Article I

Data integration and visualization system for enabling conceptual biology

In: Bioinformatics.
Vol. 21, Suppl. 1, 2005, pp. i177–i185.
Published by permission of Oxford University Press.

BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i177–i185
doi:10.1093/bioinformatics/bti1015

Data integration and visualization system for enabling conceptual biology

Peddinti V. Gopalacharyulu[1], Erno Lindfors[1,]
Catherine Bounsaythip[1], Teemu Kivioja[1], Laxman Yetukuri[1],
Jaakko Hollmén[2] and Matej Orešič[1,*]

[1]VTT Biotechnology, PO Box 1500, Espoo, FIN-02044 VTT, Finland and
[2]Helsinki University of Technology, Laboratory of Computer and Information Science,
PO Box 5400, Espoo, FIN-02015 HUT, Finland

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Integration of heterogeneous data in life sciences is a growing and recognized challenge. The problem is not only to enable the study of such data within the context of a biological question but also more fundamentally, how to represent the available knowledge and make it accessible for mining.

Results: Our integration approach is based on the premise that relationships between biological entities can be represented as a complex network. The context dependency is achieved by a judicious use of distance measures on these networks. The biological entities and the distances between them are mapped for the purpose of visualization into the lower dimensional space using the Sammon's mapping. The system implementation is based on a multi-tier architecture using a native XML database and a software tool for querying and visualizing complex biological networks. The functionality of our system is demonstrated with two examples: (1) A multiple pathway retrieval, in which, given a pathway name, the system finds all the relationships related to the query by checking available metabolic pathway, transcriptional, signaling, protein–protein interaction and ontology annotation resources and (2) A protein neighborhood search, in which given a protein name, the system finds all its connected entities within a specified depth. These two examples show that our system is able to conceptually traverse different databases to produce testable hypotheses and lead towards answers to complex biological questions.

Contact: matej.oresic@vtt.fi

1 INTRODUCTION

Historically, the decomposition of biology into different disciplines was necessary to tackle the complexity of life science systems by 'reducing' the degree of complexity down to the most basic level. With the advent of 'omics' revolution and systems biology, such separation of biology is becoming artificial (Blagosklonny and Pardee, 2002). In order to utilize the diverse life science knowledge, one first needs to address several practical and fundamental challenges of data integration. For example, different domain-specific naming conventions and vocabularies have been utilized both at the low level, such as genes and proteins, and the more complex entities, such as biological concepts. In order to be able to integrate data, one should therefore enable traversing across such diverse sources of information in an automated way.

From the early days of bioinformatics, several approaches for biological data integration have been developed. Well-known approaches include rule-based links, such as SRS (Etzold and Argos, 1993; Etzold *et al*., 1996), federated middleware frameworks, such as Kleisli system (Davidson *et al*., 1997; Chung and Wong, 1999), as well as wrapper-based solution using query optimization, such as IBM Discovery Link (Hass *et al*., 2001). In parallel, progress has been made to organize biological knowledge in a conceptual way by developing ontologies and domain-specific vocabularies (Ashburner *et al*., 2000; Bard and Rhee, 2004; Bodenreider, 2004). With the emergence of XML and Semantic Web technologies, the ontology-based approach to life science data integration has become more ostensible. In this context, data integration comprises problems like homogenizing the data model with schema integration, combining multiple database queries and answers, transforming and integrating the latter to construct knowledge based on underlying knowledge representation.

However, the ontology-based approach alone cannot resolve the practical problem of evolving concepts in biology, and its best promise lies in specialized domains and environments where concepts and vocabularies can be well controlled (Searls, 2005; Oresic *et al*., 2005). Neither can the ontologies alone resolve the problem of context, i.e. what may appear closely related in one context, may be further apart or unrelated in another (Gärdenfors, 2000). In this paper, we present our approach to data integration and context-based mining of biological data, which is based on the premise that relationships between biological

[*]To whom correspondence should be addressed.

entities can be represented as a complex network, with nodes being either low level (e.g. genes, compounds) or more complex entities, such as concepts (cell localization, biological processes), and with edges being relationships between them, either physical interactions or more complex relationships.

The paper is organized as follows: in Section 2, we describe the practical implementation of our three-tier data integration system and the design of the Java-based tool we developed for querying the data and visualizing complex relationships. In Section 3, we demonstrate the utility of the system with two query examples: (1) an integrated pathway retrieval and (2) a protein neighborhood search. In Section 4, we discuss the design and performance of the system as well as its future developments.

2 SYSTEMS AND METHODS

2.1 System design

Our data integration and visualization system is composed of three layers in which the data constitutes the back-end layer (Fig. 1). Schema mappings, ontology definitions and conceptual learning implementations occupy the middle tier and the user interface constitutes the front-end layer. The middle tier also comprises sets of algorithms and modules that process and display results of the query. Most of our local data are represented in XML format. The data are stored using XML data management system Tamino XML server (Software AG) in a Redhat Linux Advanced Server v2.1 environment. The databases are queried using Tamino XQuery (Fiebig and Schöning, 2004) which is an implementation of XQuery language. The queries are enabled through the Tamino Java API. For storing more voluminous data, such as gene-expression data and in house produced mass spectrometry data, we use Oracle 10*g* database server (Oracle, Inc.).

