Note to Jared

This is my first attempt at composing a "Design Description Document", and it is different than a lot of documents (such as proposals) in biology. As I understand it, the purpose of the document is to describe in detail **1) the goal of the project, 2) the specific requirements to reach that goal, and 3) the design decisions for specific features to meet those requirements**. In this type of project there are a lot of decisions about details that matter and make a big difference (such as whether or not to host the tool on a server). The idea is that making those decisions early on and getting feedback will help to make them better. Also the thorough plan for the project will help the project succeed towards a specific goal. That said, this document does not even come close to addressing all the details and decisions for this project. It will be necessary to make a lot of those decisions as we get to them, since problems and opportunities will probably arise during implementation.

Please consider the following subjects as you critique the document.

1) How valuable is the goal of the project? How well does the document communicate this goal?

2) How accurately does the document describe the user? How well does the document communicate the requirements of the project with respect to the user? What requirements would you include or exclude?

3) How appropriate are the design features to support the study of metabolism? What features would you include or exclude?

Thank you very much for reading the draft and giving feedback!

To Do...

1) Figures

This draft does not yet have **figures to illustrate the design of the graphical interface**. Planning out the graphical design of the interface will be necessary and important for the project, so it'll be worth it to spend the time now so that I can get feedback. Designing figures will be my next step on this document, and it could be a lot of work.

2) Schedule

This draft's description of the implementation steps and schedule in the **"Schedule"** section is incomplete. This section is tricky, since it breaks the implementation process down into detailed steps. It's also difficult to anticipate accurately how long each step will take.

Summary

Metabolism is a vast and complex system with many types of relations between many types of biological entities. With extensive connections throughout, it functions as a cooperative network. To biologists and clinicians who study metabolism, this complexity and continuity present a great challenge. Investigators need to consider the holistic context of the network in order both to design experiments reliably and interpret experimental results accurately. Current technologies primarily divide the metabolic network into distinct pathways and show these in static maps. These tools fail to represent the continuity of the metabolic network as a whole. They also fail to support exploration according to case-specific interest. The goal of this project is to develop computational methods and a tool to enable users to explore the metabolic network. In contrast to current technologies, this tool will support queries against the entire network to define custom selections. This tool will represent these selections dynamically, offering alternative views to illustrate properties and attributes clearly.

Introduction

Metabolism is an integral part of the biological systems of cells, tissues, and organisms. **Catabolic processes** degrade food to liberate energy and materials. **Anabolic processes** utilize these materials and harness this energy to synthesize all cellular components. In this way, **metabolism sustains life and all of its diverse processes**. It is little wonder that metabolic defects contribute to the pathology of many human diseases, including obesity, diabetes, cardiovascular disease, and cancer[1,2].

Metabolism is a cooperative and continuous system. In the cell, tens of thousands of different types of **genes, transcripts, proteins, and metabolites** exist in various amounts in separate compartments. These **compartments partition the cell into environments with specific purposes**. Genes and transcripts encode proteins, and proteins have many functions. **Proteins transport metabolites between compartments and catalyze chemical reactions between metabolites**, performing these tasks with exceptional precision. They control and regulate the rates and compartments of transports or reactions as well as which metabolites participate. Metabolites, in turn, are the fuel and building blocks of cells. All of these diverse functions influence each other such that **local modifications can impart pervasive effects throughout the system**.

The study of metabolism is changing. Traditionally this study involved experts who conducted **reductionist experiments** to characterize the functions of individual genes, transcripts, proteins, and metabolites. Over many years, many of these experiments have contributed sufficient knowledge to consider the **greater perspective of pathways, processes and entire systems**[3]. Modern technologies, especially the **omics technologies (genomics, transcriptomics, proteomics, and metabolomics)** measure these systems at nearly comprehensive scales and provide rich information[4]. Also, there is an increase in **interest and accessibility** such that the study of metabolism is becoming more **interdisciplinary**. Scientists and engineers from diverse backgrounds consider metabolic mechanisms and phenotypes, and metabolic profiling in the clinic is an important goal towards **precision medicine**[5,6,7,8,3,4,9].

These changes present challenges and opportunities. The modern study of metabolism needs an **holistic perspective**, such as that of the discipline of **systems biology**. Only this perspective sufficiently considers the **real context** in order to **design experiments reliably and interpret experimental results accurately**. However, the vastness and complexity of metabolism exceed even the ability of experts to understand in its entirety. Novice and expert investigators alike need tools to study the metabolic system. The goal of this project is to develop **computational methods and a tool to support the study of metabolism**.

