Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

A Visual cross-database Comparison of Metabolic Networks

M. Rohrschneider *G.Scheuermann* and Peter F. Stadler[1]

[1]Leipzig University, Department of Computer Science, Germany

Abstract

Bioinformatics research in general and the exploration of metabolic networks in particular rely on processing data from different sources. Visualization in this context supports the exploration process and helps to evaluate the data quality of the used sources.
In this work, we extend our existing metabolic network visualization toolbox and hereby address the fundamental task of comparing metabolic networks from two major bioinformatics resources for the purpose of data validation and verification. This is done on different levels of granularity by providing an overview on retrieval rates of chemical compounds and reactions per pathway on the one hand, as well as giving a detailed insight into the differences in the biochemical reaction networks on the other.
We reconstructed different subsets of the metabolism stored at the KEGG Pathway database and compare these networks against the complete metabolic network provided by the MetaCyc branch of the BioCyc database collection. Matches among the sets of chemical compounds and reactions are highlighted and propagated to higher levels of abstraction to infere pathway correspondence between the two resources.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

In recent years, the research on genomics and proteomics – in short 'omics' – has developed to one of the largest growing branches in the field of bioinformatics. The reconstruction and analysis of metabolic networks depends greatly on the findings by those disciplines. As a result, the correctness of the used biological data resources is essential and often assumed as given. The data accessible from major databases, e.g. KEGG Pathway, BioCYC, ExPASy, and others,
FIXME: citations needed
is either manually curated or computationally derived from sets of enzymes with known functionality that have been detected by microarray experiments.

FIXME: das ist hier vlt etwas schwammig, bin da kein experte. geht das trotzdem als einleitender teil?

Both processes are subject to errors. The motivation of this work is to expose and detect potential weaknesses and errors in the representation of biochemical networks. We hereby consider the metabolic network data provided by the KEGG Pathway database. It contains a set of manually drawn metabolic pathway diagrams presented as semi-static visualizations used for navigating the data on line as well as XML-like descriptions of those pathways.

In general, a metabolic network can be considered as a set of possibly overlapping metabolic pathways, which in are defined by a distinct set of chemical reactions transforming chemical compounds. Network-like structures arise by linking a reaction together that produces a certain compound, that in turn is consumed by another reaction. In terms of graph theory, the network can be considered as a bipartite graph with the chemical compounds being one class of nodes and the reactions as the other node class. The network does not only contain chains as the name pathway suggests, as reaction sequences can branch to form tree-like structures, or cycles. As for reaction nodes, the neighborhood, i.e. the set of adjacent nodes is always unique for a specific reaction. The set of reactions defining a pathway is somewhat arbitrary and is based on expert knowledge. While metabolic pathways presented by KEGG Pathway are rather large functional units, the authors in [GK06] propose an ontology consisting of considerably smaller units.

The focus of this work is the multi-scale comparison of

metabolic networks provided by the two aforementioned databases. Firstly, we describe how compounds and reactions are matched between the two networks and how this information is used to infere relationships between pathways defined by KEGG vs. the ontology defined in BioCYC. The results can be viewed by the user from a global point of view, i.e. a quantification of the node matching quality for each pathway of the KEGG network, and in more detail by interactively expanding the pathways of interest to reveal the respective reaction networks. For each reaction and compound node of the KEGG Pathway network, the type of match with a node of the BioCYC network is displayed. This is either an exact (unique) match (exactly one hit in BioCYC), ambiguous match (more than one hit), or ”no hits” at all. We combine several popular information visualization methods to navigate the presented network, such as semantic zoom, hierarchical exploration by node expansion, and focus & context techniques. Once the user has identified network points of interest based on the highlighted differences, we provide the context of the matched reaction or compound in the BioCYC graph as overlay on the current KEGG network. Additionally, the respective subset of BioCYC's pathway ontology can be viewed on demand for every reaction, compound, and pathway node. The implementation of the exploration process is inspired by Ben Shniderman's mantra of visual information-seeking [Shn96].

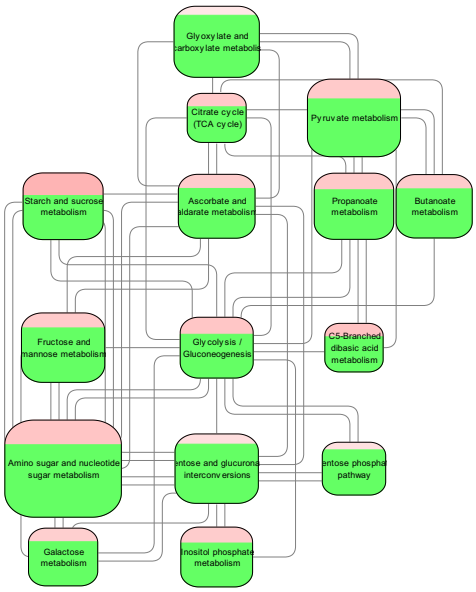

Figure 1: *Overview on the metabolic network constructed from KEGG Pathway. 15 Pathways associated with the carbohydrate metabolism are shown. The node size depicts the number of reactions and compounds of the respective pathway.The nodes' filling level reflects the 'match score', which is closely related to the ratio of matched nodes and total node number, but also penalizes ambiguous matches. The saturation of the red color hints to the relative number of nodes that could not be matched at all.*

