GraphLinker: A Visual Comparative Environment of Genomic and Metabolic Networks

Niels Hanson

October 31 2011

Introduction

Micro-bacterial communities make up a wide and diverse environment which accounts for some of the most diverse and largely unexplored biomes on the planet. Environmental genomics represents a window with which to view and explore the diversity and dynamics of these naturally occurring communities. One of the new emerging concepts in this field is the distributed nature of metabolic pathways between the multitude of different taxa within an environment. In higher order organisms, these pathways are viewed in isolation of other species, however, it is becoming increasingly apparent that these networks in the microbiome are distributed in nature between a number of different microorganisms; each organism relying on the metabolic processes of others within a traditionally isolated pathway. However, discovering specifically which taxa are contributing within a metabolic pathway is a non-trivial question that is just beginning to be investigated.

To add some formalism to the problem, we can treat microbiological taxa and metabolic pathway steps as nodes in two somehow separate, though related, networks or graphs. The presence of a specific taxa in an environment can be estimated from genomic reads of the 16S rRNA protein, using them as a putative identifiers. A collection of reads are clustered by multiple sequence alignment into operational taxonomic units (OTUs) which correspond to one or more closely related species contained in a sample. Metabolic pathway steps exist in a known network of overall metabolism with a complex yet defined order of substrates and products representing the inputs and outputs of each reaction. The presence or absence of these enzymatic steps can be inferred by via a pattern of expression detected in specific enzymatic proteins. The PathoLogic algorithm developed

by the International Standford Research Institute (SRI International) can infer the presence or absence of specific pathway steps, effectively supplying nodes, their connecting edges, and their weights in the graph. In the other graph representing taxa, OTUs can be clustered in a hierarchy using correlation algorithms, which essentially describe the edges and their weight.[1] The problem is of course to effectively infer the influence of the taxa graph on the metabolic one, which essentially boils down to a version of the Graph Isomorphism problem, a classic NP problem in computer graphing.

Graph Isomorphism is one of the twelve classic computational complexity problems proposed by Garry and Johnson back in their 1979 seminal textbook of the subject.[2] However, it is peculiar in that it remains one of the last to still have its computational complexity unsolved. It is still unclear if the problem exists in the set of polynomial or NP-complete problems.[3] The best algorithm to date runs in $2^{O(\sqrt{nlog(n)})}$ where n is the number of vertices in an undirected graph.[4] However, despite not having the complexity of an exact solution solved, several practical algorithms have been proposed which in most cases run in polynomial time despite being still exponential in the worst case.[5,6]

Bringing the problem back into biological context, previous attempts to solve a similar mapping of gene expression to pathway targets have been attempted, all-be-it with mixed success, classifying less than 50% of expressed enzymes to a pathway.[7] This could be due to a number of factors, including the KEGG pathway database being incomplete at the time with respect to the taxa in question. However, these previous attempts considered bacterial taxa in isolation, where as distributed metabolism amongst multiple microorganisms may be a more correct perspective to analyze this domain. In our model, we propose to apply hierarchical clustering to the nodes in each graph for two main reasons. The first, is to simplify the data set to gather a simpler better overview of the global relationships. The second, and more important, reason is to include clusters of prospective taxa and pathway steps into the mapping of one graph on to the other. It is thought that mappings which include multiple levels of the data hierarchy will likely be more successful, and thus reduce the risk of having the mapping fail merely because it was looking at too low a level.

The analysis of this mapping is still exploratory in nature, and at this outset it is difficult to be specific of the kinds of interaction that will be required in the visualization. Nonetheless, this proposal specifies the best guess based on previous work in the visual encoding of small world graphs and previous biological tools that have graph interaction schemes.[8] For the purposes of CPSC 533C, the scope of this project is to explore methods for visually encoding weighted associations between two graphs in the context of graph isomorphism and microgenomic data.

Personal Experience

As a masters student in the UBC Bioinformatics training program, the task of relating the microbacterial communities with their distributed metabolic pathways is a key part of my research interest in studying microbiome genomics as a whole. Currently in the Hallam Lab, we are finalizing a pathway prediction pipeline using microbacterial genomic expression data based on the PathoLogic algorithm produced by bioinformatics researchers at SRI, International.[9] Additionally, investigating bacterial community structure based on presence absence data using a number of different mathematical models based on the properties of small world graphs. Both of these research directions make up the two sets of data essential to the visualization task, making this project very unifying in terms of my research goals. Additionally, the task of visually describing graph isomorphism is an interesting problem that would be relevant any data set that requires the comparison of two related graphs representing different domains.