2.2 Design of the network visualization tool

The megNet software is a Java-based tool which affords parallel retrieval across multiple databases, with results displayed as a network. Edge attributes contain information about types of relationships, possibly quantitative or semantic information (e.g. ‘is located in’ in case of linking a protein with a complex entity, such as cell organelle). The tool retrieves biological data from the Tamino databases using Tamino Java API and data from Oracle databases using JDBC. The user interface is implemented using Java Swing libraries, with the graphs created using Tom Sawyer Visualization Toolkit 6.0 (Tom Sawyer, Inc.). The basic layout of the user interface is divided into four parts (Fig. 2):

- query section,
- network display section,
- text area displaying information on currently selecting entity and
- distance mapping section, displaying the mapping of the distance matrix into 2D space.

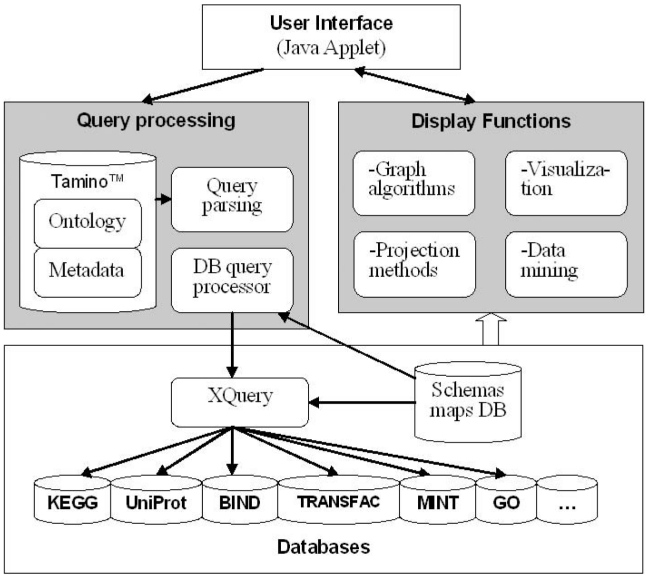


Fig. 1. Architecture of our bioinformatics data integration and visualization system.

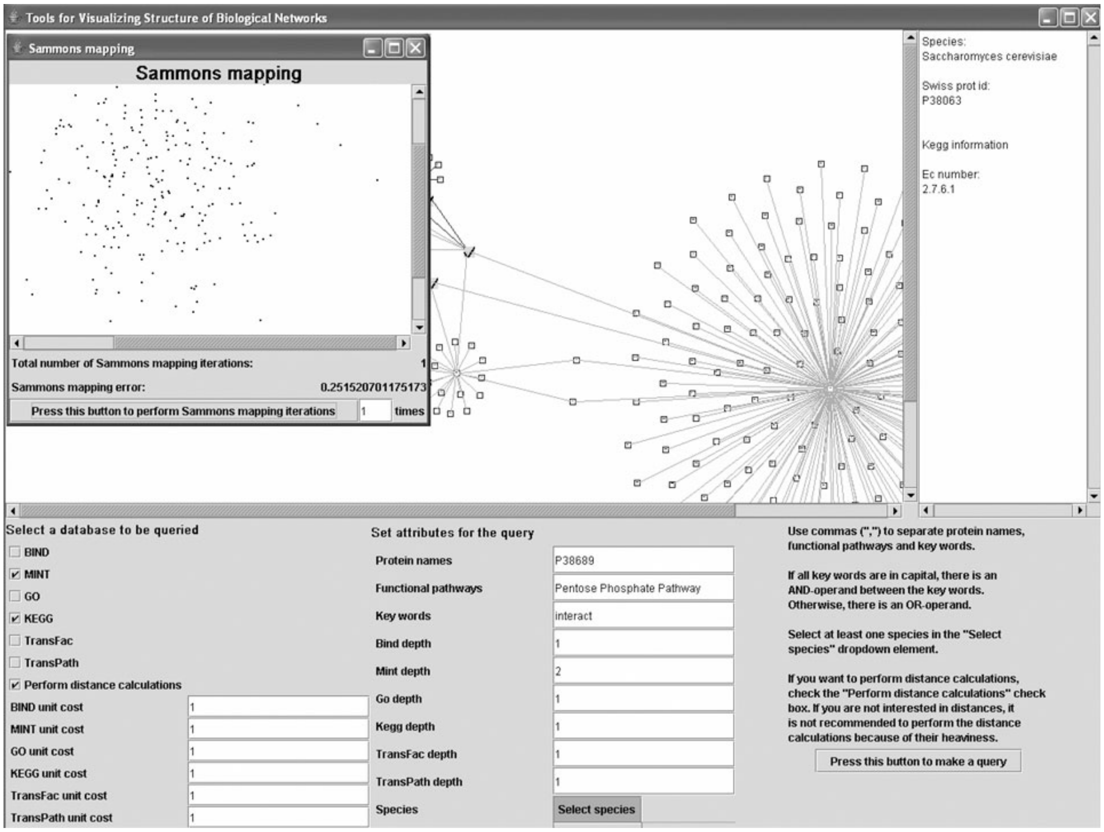
A mouse left click on a node or on an edge displays the biological information in the text area located on the right hand side. The information displayed in this text area contains the data retrieved from locally installed databases and links to external databases. The nodes can be selected to change options, such as set a new search depth for the neighbors. In the resultant graph, shape conventions are used to distinguish the type of entity underlying a node. Similarly, color codes are used to distinguish the type of relationship underlying an edge. Each node and edge shown can be checked for original source information. The resulting graph can be extracted and saved in the XML format.

