Basic Info

Project Title: Profondeur
Project Repository: https://github.com/tcameronwaller/profondeur
Author Name: Thomas Cameron Waller
Author EMail: cameron.waller@biochem.utah.edu
Author Identifier: 00409400

Background and Motivation

Metabolism sustains all of life's diverse processes, from the release of energy from food to the synthesis of genomes. It is an integral part of the biological system of the cell with extensive connections throughout. Given its central role, it is no surprise that **defects in metabolism drive the pathology of many human diseases[5,4].** These include obesity, diabetes, cardiovascular disease, and cancer. More and more studies in biological research consider metabolic phenotypes and mechanisms, and metabolic profiling is an important component of precision medicine[1,2,3,11,9]. For these reasons **it is necessary to study metabolism from a holistic perspective, in its true context within the entire biological system.**

The vast complexity of metabolism exceeds the ability of human investigators to consider it without supporting methods. As part of the biological system of the cell, metabolism forms an intricate network of relationships between genes, transcripts, proteins, and metabolites. In the human genome, approximately 70,000 genes encode 70,000 transcripts and 30,000 proteins[11]. About 5,000 of these proteins are enzymes or transporters[7,6,8] that either catalyze chemical reactions or transport molecules between cells and sub-cellular compartments. Many more proteins regulate metabolic processes via the expression and activity of relevant genes, transcripts, and proteins. Approximately 30,000 endogenous small molecular metabolites[10] complete the collection. All of these genes, transcripts, proteins, and metabolites relate to each other in multiple ways, forming the metabolic network. This network varies in different tissues throughout the human body. This stunning complexity makes it prohibitive to track all processes without advanced methods to organize and access the information of the metabolic network. We argue that **many studies certainly miss interesting biological trends because the context is too elaborate**.

The author of this project has an interest in the study of metabolism. His primary research project as a graduate researcher is in the visual analysis of the metabolic network. Indeed this course project is a small-scale start to a bigger research project in collaboration with Alexander Lex, Jared Paul Rutter, and others.

Project Objectives

The biological study of metabolism has potential to benefit the domains of biotechnology, pharmacology, and medicine. Biology's goal is to discover and describe life's functions and processes. As a basic science this domain is primarily informative, and it enables applications in many other domains. Biotechnology alters the metabolic systems of non-human hosts to synthesize hormones, antibiotics, or other drugs. Pharmacology identifies effective drug targets within the biological system, considers the broad effects of drugs on this system, and even exploits subtle side-effects of available drugs for use in new therapies. Knowledge of processes in health enables medicine to recognize perturbations, diagnose diseases, and propose appropriate therapies. **The goal of this project is to develop methods that support the study of metabolism by users in these domains (biology, biotechnology,**

pharmacology, and medicine).

In their study of metabolism, a common task in these domains is to **access information about biological entities and the relations between them**. There are many different types of these **entities (genes, transcripts, proteins, metabolites)** and **relations (code, expression, regulation, transport, catalysis, reaction)**. They each have their own specific properties. In complex organisms, genes belong to chromosomes and chromosomal loci. In eukaryotes, proteins and metabolites localize to sub-cellular compartments called organelles. Many proteins have non-protein chemical cofactors that are essential for their functions. Reactions between metabolites belong to metabolic pathways of anabolism or catabolism and also have chemical properties such as rates and reduction or oxidation. A user might need to know any of this information for an entity, a relation, or combinations thereof. For example, a user might need to know which gene encodes a transcript and protein, which protein regulates the transcription of a gene, which metabolites react via a protein's enzymatic catalysis, or which catabolic reactions in the mitochondrial compartment involve proteins that require a specific cofactor. There are also many other possibilities. **Although much of this information is within current knowledge and is available in public databases, the scale of the biological system renders this access nontrivial.**

Another common task in these domains is the **analysis of experimental data that corresponds to entities in the biological system**. Modern -omics technologies describe the biological system at nearly-comprehensive scales. Available information includes mutations and copy variations in genes (genomics), abundance and splice variations of transcripts (transcriptomics), abundance and post-translational modifications of proteins (proteomics), and abundance of metabolites (metabolomics)[11]. It is very important to analyze and interpret these data in the real context of the biological system. For example, biological trends may only be apparent with consideration for network topology. **Correct analysis and interpretation of these experimental data rely on knowledge of the biological system.** This analysis becomes a great challenge at nearly-comprehensive scales.