Requirements

The target **users** of this project's tool have diverse expertise and interest in the study of metabolism. They are scientists, engineers, or clinicians in the domains of **biology, biotechnology, pharmacology, or**

Category	Concept	Expert	Intermediate	Novice
Encoding	**Genes** encode **transcripts** via **transcription**.	X	X	X
	Transcripts encode **proteins** via **translation**.	X	X	X
	Proteins mediate both **transcription** and **translation**.	X	X	
Expression	**Proteins** regulate the extent of both **transcription** and **translation** via interaction with **genes**, **transcripts**, and other **proteins**.	X	X	
	Metabolites regulate the extent of both **transcription** and **translation** via interaction with **proteins**.	X		
Reaction	**Reactions** involve a chemical change between **metabolites**.	X	X	X
	Reactants are **metabolites** that participate in the start of a **reaction**.	X	X	X
	Products are **metabolites** that participate in the end of a **reaction**.	X	X	X
	Some **metabolites** participate in more **reactions** than others.	X	X	
	Reactions have different **rates**.	X		
	Reactions have different **directions** and likelihoods/probabilities of **reversibility**.	X		
	Pathways are sets of multiple **reactions** that carry out major processes.	X	X	
	Reactions between **metabolites** make a cooperative and continuous network.	X		
Catalysis	**Proteins** catalyze **reactions** between **metabolites**.	X	X	
Transport	**Membranes** partition the eukaryotic **cell** into separate **compartments**.	X	X	X
	Proteins mediate **transport** of **metabolites** between **compartments**.	X	X	
Regulation	Multiple copies of the same **transcript**, **protein**, or **metabolite** are present in the cell in dynamic **pools** of variable amounts.	X	X	
	Metabolites regulate the extent of **catalysis** and **transport** via interaction with **proteins**.	X		
	Proteins regulate the extent of **catalysis** and **transport** via interaction with other **proteins**.	X	X	
	Changes to a single **metabolite** or **reaction** can influence others even great distances away in the network.	X		
Technology	**'Omics** technologies identify modifications to **genes**, **transcripts**, or **proteins**.	X	X	
	'Omics technologies measure the abundances of **transcripts**, **proteins**, or **metabolites** in a sample and thereby suggest the pool amount in a cell.	X	X	
	To compare relative abundances from multiple samples, it is necessary to consider any differences in **conditions** or **parameters**.	X		
Science	**Experiments** require specific controls to study specific variables.	X	X	

Table 1: Fundamental biological, technological, and scientific concepts relevant to the study of metabolism.

medicine. **Experts** are scientists or engineers who thoroughly understand the biological basis of metabolism. Their interest is either to discover new processes within this system or to modify it for some therapeutic or synthetic benefit. They understand relevant methods and technologies and know how to use them properly in experimentation. **Intermediates** are scientists or engineers who understand general biology but are less familiar with metabolism. Their interest is to consider some aspects of metabolism that relate to another subject. They also understand relevant methods and technologies and know how to use them properly in experimentation. **Novices** are probably clinicians who understand only basic biology. Their interest is to profile the metabolism of a patient in order to diagnose a disease or prescribe an appropriate therapy. They do not understand relevant technologies or how to use them properly in experimentation. **Table 1** describes how these categories of users understand fundamental concepts in the study of metabolism.

For all users, the study of metabolism has some general requirements. Broadly, users' goals are **1) to design experiments reliably** and **2) to interpret experimental results accurately**. Both of these goals require appropriate consideration for the context of the subject in study. Since **metabolism is a continuous system that functions cooperatively**, the real context is the entire system, but its vastness and complexity are inaccessible. To handle vastness, **users need information in detail for narrow selections without neglecting the broad context**. To handle complexity, **users need to explore different types of information in different ways**. In both cases, they need **flexibility to match their interest**.

Selection Requirements

Metabolism is a vast system. The user needs to select portions of the metabolic system that are simple enough to conceptualize in detail but inclusive enough to give accurate context. These selections need to match the user's interest appropriately.

1. **Select portions of the metabolic system by how they relate to a set of entities of interest.**
   The user has interest in single or multiple entities (genes, transcripts, proteins, or metabolites) and needs information about these entities and also about other entities that relate to them. This selection must be appropriate for the user's custom, case-specific interest. It must be independent of typical pathways or other categories since interest often traverses or transcends these pathways.

2. **Select portions of the metabolic system that either have or do not have some specific property or attribute.**
   Entities have many **biological properties**. They also can have **attributes from experimental measurements** such as fold change in abundance or p-value. The user needs to identify entities that have or do not have some specific properties or attributes.

3. **Select portions of the metabolic system both by their relations and by their properties or attributes.**
   Combining selection criteria can be much more specific. In particular, the user might have a list of entities with attributes. The user might only have interest in entities with certain attributes from this list, for which they need information and context.

4. **Refine selection by inclusion or exclusion of individual entities.**
   The user needs to define very specific selections. Direct selection of individual entities might be necessary if more general selections are too ambiguous.

Exploration Requirements

Metabolism is a complex system in that many types of entities and relations have many types of properties. There is a lot of information of different types. The user needs to explore selections of the metabolic network in different ways to access different types of information according to her/his interest.

