF-kB, OCT and NKX25 are significantly overrepresented in the promoters of these genes (1 kb upstream to 100 bp downstream from the TSS), indicating their high importance in the regulation of this group of genes. In Figure 9, we show maps of the predicted TF-binding sites in

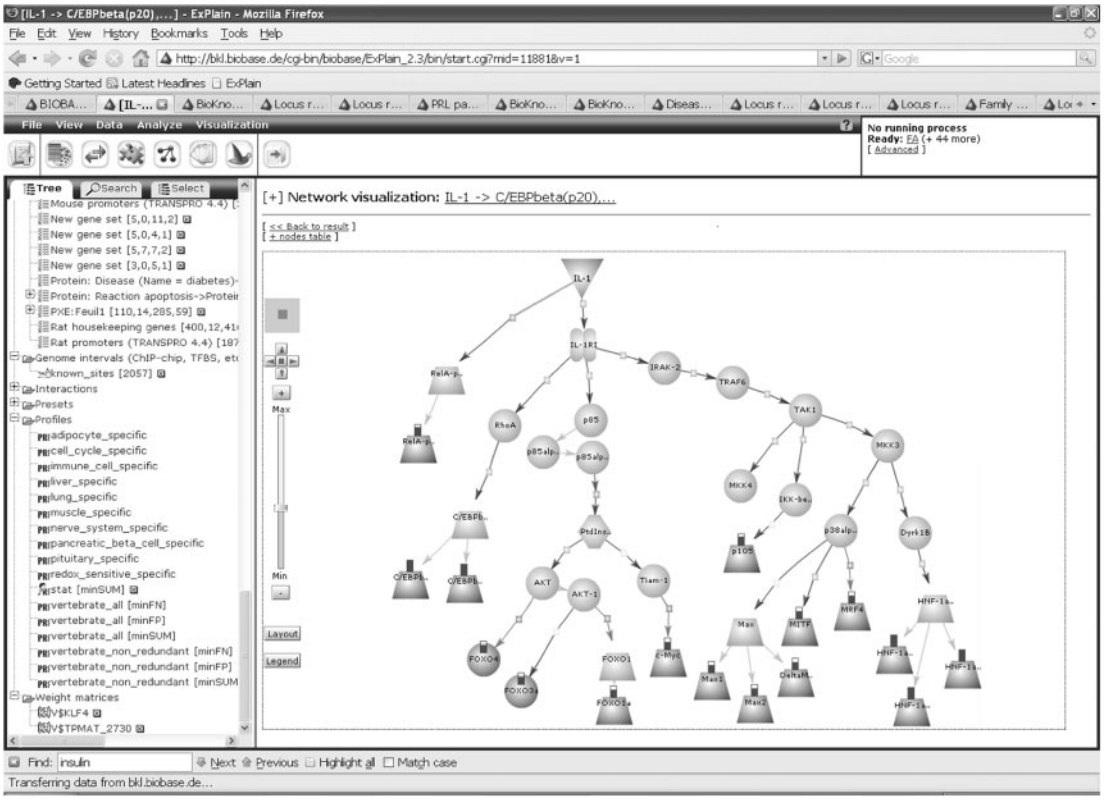

Figure 10: Application of the key-node search algorithm, revealing the IL-1 family as the most significant key node.

the promoters of first five diabetes-related genes. The full table of the predicted sites are given in Appendix 6.

The next challenging question then is identification of a common reason leading to the concerted regulation of activity of all involved TFs. Answering this question can help to build a causal understanding of the molecular reasons of the disease. Using TRANSPATH, we constructed a graph G as a representation of the signal transduction network of cells. Earlier, we developed the graph-based algorithm (key-node search algorithm) [29] to identify 'upstream' key signaling molecules that are potentially involved in activation/inhibition of the 'target set' of TFs.

We applied the key-node search algorithm to the list of TFs identified in the promoter analysis. As one can see in Figure 10, the most significant key-node identified in this analysis is the IL-1 family.

In summary, we can see that both downstream as well as upstream analysis of the set of diabetes-related genes identify strong relation with IL-1 signaling. Using the integrative workflow system, we were able to reveal this novel details of association of diabetes with the signaling pathways of IL-1. We can speculate that certain modes of action of the IL-1 signalling system, which is characterized by strong crosstalks with the insulin pathway, can be involved in autostimulating positive feedback loops leading to the formation of the diabetes pathologies.

Key Points

- Overall goal: developing data repositories into tools for personalized medicine.
- Integration of diverse data sources as well as annotation of data in an integration-friendly way, employing controlled vocabularies and ontologies.
- Integration of databases with diverse algorithm-based tools to obtain harmonious workflows.