The goal of this project is to develop computational methods for visual exploration and analysis of the metabolic network. It will address some of the specific needs of these domains. Many of the domain tasks correspond to queries against a database of network information. Analysis of experimental data in the context of network topology will require custom, probabilistic algorithms. Visual design will help the human user to recognize trends and patterns easily.

Summary

Goal: Develop computational methods for visual exploration and analysis of the metabolic network.
Users: Research scientists and engineers in biology, biotechnology, pharmacology, and medicine
Benefit: Advance knowledge of human health and ability to diagnose and treat human disease.
Domain Tasks:
1) Access information about biological entities and the relations between them.
2) Analyze experimental data in the context of the metabolic network.

Data

This project involves two categories of data. The first category includes the current knowledge about entities and relations in the biological system and their properties. This information is available in multiple public databases. This information is enormous and complex, and it is unlikely that all of the information will be available from a single source. To maintain a reasonable scope, this project will use a minor subset of this information. The second category is experimental data from -omics technologies. This sort of data is available in multiple public databases. This sort of data often has technical complications and requires processing. Again, to maintain a reasonable scope, this project probably will not use this category of data.

Data Processing

Some aspects of this project will involve substantial data processing. Information on the metabolic network will come from multiple databases. It will be necessary to organize this information in an appropriate format and assemble an accessible database. Experimental data from -omics technologies is also complex and sometimes requires extensive data processing before it is informative or interpretable.

Visualization Design

The prospective users of this project need to access information about biological entities and the relations between them. This domain task of accessing information involves multiple abstract tasks.

It is necessary to access information that matches the user's interest. With several types of entities, several types of relations, and many instances of each, the information of this network is vast and complex. The user needs to consider subsets of the metabolic system in appropriate detail (according to interest) without neglecting potentially-relevant context. Therein is a challenge, simplifying the information without losing important aspects. This process of selecting a relevant subset of the network will likely require multiple levels of user interaction. **Queries against a comprehensive database will allow the user to define her or his own interest in a fairly broad or general scope.** These queries will need to accommodate many variables in order to provide sufficient specificity. **A visual, interactive interface for dynamic queries will help the user to specify subsets of interest.** Search algorithms will then select relevant subsets of the data, ranking several possibilities according to query criteria. Then the user will be able to select from a list of search results. **A visual interface will help the user to distinguish between search results and select whichever is most appropriate.** After selecting a result from a query search, the user will need to refine the subset further according to her or his specific interest. The user will need to select which types of entities and relations to represent. The user might even need to select specific entities and relations to represent. **A visual interface for a selection menu and an interactive visualization of the network data will help the user to refine interest subsets.**

It is necessary to communicate the relevant information to the user clearly and efficiently. The user may need to study small or large subsets of the metabolic network at various degrees of detail. These subsets will include a lot of information. From these subsets, users need to recognize connectivity and many properties of entities and their relations. **A visual representation of the network, such as a node-link diagram, will communicate this information to the user.** This diagram will require a dynamic layout. Alternative visual representations of the network might also work. Also, due to the complexity of the metabolic network and the limitations of available visual channels on a two-dimensional display, it will probably be necessary to represent information in multiple views. The user will be able to switch or scroll between views. Animations and other strategies will help to preserve the user's sense of context through these transitions.

It is necessary to associate experimental data with entities and relations from the metabolic network and to identify trends in these data. Keys (such as names of genes, proteins, or metabolites) will associate experimental data to specific entities in the database. Search algorithms will then handle experimental data similarly to other properties or attributes to recognize pathways with significant differences.