2.3 Databases and data curation

Data from various public data sources were collected into our local database. Table 1 lists the data sources utilized in the examples of this paper.

In order to add a specific bioinformatics database into our system, it has to be passed first through a curation stage. A typical data curation flow is explained below in the form of a pseudoalgorithm:

(1) Decide on a data source to be set up and download the data typically using ftp. If the downloaded data are already in XML format go to step (3) otherwise go to (2).


Tools for Visualizing Structure of Biological Networks
Sammons mapping
Sammons mapping
Press this button to perform Sammons mapping iterations
times
Press this button to make a query
Select species

Total number of Sammons mapping iterations: 1

Sammons mapping error: 0.251520701175173

1 times

Species:
Saccharomyces cerevisiae

Swiss prot id:
P38063

Kegg information

Ec number:
2.7.6.1

Select a database to be queried

- □ BIND
- ☑ MINT
- □ GO
- ☑ KEGG
- □ TransFac
- □ TransPath
- ☑ Perform distance calculations

BIND unit cost	1
MINT unit cost	1
GO unit cost	1
KEGG unit cost	1
TransFac unit cost	1
TransPath unit cost	1

Set attributes for the query

Protein names	P38689
Functional pathways	Pentose Phosphate Pathway
Key words	interact
Bind depth	1
Mint depth	2
Go depth	1
Kegg depth	1
TransFac depth	1
TransPath depth	1
Species	Select species

Use commas (",") to separate protein names, functional pathways and key words.

If all key words are in capital, there is an AND-operand between the key words. Otherwise, there is an OR-operand.

Select at least one species in the "Select species" dropdown element.

If you want to perform distance calculations, check the "Perform distance calculations" check box. If you are not interested in distances, it is not recommended to perform the distance calculations because of their heaviness.

Fig. 2. Screenshot of the megNet network visualization tool. Node shapes represent their types (e.g. protein, gene), and edge colors represent types of relationships. The Sammon's mapping window displays the mapping based on specified distance metrics.

(2) Study the structure of the non-XML data and define XML schemas to capture the logical structure of the data. Go to step (4).

(3) If the document structures have been defined using DTD then convert the DTD to W3C Schema. If the XML schema is available from the source itself, if necessary, make changes to it to fit the requirements of the implementation (e.g. change the target namespace to Tamino namespace and define a prefix for the original target namespace).

(4) Define physical properties, such as indices and doctype for the logical schema to construct a Tamino Schema Definition document, i.e. TSD schema. If the previous step was (2) go to (5) or else go to (6).

(5) Develop parsers to convert the non-XML data into an XML format. A typical development phase is always followed by several test and feedback loops that involve an extensive use of XML data validation as well as human reading. Go to (7).

(6) Develop parsers to convert the distributed XML format to the required XML format.

(7) Load the resulting XML documents using mass-loading tool of the Tamino Server.

It must be noted that not every field in the source database is integrated. It is the task of the curator to capture its relevant subparts as well as to define appropriate semantics for the integrated database. Table 1 shows the XML Document Classes captured from databases used in this paper. In the course of implementing the above steps we make use of XMLSPY software (Altova, Inc.) and Tamino Schema Editor software (Software AG) for the construction and validation of logical and physical schemas, respectively. The development of parsers is usually implemented in Perl programming language and in some cases using Java.

2.4 Database traversals with schema maps

Resolving even simple biological relationships containing only a few biomolecular components often requires traversing multiple databases (Fig. 3). In order to enable such traversals within our system, we developed a database of schema maps (henceforth called maps database), which maps across different names used for the same entities across multiple databases. At the current state of development, the maps database

Table 1. Databases used in the present study

Database	Version or release date	XML document class	No. of entries
Uniprot/Swiss-Prot (Bairoch *et al*., 2005)	44.0	Uniprot	153 871
NCBI PubChem[a] (NCBI, 2004)	January 4, 2005	PC-substances	788 730
KEGG (Kanehisa *et al*., 2004)	August 2004	Pathways	11 380
LIGAND (Goto *et al*., 2002)		Gene	705 802
		Enzyme	4327
		Compound	11 116
		Glycan	10 302
TRANSFAC (Matys *et al*., 2003)	8.4	Gene	7796
		Factor	5919
		Site	14 782
TRANSPATH (Krull *et al*., 2003)	5.3	Network	72 769
Logical classes of data and entries: Pathway—333 Gene—4989 Molecule—20 164 Reaction—23 065 Annotation—24 218			
BIND (Bader *et al*., 2003)	August 27, 2004	BIND-submit	90 580
MINT (Zanzoni *et al*., 2002)	2.1	Entryset	18 951
IntAct (Hermjakob *et al*., 2004)	September 7, 2004	Entryset	37
Gene Ontology (Ashburner *et al*., 2000) assocdb XML version	January 4, 2004	GO	18 078