Proposed Information Visualization Solution

A Mathematical Description of the Graphs

Let there be two graphs, G and G', with nodes, n and n', edges, $e(n_i, n_j)$ and $e'(n'_i, n'_j)$, and clusters $c(n_1, n_2, ...)$ and $c'(n'_1, n'_2, ...)$. Let a connecting edge between graphs be expressed by $e_c(n_i|c_i, n'_j|c'_j)$. In all above cases, let $i \neq j$. Clusters of nodes or clusters from when classified by some general hierarchical clustering algorithm. See Figure 1 for a more visual description of the layout.

Biological Mapping

A quick summary showing the mapping of the graph to their biological features.

G Graph of putative taxa in the form of OTUs: a node, n, represents the OTUÕs presence in the sample set being displayed.

G' Graph of putative metabolic pathways: a node, n', represents a metabolic step called by PathoLogic

$e(n_i, n_j)$ Edges between nodes i and j of graph G. This represents a co-occurrence between two OTUs across samples within an environment. It could be weighted or unweighted depending on the dataset.

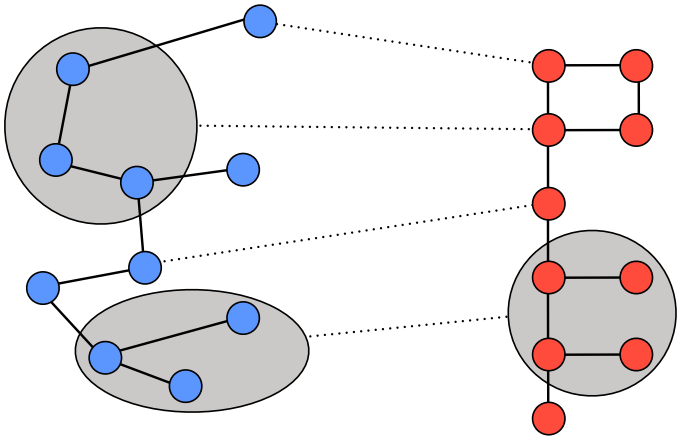

Figure 1: A schematic describing the graph isomorphism problem and its mapping into the biological domain of operational taxonomic units (blue) and metabolic pathway units (red). Notice that edges within a graph are solid, while meta-edges connecting the two graphs are dotted. Clusters or meta-nodes represent a hierarchical clustering on a set sub-set of nodes, but can also treated as a target for meta-edges. These clusters represent groups of putative co-occurring bacteria in the OTU graph, while they represent a cohesive metabolic pathway in the other.

$e'(n'_i, n'_j)$ Edge between nodes i and j of graph G'. The edge represents a connection between two metabolic steps. It could be weighted or unweighted depending on the dataset.

$c(n_1, ..., n_k)$ A cluster or meta-node of k nodes of graph G. A cluster of correlation between taxa; a putative microcommunity.

$c'(n'_1, ..., n'_k)$ A cluster or meta-node of k nodes of graph G'. A cluster of correlation between metabolic steps; a putative cohesive metabolic pathway.

$e_c(n_i|c_i, n'_j|c'_j)$ A connecting or meta-edge between a node or cluster of graph G to a node or cluster of G'. Weight determined by the weighting algorithm proposed by Ogata et. al. This edge suggests a putative association between OTUs in G to metabolic steps in G'

Visual Encoding

The two graphs will be displayed in the same view though separated by some minimal central whitespace (Figure 1). We initially expect the nodes to be laid out in a force-directed graph according to edge weight and have different colours according to the graph they come from for clarity. In first schematic, Figure 1, nodes in the the OTU graph are blue and taxonomic nodes are red. Edges connecting nodes within a graph straight, full lines, while meta-edges between graphs are dotted. Hierarchical clusters are represented by background ellipses containing all child nodes beneath them. In our biological problem, one layer of hierarchical clustering might be sufficient for the problem, however, many other problems might require higher levels. In which case, the hierarchy might be represented by nested concentric ellipses akin to the Grouse hierarchy visualization tool (Figure 2).[10]

OTU graphs can contain thousands of nodes, while in metabolic pathways there are likely to be hundreds. This makes this a moderately difficult graph drawing problem in terms of screen real estate. Two noise/data reduction methods are proposed, collapsing meta-node clusters and 'ghosting' or removal of nodes and edges that do not have meta-edges (Figure 2). This way the only edges and nodes that remain are those that connect the two graphs, highlighting the connections found. The connections are also highlighted via the thickness of the lines, encoding the strength of the particular association between the two nodes or clusters. This is further improved by moving all meta edges and their nodes along the centre of the screen, giving them a 'front and centre' visual display.