Must-Have Features

- interactive selection menu to include or remove specific types of entities and relations

- visual representation for subset of the metabolic network
- interaction with the network representation to include or remove specific elements
- visual representation of multiple types of properties on the network representation

Optional Features

This course project only entails 6 weeks of time for 1 developer. To maintain a reasonable scope, several features will be optional. These features will be more important for a longer-term, more complete project.

- comprehensive database for current knowledge of the metabolic network
- dynamic query interface
- search algorithms to identify pathways between entities matching query criteria
- analytical algorithms to identify network aspects with interesting trends in experimental data
- interface for selection from list of search results
- visual representation of experimental data on the network representation

Project Schedule

This course project entails 6 weeks of time for 1 developer.

Weekly Goals

1. 24 October - 30 October
    - visual design of interface, especially network representation
2. 31 October - 6 November
    - visual design of interface, especially network representation
    - collection, processing, and organization of network data
    - implementation of visual interface
3. 7 November - 13 November
    - implementation of visual interface
    - milestone review
4. 14 November - 20 November
    - implementation of visual interface
    - implementation of interaction
5. 21 November - 27 November

- implementation of visual interface
- implementation of interaction

6. 28 November - 2 December
    - screen-cast tutorial video
    - website
    - final submission

References

[1] K. Beebe and A. D. Kennedy. "Sharpening Precision Medicine by a Thorough Interrogation of Metabolic Individuality". In: *Comput Struct Biotechnol J* 14 (2016), pp. 97–105.

[2] M. Benson. "Clinical implications of omics and systems medicine: focus on predictive and individualized treatment". In: *J. Intern. Med.* 279.3 (Mar. 2016), pp. 229–240.

[3] F. S. Collins and H. Varmus. "A new initiative on precision medicine". In: *N. Engl. J. Med.* 372.9 (Feb. 2015), pp. 793–795.

[4] D. W. Haslam and W. P. James. "Obesity". In: *Lancet* 366.9492 (Oct. 2005), pp. 1197–1209.

[5] A. A. van der Klaauw and I. S. Farooqi. "The hunger genes: pathways to obesity". In: *Cell* 161.1 (Mar. 2015), pp. 119–132.

[6] N. Pornputtapong, I. Nookaew, and J. Nielsen. "Human metabolic atlas: an online resource for human metabolism". In: *Database (Oxford)* 2015 (2015), bav068.

[7] N. Swainston et al. "Recon 2.2: from reconstruction to model of human metabolism". In: *Metabolomics* 12 (2016), p. 109.

[8] I. Thiele et al. "A community-driven global reconstruction of human metabolism". In: *Nat. Biotechnol.* 31.5 (May 2013), pp. 419–425.

[9] E. J. Topol. "Individualized medicine from prewomb to tomb". In: *Cell* 157.1 (Mar. 2014), pp. 241–253.

[10] D. S. Wishart et al. "HMDB 3.0–The Human Metabolome Database in 2013". In: *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D801–807.

[11] X. Zhang, J. A. Kuivenhoven, and A. K. Groen. "Forward Individualized Medicine from Personal Genomes to Interactomes". In: *Front Physiol* 6 (2015), p. 364.

Date: 11 October 2016
Name: T Cameron Waller
Number: 1

Must be dynamic with useful feed-back.

Query Interface

Query supports pathways by name but also non-arbitrary option.