Acknowledgements

We are indebted to Prof. H. Tanaka (Tokyo Medical and Dental University, Department of Bioinformatics) who first introduced us to the concept of systems pathology. TRANSFAC, TRANSPATH and CYTOMER are registered trademarks; BKL, BIOBASE Knowledge Library, HumanPSD, HumanPSD Disease View and ExPlain are trademarks of BIOBASE GmbH, Germany.

FUNDING

The Eurodia project (EU grant no. 518153).

References

1. Dietel M, Sers C. Personalized medicine and development of targeted therapies: the upcoming challenge for diagnostic molecular pathology. A review. *Virchows Arch* 2006;**448**:744–55.
2. van der Greef J, Martin S, Juhasz P, *et al*. The art and practice of systems biology in medicine: mapping patterns of relationships. *J Proteome Res* 2007;**6**:1540–59.
3. Hodges PE, Carrico PM, Hogan JD, *et al*. 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* 2002;**30**:137–41.
4. Matys V, Kel-Margoulis O, Fricke E, *et al*. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;**34**:D108–10.
5. Krull M, Pistor S, Voss N, *et al*. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res* 2006;**34**:D546–51.
6. Lehmann JM, Moore LB, Smith-Oliver TA, *et al*. An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma). *J Biol Chem* 1995;**270**:12953–6.
7. Guilherme A, Virbasius JV, Puri V, *et al*. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. *Nat Rev Mol Cell Biol* 2008;**9**:367–77.
8. Takashi K, Toshimasa Y, Naoto K, *et al*. Adiponectin and adiponectin receptors in insulin resistance, diabetes, and the metabolic syndrome. *J Clin Invest* 2006;**116**:1784–92.
9. Chua HN, Sung WK, Wong L.. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics* 2007;**23**:3364–73.
10. Boyle J, Cavnor C, Killcoyne S, *et al*. Systems biology driven software design for the research enterprise. *BMC Bioinformatics* 2008;**9**:295.
11. Rubin DL, Shah NH, Noy NF., Biomedical ontologies: a functional perspective. *Brief Bioinform* 2008;**9**:75–90.
12. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.
13. Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Res* 2008;**36**:D440–4.
14. Johnson RJ, Williams JM, Schreiber BM, *et al*. Analysis of Gene Ontology features in microarray data using the Proteome BioKnowledge Library. *In Silico Biol* 2005;**5**: 0035.
15. Wingender E, Crass T, Hogan JD, *et al*. Integrative content-driven concepts for bioinformatics 'beyond the cell'. *J Biosci* 2007;**32**:169–80.
16. Michael H, Chen X, Fricke E, *et al*. Deriving an ontology for human gene expression sources from the CYTOMER database on human organs and cell types. *In Silico Biol* 2004; **5**:0007.
17. Dönitz J, Goemann B, Lizé M, *et al*. EndoNet: an information resource about regulatory networks of cell-to-cell communication. *Nucleic Acids Res* 2008;**36**: D689–94.
18. Mulder NJ, Kersey P, Pruess M, *et al*. In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol Biotechnol* 2008;**38**:165–77.
19. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996;**266**:114–28.
20. Kaps A, Dyshlevoi K, Heumann K, *et al*. The BioRS™ Integration and retrieval system: an open system for distributed data integration. *J Integrative Bioinformatics* 2006; **3**:44.
21. Prlić A, Down TA, Kulesha E, *et al*. Integrating sequence and structural biology with DAS. *BMC Bioinformatics* 2007; **8**:333.
22. InforSense Limited. *The Inforsense Platform*. http://www.inforsense.com/products/core_technology/inforsense_platform/(12 August 2008, date last accessed).
23. Oinn T, Addis M, Ferris J, *et al*. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 2004;**20**:3045–54.
24. Devaraj S, Jialal I. Low-density lipoprotein postsecretory modification, monocyte function, and circulating adhesion molecules in type 2 diabetic patients with and without macrovascular complications: the effect of alpha-tocopherol supplementation. *Circulation* 2000;**102**:191–6.
25. Spranger J, Kroke A, Möhlig M, *et al*. Inflammatory cytokines and the risk to develop type 2 diabetes: results of the prospective population-based European prospective investigation into cancer and nutrition (EPIC)-Potsdam study. *Diabetes* 2003;**52**:812–7.
26. Lee SH, Ihm CG, Sohn SD, *et al*. Polymorphisms in interleukin-1 beta and Interleukin-1 receptor antagonist genes are associated with kidney failure in Korean patients with type 2 diabetes mellitus. *Am J Nephrol* 2004;**24**:410–4.
27. 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.
28. 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.
29. 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.