2. Related Works

[CFF*06] [CAD*10] [KZM*04] [GK06] [KPR02] [RKS11] [AKK*09] [FHK*09] [SDMW09] [KS07] [Sch03] [RHR*09] [SOR*11] [AEBG*08] [Bar95] [TB00] [DZdS08] [WEK01]

3. Data Resources, Preprocessing, and Data Structure

KGML is an exchange format for the KEGG pathway maps that are manually drawn and updated. It enables automatic drawing of KEGG pathways and provides facilities for computational analysis and modeling of protein and chemical networks.

4. Graph Layout

5. Network Comparison

5.1. Graph Matching

KEGG graph as template
matching process
compounds of KEGG against BioCyc compounds: given two sets of synonyms describing the chemical compound
reactions: 2-step-process
(1) pattern matching using local topology of unified reaction network graph
robust and reveals inconsistencies of reaction definitions
(2) for multiple hits: use ec nomenclature of reaction to select the correct match

[DZdS08] Comparison of metabolic reactions has been mostly based on Enzyme Commission (EC) numbers, which are extremely useful and widespread, but not always straightforward to apply, and often problematic when an enzyme catalyzes several reactions, when the same reaction is catalyzed by different enzymes, when official full EC numbers are unavailable or when reactions are not catalyzed by enzymes

5.2. Visual Representation and Exploration

node coloring to identify match types
(compound matches, primary and secondary reaction matches)
match score summary visualized in parent (pathway) nodes: saturation and filling level

verification by displaying context information of biocyc graph
display pathway ontology of pathways associated with se-

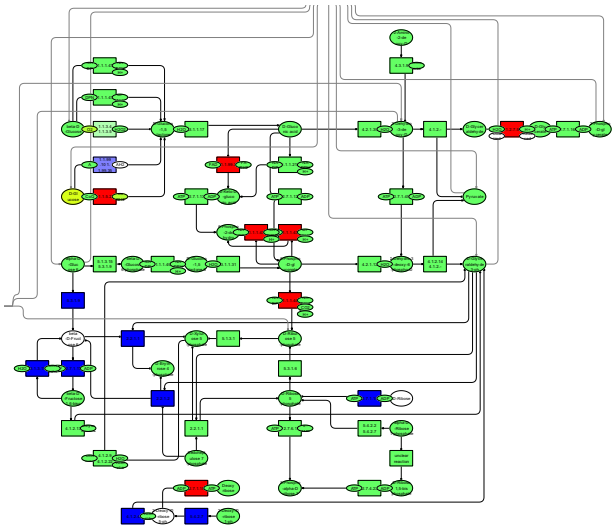

Figure 2: *Match results on the detailed reaction network of the Pentose-Phosphate-Pathway. A color on the scale between yellow and green depicts the local match score, i.e. the inverse of the number of hits in the BioCyc graph. A white node color for compounds and red for reactions indicates, that the compound or reaction could not be found in the BioCyc graph. A reaction node drawn in blue indicates, that only a subset of adjacent compound nodes could be found in the BioCyc graph. In those cases, the reaction matches are much less reliable, but could very often veryfied using the EC number of the associated enzyme.*

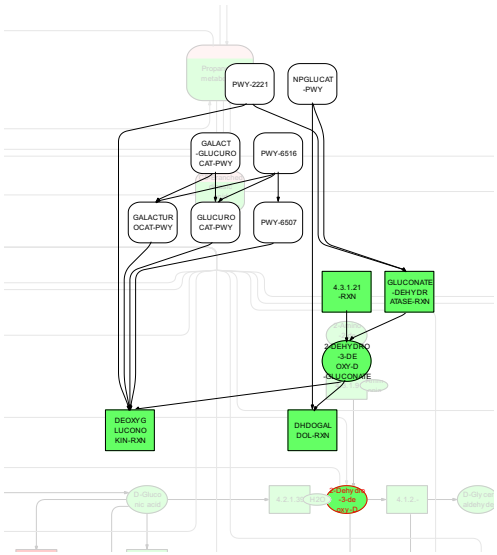

Figure 3: *For the compound 2-Dehydro-3-deoxy-D-gluconate of the PPW, the neighborhood (green reaction nodes) in the BioCyc graph together with the associated pathway ontology (white, rounded rectangles) are shown.*

lected reaction or compound node
for kegg pathways: display pathway ontology associated with matched reactions contained in that pathway.