[a] NCBI PubChem (Accessed on January 10, 2005) http://pubchem.ncbi.nlm.nih.gov/

contains protein entities, indexed by UniProt identifiers. An example of such a map is shown in the XML code in Table 2. For creating such a map, we developed a Perl program to extract data from the Uniprot XML documents. We further extended this data with the GenInfo identifiers used in the BIND database (Bader *et al*., 2003) for each interacting protein. This data is obtained by applying the 'SeqHound-GetDefline' function of the SeqHound API (Michalickova *et al*., 2002). The HTTP method call for this 'SeqHound' function has been implemented using LWP module of the Perl programming language.

The database traversals can be achieved by applying simple join operations involving the maps database. Since the maps database records contain identifiers and names of an entity from all databases, it is ensured that the join operation between appropriate databases and rightly chosen entities would always return a non-empty result. The querying of a database independent of the names used in it can be achieved by writing queries to first search the maps database to find out the name/Id number of the entity in the original database and then search the original database with the correct name/Id number. Considerable challenge for any biological data integration is the often-changing structures of the data in the public databanks (Critchlow *et al*., 2000). We address this problem at the 'Logical schema construction level' of our data curation cycle by keeping our logical schemas to be as minimal as possible, yet useful enough to be able to observe the associations between all the data sources.

2.5 Similarity measures and graph projection

Property of similarity plays an essential role in human perception and formation of new concepts. The problem of evaluating similarity (or inversely, distance) between two entities or concepts appears more difficult when considering several 'quality dimensions' (Gärdenfors, 2000). In the domain of biology, the 'quality dimensions' could mean relationships of different types, i.e. chemical reactions, protein–protein interactions, gene sequence comparison or more complex relationships like protein localization, gene–phenotype association or compound properties.

Although distances within the molecular networks can be intuitively set to the length of the shortest path between the molecules, distance measure is less obvious for relationships, such as in ontologies. It was shown that Gene Ontology (GO) could be represented as a graph, and the distance measures in such a case were already studied (Lee *et al*., 2004). For the ontology trees, we assign a distance based on the closest common ancestor in the graph. When combining multiple relationships and corresponding distance measures, reasonable normalization of distance values has to be set in order to be able to compare across heterogeneous data sources. The distances between entities that do not have a direct relationship are then calculated as the

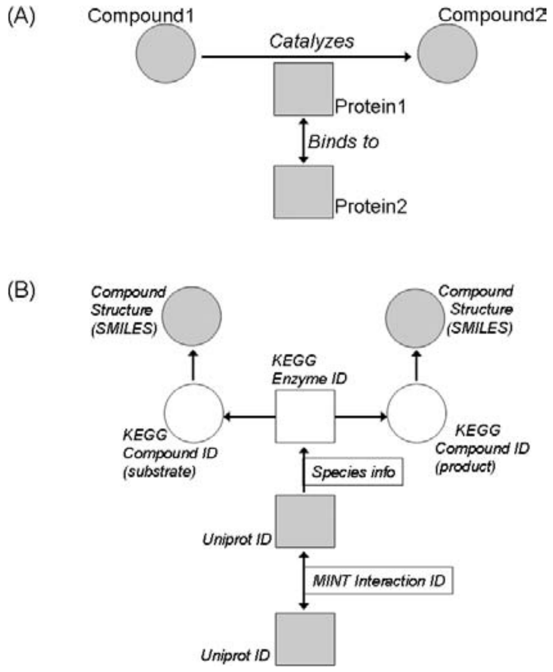


Fig. 3. (**A**) Schematic representation of relationships between two compounds and two proteins. (**B**) Same representation as hypothetically resolved via traversals across multiple databases.

Table 2. XML document from maps database for Uniprot protein entry AG35_VACCV, with links to indices from databases, such as EMBL, PIR, INTERPRO and Pfam

```
<?xml version="1.0" encoding="utf-8"?>
<protein created="1988-04-01" dataset="Swiss-Prot" ino:id="3426"
updated="2004-07-05">
  <primaryid>P07242</primaryid>
  <entry>AG35_VACCV</entry>
  <name>Envelope protein</name>
  <synonym>Protein H5</synonym>
  <synonym>Protein H6</synonym>
  <organism>
    <name>Vaccinia virus (strain WR)</name>
    <dbref id="10254" type="NCBI Taxonomy"/>
  </organism>
  <gene>
    <name>AG35</name>
    <synonym>H5R</synonym>
  <dbref id="M13209" type="EMBL">
    <property type="protein sequence ID"
    value="AAB59841.1"/>
  </dbref>
  <dbref id="M23648" type="EMBL">
  <property type="protein sequence ID"
    value="AAA47962.1"/>
  </dbref>
</gene>
<dblinks>
<dbref id="F24481" type="PIR">
  <property type="entry name" value="QQVZH6"/>
</dbref>
  <dbref id="IPR004966" type="InterPro">
  <property type="entry name" value="Pox_Ag35"/>
</dbref>
<dbref id="PF03286" type="Pfam">
  <property type="entry name" value="Pox_Ag35"/>
</dbref>
  <dbref id="138380" type="GenInfo"/>
</dblinks>
</protein>
```