It is unclear how connected the our data is going to be, however, it is likely that they will be 'small world' graphs with a high clustering index and a small average path length. Since edge crossing and density are the primary impediments people have when analyzing graphs, the replacement of many weak edges with one edge in a cluster might be important.[11] User interaction and movement, important though not within the main scope of this proposal, will likely be limited to basic panning and zooming over the contents. Other features beyond scope, inspired by the Prawn visualization display, include guaranteed visibility of selected nodes at the edge of the frame at their tangential location, text scaling to ensure ease of reading, and when expanding a meta-node or cluster, moving the surrounding nodes such that this expansion does not occlude them with an occlusion avoidance algorithm.[12]

One should note that this represents the initial specification of the design and visual encoding given the current understanding of the problem. It is by no means permanent and is quite likely, given the exploratory nature of the data, to change as development continues due to certain programming, display, or computational constraints that arise.

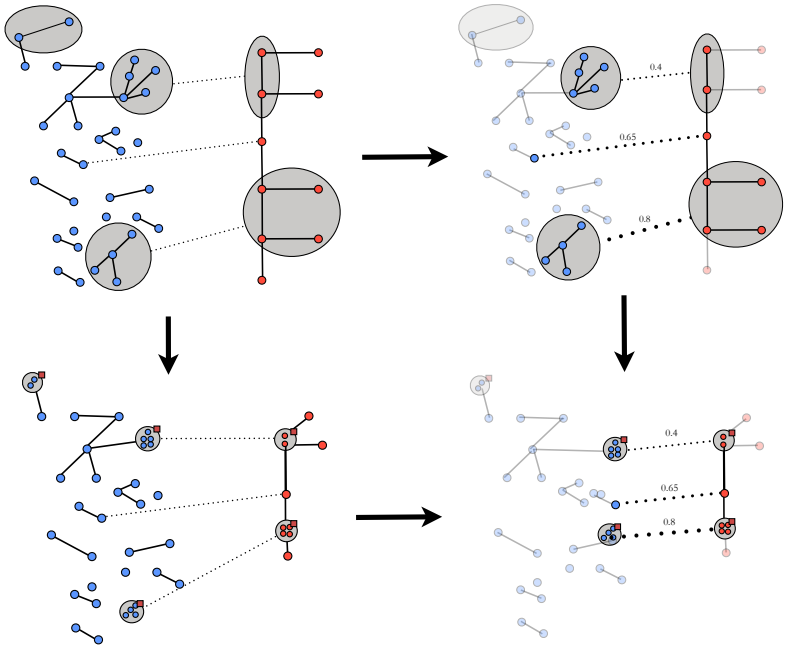

Figure 2: A proposed visual encoding for graph comparison. Graphs are separated via some central whitespace, however, due to the likely number of nodes of OTU and metabolic networks some simplifying measure are propose. Meta-nodes or clusters in each graph can be scaled down by removing the internal connecting edges and scaling the size of all nodes proportionately down (bottom left). The node positions however remain relative to their original configuration with edges as to provide a thumbnail into their underlying structure. To highlight edges and nodes that are connected between the two graphs, all other nodes are ghosted and nodes connected by meta-edges are brought to the centre of the screen (top right). Both cluster-scaling, and node ghosting can be combined to further highlight the connections between the two graphs.

Proposed Implementation

The open source visualization environment, Cytoscape, is a popular framework within the biological community for complex network analysis, and is likely a first chose for the implementation.[13] A number of biological plugins have already been implemented with large support from the scientific community.[8] Cytoscape, is also good in that it has both PC and Mac implementation as deployment to both is important in the case of the biological systems community. Recently, Cytoscape has also developed a web based

version, Cytoscape Web, for web deployment. However, it is unclear if this would be effective for our needs, considering the size of our graphs that we are likely to require.[14]

The fallback implementation, should it be required, is to have basic functionality demonstrated in Processing, a visualization development package written in Java. Priority will be placed on showing the meta-nodes connecting the two graphs as this is the focus of the visualization. Expanding/Contracting hierarchies and specializing the visualization for small world graphs will likely help, but the concepts have been well thought out in previous work.[10,11,15–17]

Schedule & Milestones

Here we outline the schedule for development of GraphLinker.

Previous Work

The using graphs to interpret information and their visualization issues has in some ways always existed from the time of Euler. However, the formalism of the graph visualization community really grew around yearly Symposia on Graph Drawing which started in 1992 in Rome. Since then, research has identified three general overlapping problems of graph drawing: node and edge occlusion by density, readable edge layout, and computational complexity of drawing algorithms.[18] The intuitive impact of graphs and their cognitive interpretation has unfortunately been less formally studied when compared with other aspects of visualization like colour. This makes the area more holistic and reliant on usability studies for validation.[18]

Many real world graphical datasets have the global property of large number of highly distributed clusters with a small average path-length. This is known as the 'small world' or power law property in the literature, and has been be shown to occur many different datasets across multiple knowledge domains. Recently, there have been a number of attempts to optimize the visualization to graphs having these properties though the specialized use of force-directed layouts.[11] These forced-based approaches, though initially too computationally complex and non-deterministic, have had a number of proposed algorithmic optimizations that make the problem tractable.[18] More recent developments have been to apply force-directed approximations to online (live) graphs to facilitate user interaction.[19] Ham and Wijk have used both semantical and geometrical distortions to create scalable, interactive visualizations of small world graphs.[16]