6. Results

table of matching scores for kegg pathways associated with carbohydrate metabolism
details: false positive reaction matchings (mismatch)
ambiguous matches
no hits for reactions, verified by presenting the corresponding biocyc context

7. Conclusion and Future Work

Vorverarbeitung Framework
Interaktions- und Vis-Techniken, um Unterschiede zwischen Netzwerken zu erkennen. Manche Unterschiede sind erklaerbar, andere nicht plausibel. Dies ist vom Anwender zu beurteilen und setzt biochemisches Vorwissen voraus.
informierte Datenbereinigung - use graph editor to manually curate or manipulate metabolic network data. This can be

References

[AEBG*08] ALBRECHT M., ESTRELLA-BALDERRAMA A., GEYER M., GUTWENGER C., KLEIN K., KOHLBACHER O., SCHULZ M.: 08191 working group summary – visually comparing a set of graphs. In *Graph Drawing with Applications to Bioinformatics and Social Sciences* (Dagstuhl, Germany, 2008), Borgatti S. P., Kobourov S., Kohlbacher O., Mutzel P., (Eds.), no. 08191 in Dagstuhl Seminar Proceedings, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[AKK*09] ALBRECHT M., KERREN A., KLEIN K., KOHLBACHER O., MUTZEL P., PAUL W., SCHREIBER F., WYBROW M.: On open problems in biological network visualization. In *Graph Drawing* (2009), pp. 256–267.

[Bar95] BARRETT A. J.: Enzyme nomenclature. recommendations 1992. *European Journal of Biochemistry 232*, 1 (1995), 1–1. doi:10.1111/j.1432-1033.1995.tb20774.x.

[CAD*10] CASPI R., ALTMAN T., DALE J., DREHER K., FULCHER C., GILHAM F., KAIPA P., KARTHIKEYAN A., KOTHARI A., KRUMMENACKER M., LATENDRESSE M., MUELLER L., PALEY S., POPESCU L., PUJAR A., SHEARER A., ZHANG P., KARP P.: Metacyc: The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research 38* (2010), D473–D479. doi:10.1093/nar/gkj128.

[CFF*06] CASPI R., FOERSTER H., FULCHER C. A., HOPKINSON R., INGRAHAM J., KAIPA P., KRUMMENACKER M., PALEY S., PICK J., RHEE S. Y., TISSIER C., ZHANG P., KARP P. D.: Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research 34*, Database issue (2006), D511–D516. doi:10.1093/nar/gkj128.

[DZdS08] DIOGO A. L., ZHANG Q.-Y., DE SOUSA J. A.: Genome-scale classification of metabolic reactions and assignment of ec numbers with self-organizing maps. *Bioinformatics 24*, 19 (2008), 2236–2244.

[FHK*09] FUNG D. C. Y., HONG S.-H., KOSCHÜTZKI D., SCHREIBER F., XU K.: Visual analysis of overlapping biological networks. In *IV* (2009), pp. 337–342.

[GK06] GREEN M. L., KARP P. D.: The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research 34*, 13 (2006), 3687–3697. doi:10.1093/nar/gkl438.

[KPR02] KARP P. D., PALEY S., ROMERO P.: The pathway tools software. *Bioinformatics 18*, 1 (2002), S225–S232.

[KS07] KLUKAS C., SCHREIBER F.: Dynamic exploration and editing of kegg pathway diagrams. *Bioinformatics 23*, 3 (2007), 344–350.

[KZM*04] KRIEGER C. J., ZHANG P., MUELLER L. A., WANG A., PALEY S., ARNAUD M., PICK J., RHEE S. Y., KARP P. D.: Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research 32*, Database issue (2004), D438–D442. doi:10.1093/nar/gkh100.

[RHR*09] ROHRSCHNEIDER M., HEINE C., REICHENBACH A., KERREN A., SCHEUERMANN G.: A novel grid-based visualization approach for metabolic networks with advanced focus&context view. In *Graph Drawing* (2009), pp. 268–279.

[RKS11] ROHN H., KLUKAS C., SCHREIBER F.: Visual analytics of multimodal biological data. In *IMAGAPP/IVAPP* (2011), pp. 256–261.

[Sch03] SCHREIBER F.: Visual comparison of metabolic pathways. *J. Vis. Lang. Comput. 14*, 4 (2003), 327–340.

[SDMW09] SCHREIBER F., DWYER T., MARRIOTT K., WYBROW M.: A generic algorithm for layout of biological networks. *BMC Bioinformatics 10* (2009), 375.

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (1996), pp. 336–343.

[SOR*11] SMOOT M. E., ONO K., RUSCHEINSKI J., WANG P., IDEKER T.: Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics 27*, 3 (2011), 431–432.

[TB00] TIPTON K., BOYCE S.: History of the enzyme nomenclature system. *Bioinformatics 16*, 1 (2000), 34–40.

[WEK01] WIESE R., EIGLSPERGER M., KAUFMANN M.: yfiles: Visualization and automatic layout of graphs. In *Proceedings of the 9th International Symposium on Graph Drawing (GD 2001)* (2001), Springer-Verlag.