lengths of the shortest paths with the distance-weighted edges (Fig. 4). The normalization of distances for each new data source is, in practice, handled by the bioinformaticians performing data curation. This assures that the system users do not need to know the specifics of the underlying data representation.

After distance normalization, it is ultimately up to the user to assign importance and therefore distance bias to any particular relationship type, by which context sensitivity can be achieved (Gärdenfors, 2000), as illustrated in Figure 4. When visualizing such complex data, we often need to project them into a lower dimensional space. In doing so it is important to preserve distances, i.e. two samples that are close to each other in the original space have to stay close when projected, or vice versa, two entities that are close to each other in the projected space must have come from the samples that were close to each other in the original space. It is the idea behind Sammon's mapping (Sammon Jr, 1969), which is implemented in our visualization tool. Visual configuration of entities is estimated with a gradient descent type of algorithm on a cost function based on the interpoint distances between the entities in the original space and the introduced discrepancies when applying the dimensionality-reducing mapping. In this way, the visual configuration approximates the original relationships in the complex networks. This kind of distance preservation is also used in the Kohonen's self-organizing maps (Kohonen, 2001) and multi-dimensional scaling (Torgerson, 1952).

3 EXAMPLES

3.1 Integrated pathway retrieval

Metabolic pathways and protein interaction networks have been studied extensively in the context of topology and modularity (Jeong *et al.*, 2000, 2001). When attempting to model real biological phenomena, it is becoming clear that one needs to understand the cross-talk across different levels of biological organization, for example, between metabolic pathways and cell signaling (Papin and Palsson, 2004).

One of the primary motivations for the development of our bioinformatics system was the need to facilitate the study of available information in the context of biological questions.

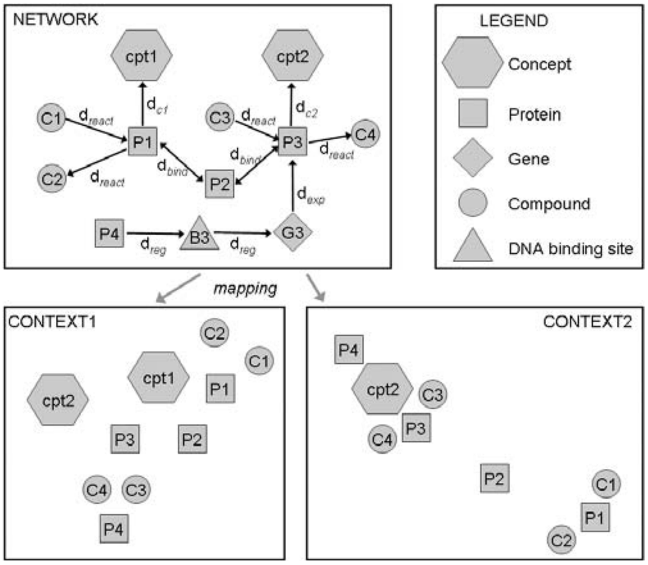


Fig. 4. Illustrative example of using graph projection in exploratory analysis of biological networks. In CONTEXT1 we are weighting all types of relationships similarly, so the nodes are clustered based on shortest path length between the edges. In CONTEXT2, we are interested only in concept cpt2, and assign lower distance value to nearest neighbors in metabolic pathways compared with other interactions.

One such application is the study of metabolic pathways, enriched with information about known molecular interactions at the level of protein–protein interactions, regulatory and signaling networks. As an example, we created the following query: ‘Glycolysis/Gluconeogenesis AND Pentose phosphate pathway AND TCA cycle IN *S.cerevisiae*’. The query was set up to first search the KEGG and retrieve the primary components of the pathways, i.e. enzymes and compounds. The database traversals were then used to search protein–protein interaction databases BIND and MINT for interactions of the enzymes with the nearest neighbor proteins (i.e. interaction search depth was set to 1). The resulting networks show surprisingly high level of connectivity across different stages of linear metabolic pathways via protein–protein interactions (Fig. 5). Specifically, in the zoomed-in region of Figure 5, we focus on two enzymes from the glycolysis pathway: phosphoglycerate kinase (PGK; EC 2.7.2.3) and acetate-CoA ligase (ACS; EC 6.2.1.1). ACS catalyzes formation of acetyl-CoA from acetate, which is a starting point in the TCA cycle, while PGK catalyzes acetylation of 3-phospho-D-glycerate, which is a part of the second phase of glycolysis. Both enzymes appear to aggregate with SRB2, based on the evidence from the yeast two-hybrid pooling approach (Ito *et al*., 2001). Notably, SRB2 is involved in transcriptional initiation (Thompson *et al*., 1993). This could mean that PGK and ACS, enzymes at two different stages of glycolysis, are coregulated. While the evidence from high-throughput yeast two-hybrid assays needs to be taken with caution due to possibly high number of false positive aggregation hits (Mrowka *et al*., 2001), our results do point toward a testable hypothesis for the future research.