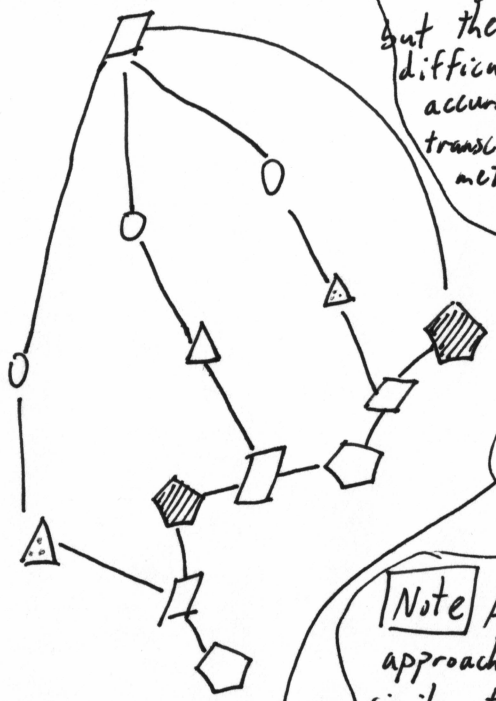
Notes

Study literature for examples of visualization of multi-dimensional networks (such as PathFinder). How to handle the dimensionality?

- Queries determine the scope of nodes and links of interest (interest set)
- Use path ranking by attributes (flux rate, exp data) or by path length or something.

Main View

Maybe show the entire set initially? ~~or~~ Or start simple?



The genes are connected, but the connection is difficult to represent accurately without transcripts, proteins, and metabolites. Really genes could relate in multiple ways that would be ambiguous without the other node types to clarify.

Note: An alternative approach: use a strategy similar to Entourage. Use separate views for networks of different dimensions (types of nodes and links) with links/portals between these sub-network views.

I think the understandability of this strategy would be limited. Think about it. Some connections go through others as above.

The idea

1) Query the network to define interest set of nodes and links.

Query returns multiple, ranked sets and user selects (like PathFinder)

Also support experimental data-dependent auto-query recommendations.

2) User selects nodes, links, and relations from interest set to show according to desired detail.

Support variety of visual encodings → For example which relations should be links and which should be regions (containment)?

Network Builder

Node Selector

Link Selector

Network Dimensions

The set of nodes and links is dependent on the query. Query defines nodes and links in interest set. Then select more specifically from that set.

Node Selector

(select to include)/turn on/off

- Genes
- Transcripts
- Proteins
- Metabolites

Maybe also allow drag-and-drop ranking to prioritize

Consider how to maintain point-of-reference as the network representation changes. Let user specify reference node to highlight or anchor as network changes. ↑ also be give automatic recommendation from query or node attributes.

Link Selector

- En Code
- Expression
- Regulation
- Transport
- Catalysis
- Reaction

Data Mapper

> select which data to show and how.

Name: T Cameron Waller

Date: 11 October 2016

Number: 2

#1. Note: The main idea here is to represent the multiple dimensions of the biological network (types of links between nodes) in separate views. The strategy is sort of similar to that of Entourage. The * distinct sub-networks would be linked.

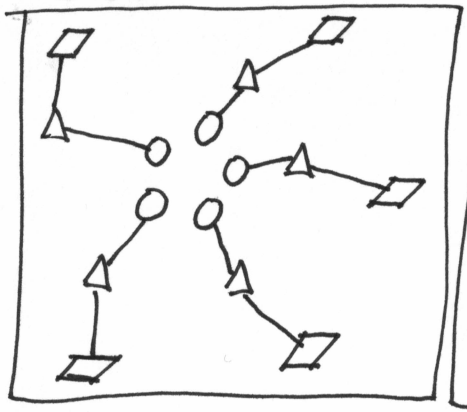
I think the number of types of nodes and links makes this strategy too complex and busy. ~~For example, you would have to show~~ At least it might be too complex to show all sub-network types. Maybe simplify with fewer.

Query Interface

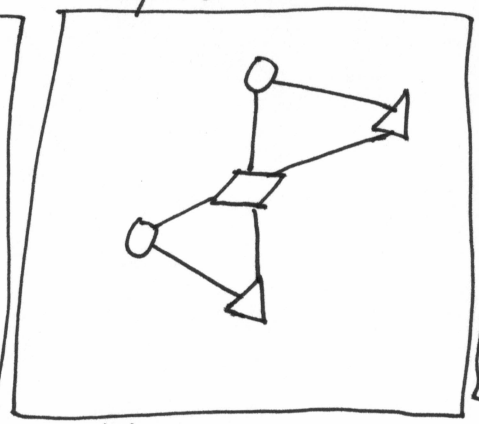
Ranked Paths or interest-based network subsets.