Another aspect of graph layout has been the visualization of hierarchical clustering. A number of classic layouts are highlighted by Herman in his review, however, the work

Milestone	Date
1. Become accustomed with the most recent version of Cytoscape, implementation of previous plug-ins etc.	October 29th
2. Adding edge weights to edges, displaying two separate graphs at the same time.	November 1st
3. Cluster nodes by containment.	November 4th
4. Linking nodes on two graphs by edges of a different style. Change edge width due to weight of association.	November 6th
5. Bring meta edges and associated nodes to the centre, sort vertically by edge weight.	November 8th
6. 'Ghost' or remove nodes and edges not associated with meta-edges.	November 10th
7. Have basic demonstration of functionality ready to go for project update presentation.	November 12th
8. If four of the last 7 tasks are not complete, change to 'Plan B' of implementing basic functionality in Processing.	November 15th
9. Basic Pan & Zoom should work. Keep highlighted nodes in the main view at a tangential angle.	November 18th
10. Have all meta-edge functionality working. Move on to extra features of multi-nested hierarchical clustering.	November 20th
11. Collapsing and expanding clusters or meta-nodes.	November 22nd
12. Implement occluding avoidance algorithm when expanding or collapsing nodes	November 25th
13. Final extras and features isolated. Have second demonstration model ready to present.	November 30th
14. Begin final writeup. Start replacing theoretical data with legitimate calls from biological sources.	December 1st

of Archambault *et. al.* has investigated into the visualization of small world graphs from a number of additional different angles. In TopoLayout he proposed a multi-level algorithm that draws undirected graphs based on the topological features and improves upon a number of different layout algorithms.[17] Grouse, another visualization environment, tackles the hierarchical normally shown in a tree layout by collapsing nodes into concentric meta-nodes with meta-edges and transforms the visualization of the graph up the structural hierarchy.[10] Furthermore, Archambault has recently touched upon graph comparison through the overlapping of two graphs in difference maps when nodes of the graph exist in the same set of objects.[15] These are all useful and relevant advances, however, to our knowledge this proposal represents the first attempt at the visual mapping of two related graphs from different domains.

References

1. Chaffron, S., Rehrauer, H., Pernthaler, J. & Mering, von, C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res 20, 947-959 (2010).
2. Garey, M.R. & Johnson, D.S. Computers and intractability. 338 (W.H. Freeman & Company: 1979).
3. Mulzer, W. & Rote, G. Minimum-weight triangulation is NP-hard. J. ACM 55, 1-29 (2008).
4. Johnson, D.S. The NP-completeness column. ACM Trans. Algorithms 1, 160-176 (2005).
5. Schmidt, D.C. & Druffel, L.E. A Fast Backtracking Algorithm to Test Directed Graphs for Isomorphism Using Distance Matrices. J. ACM 23, 433-445 (1976).
6. McKay, B. Practical Graph Isomorphism, Congressus Numerantium, 30. 88 (Utilitas Mathematica Pub. Inc.: 1981).
7. Ogata, H., Fujibuchi, W. & Goto, S. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic acids ... (2000).
8. Barsky, A., Munzner, T., Gardy, J. & Kincaid, R. Cerebral: visualizing multiple experimental conditions on a graph with biological context. IEEE Trans Vis Comput Graph 14, 1253-1260 (2008).
9. Karp, P. & Paley, S. The pathway tools software. Bioinformatics (2002).
10. Archambault, D. & Munzner, T. Grouse: Feature-based, steerable graph hierarchy exploration. Proc of Eurographics/IEEE VGTC ... (2007).
11. Auber, D. & Chiricota, Y. Multiscale visualization of small world networks. ... Visualization (2003).

12. Chan, A. Prawn: An Interactive Tool for Software Visualization. University of British Columbia (2000).
13. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13, 2498-2504 (2003).
14. Lopes, C.T. et al. Cytoscape Web: an interactive web-based network browser. Bioinformatics 26, 2347-2348 (2010).
15. Archambault, D. Structural differences between two graphs through hierarchies. Proceedings of Graphics Interface 2009 (2009).
16. van Ham, F. Interactive visualization of small world graphs. ... Visualization (2004).
17. Archambault, D. & Munzner, T. Topolayout: Multilevel graph layout by topological features. IEEE Transactions on ... (2007).
18. Herman, I. & Melan◌on, G. Graph visualization and navigation in information visualization: A survey. Visualization and ... (2000).
19. Frishman, Y. Online dynamic graph drawing. IEEE Transactions on Visualization and ... (2008).