3.2 Protein neighborhood search

Assignment of protein function is a non-trivial task owing to the fact that the same proteins may be involved in different biological processes, depending on the state of the biological system and protein localization (Camon *et al*., 2004). Therefore, protein function is context dependent.

The ‘protein neighborhood’, i.e. the entities of the network close to the protein, mode provide an insight about the protein function and its mode of action. The entities in our case can be molecules, genes or more complex concepts, and the proximity is measured by applying the distance measure. As an example, we searched the neighborhood of mannose-6-phosphate isomerase for *Saccharomyces cerevisiae* (PMI40; UniProt Id: P29952), which catalyzes the conversion between fructose 6-phosphate and mannose 6-phosphate and thus connects glycolysis with the cell wall synthesis in *S.cerevisiae* (Smith *et al*., 1992). The search involved concurrent retrieval of relationships for the following databases: UniProt, KEGG, BIND, MINT and GO Biological Process. For any nearest neighbor protein–protein association, such as protein–protein interaction or sharing the same GO class at the lowest level, the distance was set to 1. In the case of metabolic pathways, weight of each edge was set to 0.5 in the direction of possible reaction. The search depth was set to two nearest proteins if the first of the edges was a protein–protein interaction, and to the nearest protein otherwise. This included cases where the nearest protein was connected to the search protein via the compound in metabolic pathways or the lowest level GO term. Figure 6 shows the resulting graphs and Sammon’s mapping of the nearest protein neighbors of PMI40.

The zoomed-in window shows one region of potential interest, which includes protein–protein interactions between the PMI40 and NUP100 (UniProt Id: Q02629), a subunit of the nuclear pore complex, as well as between alpha-1,6-mannosyltransferase (MNN10; UniProt Id: P50108) and NUP100. According to GO (GO:0000032), both PMI40 and MNN10 are also involved in cell wall mannoprotein synthesis. While PMI40 is a ‘gate’ between cell wall synthesis and glycolysis, i.e. cell decision point between growth or energy production, MNN10 is a part of the protein complex in mannoprotein synthesis toward the end of the cell wall biosynthesis pathways. Examination of interaction entries (BIND Ids 137 955 and 137 823) suggests that NUP100 protein, which is a part of nuclear pore complex, binds to the PMI40 and MNN10 open reading frames (Casolari *et al*., 2004). This and other evidence by Casolari *et al*. provide support for the

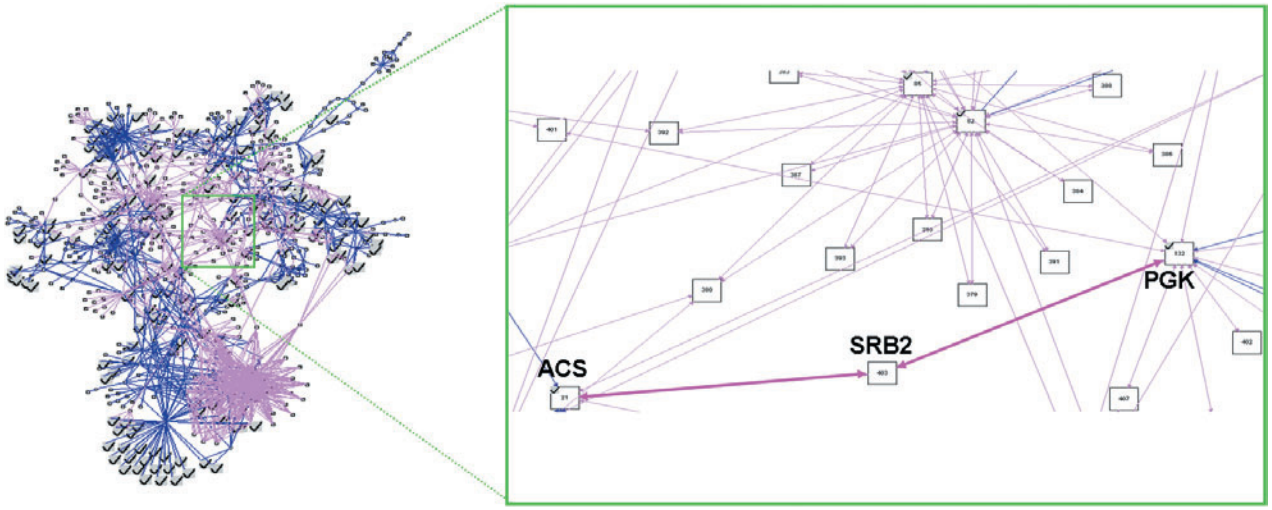


Fig. 5. Integrated pathway retrieval using megNet network visualization tool, with the query for ‘Glycolysis/Gluconeogenesis AND Pentose phosphate pathway AND TCA cycle IN *S.cerevisiae*’. Metabolic pathways are shown with blue edges, protein–protein interactions with pink. Proteins are represented with squares, compounds with circles. Surprisingly, high level of connectivity via protein–protein interactions is found across different modules of the metabolism. The zoomed-in region shows a specific connection between Acetate-CoA ligase (ACS) and Phosphoglycerate kinase (PGK) via interactions with SRB2, which is known to be involved in transcriptional initiation. The interactions discussed are highlighted for clarity.