Network Dimension Views

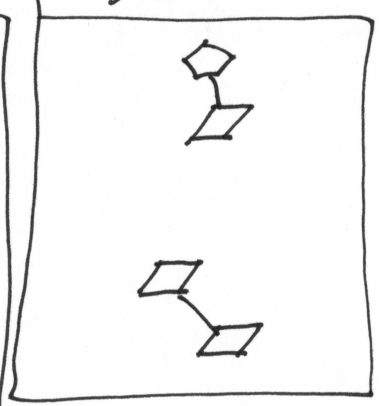
Code



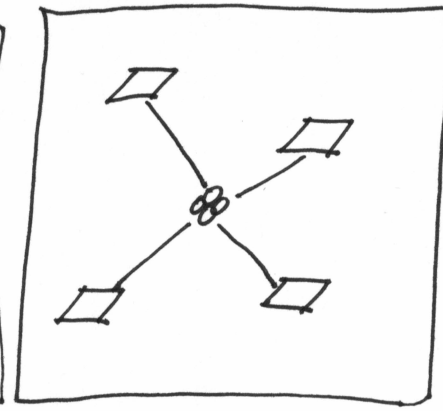
Expression



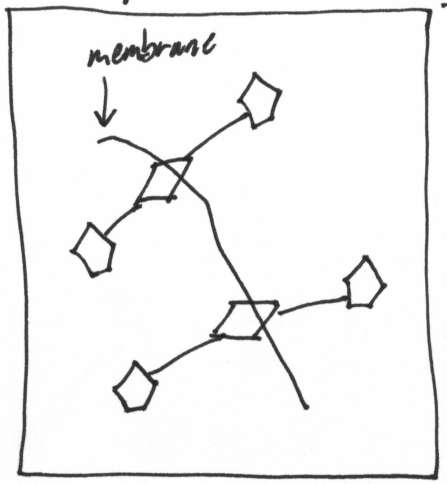
Regulation



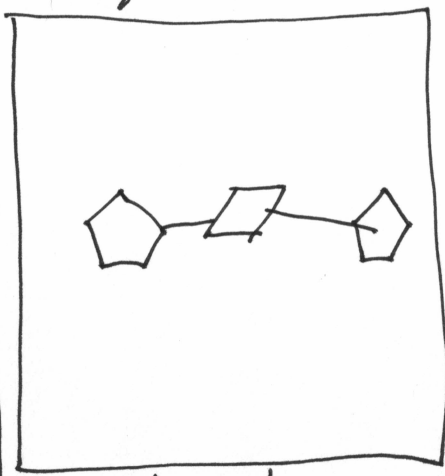
Cofactor



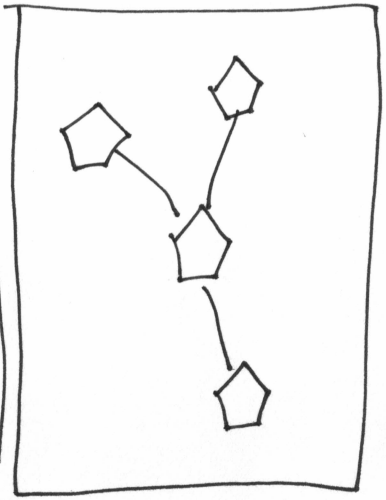
