Visualizing the Results of Metabolic Pathway Queries

Allison P. Heath[1], George N. Bennett[2], and Lydia E. Kavraki[1,3,4]

[1]Department of Computer Science, [2]Department of Biochemistry and Cell Biology,
[3]Department of Bioengineering, Rice University, Houston, TX 77005, USA
[4]Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine,
Houston, TX, 77005, USA

1 Introduction and Problem Definition

Biology contains a wealth of network data, such as metabolic, transcription, signaling and protein-protein interaction networks. Our research currently focuses on metabolic networks, although similar ideas may be applied to other biological networks. Metabolic networks consist of the chemical compounds and reactions necessary to support life. Traditionally, series of successive metabolic reactions have been organized into simple metabolic pathways and manually drawn. However, as we move into the era of systems biology, it is becoming apparent that automated ways of processing and visualizing metabolic networks must be developed.

Our main goal is to create helpful visualizations of a large number of small metabolic networks or paths for biological researchers. This is a distinct problem from previous work on visualizing large metabolic networks and single pathways [1,2,3,4,5]. As a concrete example, a common query is to find all of the paths between two chemicals in the network. Using a methodology we developed, we find 27,912 paths of length 13 from L-2-aminoadipate to L-lysine in KEGG. The number of paths quickly scales with length; there are 693,943 paths of length 15 between the same two compounds. Displaying a list of all of these pathways or merging all of these pathways together produces an unsatisfactory visualization.

2 Approach and Results

While merging all of the pathways together produces a poor visualization for biological researchers, it does reveal that the main variation between the pathways is the reactions, not the compounds. Therefore, we investigated clustering the results. We define a distance measure based on the similarity of the chemical compounds in the pathways: $\frac{|c(X) \oplus c(Y)|}{|c(X)|+|c(Y)|}$, where $c(X)$ and $c(Y)$ are the set of chemical compounds in path X and Y. Using this distance measure, we cluster using a simple leader algorithm. This algorithm builds a list of the paths in random order, then selects the first path and designates it a cluster center. It then iterates over the remaining paths. If the next path is less than a predetermined maximum distance from a cluster center, it is added to the nearest cluster. Otherwise, it is designated as a new cluster center. This algorithm is fast and does not require knowledge of the number of clusters. While it is a localized algorithm, it can be run multiple times and the clusters can be compared to see how consistent they are.

I.G. Tollis and M. Patrignani (Eds.): GD 2008, LNCS 5417, pp. 441–442, 2009.
© Springer-Verlag Berlin Heidelberg 2009

For our example data consisting of 27,912 paths of length 13 from L-2-aminoadipate to L-lysine this method appears to work relatively well. At a distance cutoff of 0 we get 81 clusters ranging in size from 8 to 140 nodes. At a distance cutoff of 0.2 we get 35 clusters ranging in size from 8 to 146. Visualizing these clusters using Cytoscape produce qualitatively decent results. However, further investigation of distance measures and clustering technique will likely be needed for larger or more dissimilar result sets.

3 Discussion

We demonstrate that simple clustering methods can help reveal the structure of the data and create simpler, more useful visualizations. In addition to information obtained from clustering the results, there is a wealth of external biological information that can assist and enrich the visualization. In order to be fully useful, the display should enable the user to interact with the results. We have begun work on a Cytoscape plugin which should enable interactive features. We hope to combine these ideas together to create useful visualizations of many small metabolic networks or pathways. However, many open questions remain to be investigated on visualizing biological pathway data.

Acknowledgements

This work is supported in part by a Hamill Innovation Grant and NSF CCF 0523908. APH is supported by a NSF Graduate Research Fellowship.

References

1. Bourqui, R., Cottret, L., Lacroix, V., Auber, D., Mary, P., Sagot, M.F., Jourdan, F.: Metabolic network visualization eliminating node redundance and preserving metabolic pathways. BMC Syst. Biol. 1, 29 (2007)
2. Brandes, U., Dwyer, T., Schreiber, F.: Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions. Journal of Integrative Bioinformatics 1, 1 (2004)
3. Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J., Giegerich, R.: PathFinder: reconstruction and dynamic visualization of metabolic pathways. Bioinformatics 18, 124–129 (2002)
4. Kojima, K., Nagasaki, M., Miyano, S.: Fast grid layout algorithm for biological networks with sweep calculation. Bioinformatics 24(12), 1433–1441 (2008)
5. Schreiber, F.: High quality visualization of biochemical pathways in biopath. Silico. Biol. 2(2), 59–73 (2002)