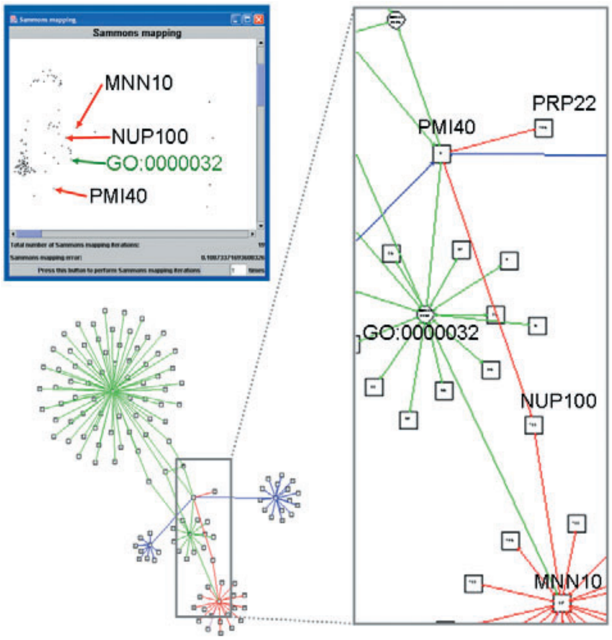


Fig. 6. Network neighborhood of mannose-6-phosphate isomerase (PMI40) in *S.cerevisiae*. Metabolic pathway relationships are shown in blue, protein–protein interactions in red, and GO associations in green. Both PMI40 and MNN10 are involved in cell wall mannoprotein synthesis (GO:0000032). NUP100 protein, which is part of the nuclear pore complex, appears to interact with the PMI40 and MNN10 genes.

‘gene-gating’ hypothesis, which suggests that the interaction of the nuclear pore complex with different genes might serve as a level of gene regulation (Blobel, 1985). It remains to be tested whether PMI40 and MNN10 are indeed coregulated in relation to cell decision-making between energy production versus growth.

4 DISCUSSION

Our integration approach is based on the premise that relationships between biological entities can be represented as a complex network. The information in such networks forms a basis for exploratory mining. Distances between different nodes in an integrated network play a central role in our framework. In order to calculate distances, one first needs to define distance measures across heterogeneous types of information. We are taking a pragmatic approach by letting the user define the distances as a part of the query. This is reasonable since the distance basically defines the context of the questions posed by the user and allows biasing the similarity toward particular types of relationships, or toward relationships in a specific context. Once the distance measure is specified, we can map the nodes of the graph into a lower dimensional space. As the mapping is approximate, there will be some distortion while doing the mapping. Therefore, in our opinion the exact form of distance measure is not a critical issue, so long as it underlines the relationships in the concept graph. In fact, selection of distance measure may reflect a subjective choice and as such will be subject to debate. It is ultimately the end result of mining that determines the utility of specific distance measure.

Presently, we are using Sammon's mapping for that purpose, which maps the graph non-linearly into lower dimensional space while preserving the internode distances across the network. One disadvantage of Sammon's mapping is that addition of the nodes requires new computation of the mapping on the complete network, and is therefore not well suited for interactive addition of new nodes. Other mappings, such as other types of multidimensional scaling methods (Torgerson, 1952) or self organizing maps (Kohonen, 2001), are also considered for future implementations. In particular, we will investigate the non-metric multidimensional scaling method (Cox and Cox, 2001), which is focused on preserving the order of similarities.

The two illustrative examples shown in the paper provide evidence for the usefulness of our approach. In the case of integrated pathway retrieval, we found large level of interconnectivity across different stages and modules of the metabolic pathways via protein–protein interactions, which raises questions about merit of studying the topology of metabolic networks outside the scope of other biological networks. Specifically, we found evidence of possible coregulation of enzymes at early and late stages of glycolysis pathway, which needs to be further investigated experimentally. In the case of protein neighborhood search, we were able to retrieve relationships and potential mechanisms that would not have been easily found through browsing databases separately. We believe our protein neighborhood search is a powerful tool for visual protein annotation in a context dependent manner.

Our approach is not limited to pathway databases and ontologies alone. We are currently extending the system in two directions. First, we aim at complementing the knowledge extracted from structured and semistructured data with the knowledge extracted from literature. Currently, we are implementing a text mining tool to retrieve from literature relationships between entities of interest, with primary focus on biomedical domain (Oresic *et al*., 2005). The discovered relationships will be, similarly as described in this paper, represented as a network. Second, genome information and experimental data such as metabolic profiles or gene-expression data can also be included. The distance measures in such cases are related to the level of association (e.g. correlation coefficient) or in the case of gene sequence comparison, to the alignment score. Combining molecular profile data with ontology information using database traversals has already been attempted (Oresic *et al*., 2004), but without the distance calculations.