Transport



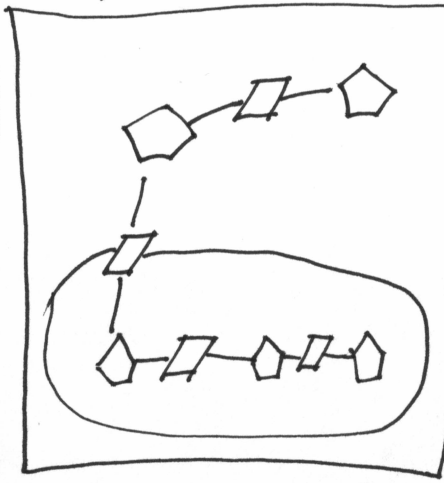
Catalysis



Reaction



Compartment



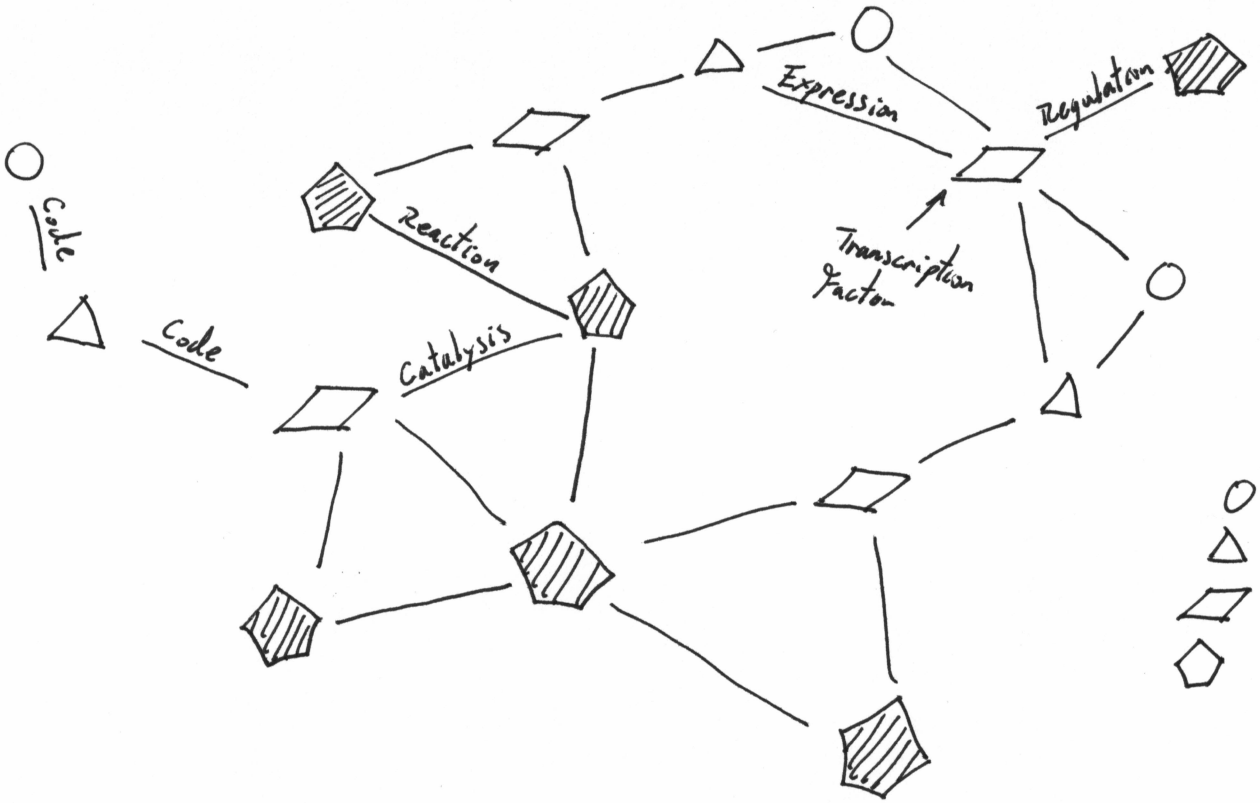
- Gene
- Transcript
- Protein
- Metabolite
- Cofactor

Example of relations that only make sense when considering multiple dimensions simultaneously.