We have presented an integrated database and software system that enables retrieval and visualization of biological relationships across heterogeneous data sources. We have demonstrated its merit on two practical examples: protein neighborhood search and integrated pathway retrieval. Owing to light-weight design of the system, it is relatively easy to incorporate new types of information and relationships. We believe our approach facilitates discovery of novel or unexpected relationships, formulation of new hypotheses, design of experiments, data annotation, interpretation of new experimental data, and construction and validation of new network-based models of biological systems.

ACKNOWLEDGEMENTS

M.O. was in part funded by Marie Curie International Reintegration Grant. M.O. and J.H. were in part funded by the Academy of Finland SYSBIO Programme.

REFERENCES

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S. and Eppig,J. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bader,G.D., Betel,D. and Hogue,C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.

Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bard,J.B.L. and Rhee,S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.*, **5**, 213–222.

Blagosklonny,M.V. and Pardee,A.B. (2002) Conceptual biology: unearthing the gems. *Nature*, **416**, 373.

Blobel,G. (1985) Gene gating: a hypothesis. *Proc. Natl Acad. Sci. USA*, **82**, 8527–8529.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.

Casolari,J.M., Brown,C.R., Komili,S., West,J., Hieronymus,H. and Silver,P.A. (2004) Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell*, **117**, 427–439.

Chung,S.Y. and Wong,L. (1999) Kleisli: a new tool for data integration in biology. *Trends Biotechnol.*, **17**, 351–355.

Cox,T.F. and Cox,M.A.A. (2001) *Multidimensional Scaling*, Chapman and Hall/CRC, Boca Raton.

Critchlow,T., Fidelis,K., Ganesh,M., Musick,R. and Slezak,T. (2000) DataFoundry: information management for scientific data. *IEEE Trans. Inf. Technol. Biomed.*, **4**, 52–57.

Davidson,S.B., Overton,C.G., Tannen,V. and Wong,L. (1997) BioKleisli: a digital library for biomedical researchers. *Int. J. on Digital Libraries*, **1**, 36–53.

Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *CABIOS*, **9**, 49–57.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods enzymol.*, 114–128.

Fiebig,T. and Schöning,H. (2004) Software AG's Tamino XQuery Processor. *XIME-P 2004*, 19–24.

Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.

Gärdenfors,P. (2000) *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, MA.

Hass,L.M., Schwartz,P.M. and Kodali,P. (2001) DiscoveryLink: a system for integrated access to life science data sources. *IBM Systems Journal*, **40**, 489–511.

Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jeong,H., Mason,S.P., Barabási,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jeong,H., Tombor,B., Albert, R., Oltvai,Z.N. and Barabási,A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kohonen,T. (2001) *Self Organizing Maps*, Springer Verlag.

Krull,M., Voss,N., Choi,C., Pistor,S., Potapov,A. and Wingender,E. (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.*, **31**, 97–100.

Lee,S.G., Hur,J.U. and Kim,Y.S. (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381–388.

Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Michalickova,K., Bader,G., Dumontier,M., Lieu,H., Betel,D., Isserlin,R. and Hogue,C. (2002) SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics*, **3**, 32.

Mrowka,R., Patzak,A. and Herzel,H. (2001) Is there a bias in proteome research? *Genome Res.*, **11**, 1971–1973.

Oresic,M., Clish,C.B., Davidov,E.J., Verheij,E., Vogels,J.T.W.E., Havekes,L.M., Neumann,E., Adourian,A., Naylor,S., Greef,J.V.D. *et al*. (2004) Phenotype characterization using integrated gene transcript, protein and metabolite profiling. *Appl. Bioinformatics*, **3**, 205–217.

Oresic,M., Gopalacharyulu,P.V., Lindfors,E., Bounsaythip,C., Karanta,I., Hiirsalmi,M., Seitsonen,L. and Silvonen,P. (2005) Towards an integrative and context sensitive approach to *in silico* disease modelling. *ERCIM News*, 25–26.

Papin,J.A. and Palsson,B.O. (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.*, **227**, 283–297.

Sammon,J.W.Jr. (1969) A nonlinear mapping for data structure analysis. *IEEE Trans. Comp.*, **C-18**, 401–409.

Searls,D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Disc.*, **4**, 45–48.

Smith,D., Proudfoot,A., Friedli,L., Klig,L., Paravicini,G. and Payton,M. (1992) PMI40, an intron-containing gene required for early steps in yeast mannosylation. *Mol. Cell. Biol.*, **12**, 2924–2930.

Thompson,C.M., Koleske,A.J., Chao,D.M. and Young,R.A. (1993) A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell*, **73**, 1361–1375.

Torgerson,W.S. (1952) Multidimensional scaling: I. theory and method. *Psychometrika*, **17**, 401–419.

Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a molecular interaction database. *FEBS Lett.*, **513**, 135–140.