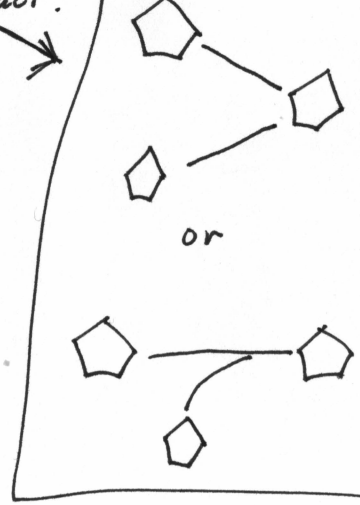
Note: Another problem is that it is difficult or confusing to try to separate some dimensions of the network. For example, transport makes little sense without compartment. Networks relate to each other via interactions between their proteins

Name: T Cameron Waller
Date: 11 October 2016
Number: 3

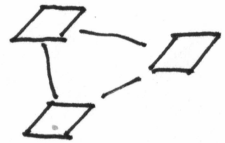
How do I represent multiple reactants going to the same product? That problem gets tricky when there are many reactants or many products.



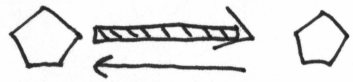
Gene
Transcript
Protein
Metabolite.



How do I represent protein complexes?



How do I represent reaction directionality or reversibility?



Maybe that is an optional level of detail.

Note

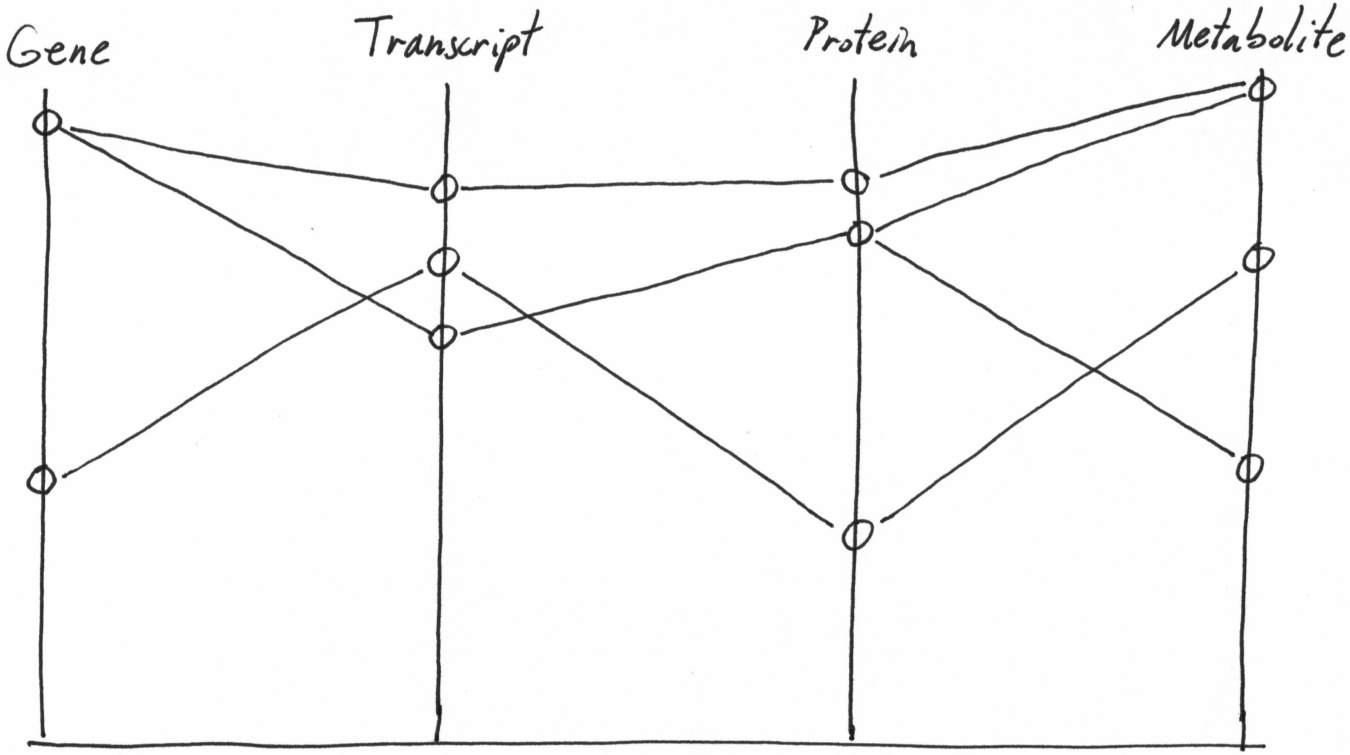
There are so many types of nodes and links (dimensions of the network) that ~~it beco~~ visualization becomes very complex very quickly, even with a small number of nodes.

How do I represent different types of nodes and links clearly?
- color? Line type for links? ~~tight~~ Line width for links?

The complexity is so great that it ~~is~~ will be very important to help the user to specify nodes and links of interest. Help user do that within context of entire network.

Name : T Cameron Waller
Date : 11 October 2016
Number : 4

Note: I had been encoding node type by shape. The idea here is to encode node type by position on a parallel coordinates chart. Position can also represent ranking of nodes by attributes, such as experimental data.



Note] It is very effective to distinguish between different nodes of the same type and ~~nodes~~ between nodes of different types.

It is very effective to represent one-to-many relationships, such as between an enzyme and multiple metabolites that are reactants, products, or allosteric regulators.

It is very effective to rank nodes by attributes and observe trends in those attributes.

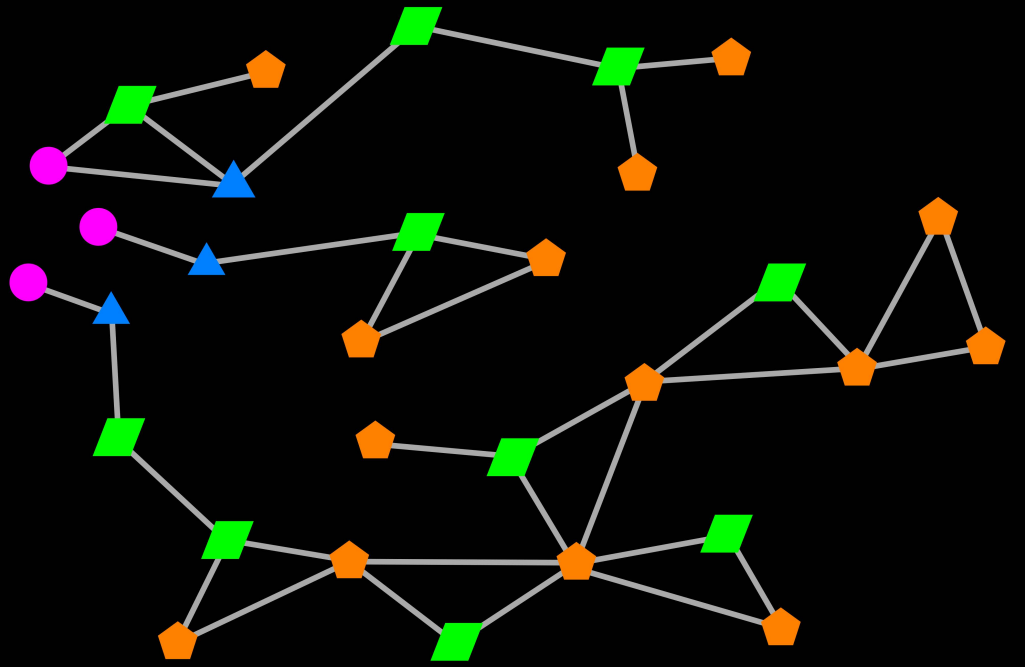
Dynamic Query

Common or
Custom Pathways

Rank by Attributes

Results

Network View



3D Context View

Selection

Nodes

- Gene
- Transcript
- Protein
- Metabolite

Links

- Code
- Expression
- Regulation
- Transport
- Catalysis
- Reaction