1. **Display summary information for the entire metabolic system and for the selection.**
   The user needs a **concise orientation to the context of the system**. The user also needs to compare the smaller selection to the larger system.

2. **Represent selections of the metabolic system clearly.**
   The user needs to select any individual entity in the representation to either access information in more detail or to modify this entity's representation. The user needs to recognize both near and distant relations between entities clearly. The user needs to trace paths of relations between entities easily for both near and distant relations. The representation must be able to visualize different properties of the entities in the context of their relations. The representation must be automatic and robust in order to accommodate custom selections. The representation must also be dynamic to allow the user to interact with entities and to enable the user to change the representation to illustrate different properties or attributes.

3. **Associate entities clearly with their properties or attributes.**
   Properties of entities are important aspects of the metabolic system. The user needs to recognize a variety of these properties easily. Also, the relations and properties of the metabolic system give important context for attributes from experimental measurements. The user needs to analyze these attributes within the context of the metabolic system.

4. **Display information in detail for entities in response to user interest and interaction.**
   Even if it is just for a single entity or a few entities from the selection, the user needs to access some information in detail. This information can be of different types. The user needs some way to specify and access this information of interest.

Practical Requirements

1. **Accommodate custom models of metabolic systems.**
   Metabolic systems vary in different species and even in different tissues or conditions in the same species. The user needs to use this tool to study her/his custom model of the metabolic system.

2. **Maximize accessibility of the tool.**
   The user will be unlikely or unable to use a tool that requires installation of custom programs or packages. The user will also be unlikely or unable to use a tool that requires computational resources beyond those of a typical personal computer.

3. **Simplify the user interface.**
   The user is familiar with and able to use common programs with graphical user interfaces such as office programs and internet browsers. The user has little knowledge or skill in computational programming. The user will be unlikely or unable to use a tool that uses a script-like interface. The tool needs to have a simple and attractive graphical user interface. User interaction with this interface must be intuitive. The user is unlikely to consult documentation for the tool.

4. **Minimize the need for long-term maintenance.**
   There is no guarantee that this tool will have support for long-term maintenance. The tool needs to be sufficiently functional and robust without maintenance for a period of about 5 years.

Current Technology

This section includes a concise depiction of the current technology that is relevant to this project with summaries and evaluations of a set of tools. This set of tools is not comprehensive.

Category	Property	Value
Compartment	Count	**9**
Metabolite	Count of chemically unique	**2652**
	Count in extracellular space	642*
	Count in cytoplasm	1878*
	Count in mitochondrion matrix	500*
	Count in mitochondrion intermembrane space	250*
	Count in nucleus	165*
	Count in endoplasmic reticulum	570*
	Count in peroxisome	435*
	Count in lysosome	302*
	Count in golgi apparatus	317*
	Count of total compartmental	**5324**
Reaction	Count of reactions	**7785**
	Count of genes	**1675**
	Count of transcripts	2194*

Table 2: Specifications of the **Recon 2.2 model of human metabolism**[10,11]. Symbol "*" indicates approximate values.

Databases

Many databases provide information about biological entities that is relevant to the metabolic system. The Universal Protein Resource (UniProt, http://www.uniprot.org/)[12] provides information about proteins, their properties and functions. The Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/)[13], and the MetaCyc Metabolic Pathway Database (http://www.metacyc.org/)[14], provide information about metabolites and reactions.

Metabolic Models

Systems biology uses computational models[15,16,17] to study the vastness and complexity of biological systems. These models are compilations of current knowledge. Relevant information comes from annotations of genes in genomes, from databases about biological entities, and from many biochemical and molecular biological studies that characterize the functions of individual genes, transcripts, proteins, and metabolites. **Metabolic models** are **collections of all known reactions**. These reactions include information for **reactant and product metabolites** with appropriate **stoichiometry to balance mass and charge**. They include information about **rates, directionality, and reversibility**. They also include information about **cellular compartments and transport** of metabolites between compartments. In cases of catalysis or transport, they include information about the **genes that encode the transcripts and proteins** to facilitate the process. Metabolic models are **specific to species, tissues, and even conditions**. Active communities of experts develop, curate, and maintain models that are available from multiple repositories[18,19,20]. Robust models are

available for many organisms, including *Saccharomyces cerevisiae* (yeast), *Mus musculus* (mouse), and *Homo sapiens* (human). The most current model for human metabolism is Recon 2.2[10,11] (**Table 2**).

Tools are also available for managing these metabolic models. The Systems-Biology-Markup-Language (SBML) (http://sbml.org) relates to the Extensible-Markup-Language (XML) and is an open standard and format for the representation of computational models for biological systems. libSBML[21] (http://sbml.org/Software/libSBML) is a reference for interpreting information in SBML format. COBRApy[22] (https://opencobra.github.io/cobrapy/) is an open-source tool for managing metabolic models in Python. libSBML and COBRApy together are useful to convert information for metabolic models from SBML format to JavaScript Object Notation (JSON) format.

Visualization and Exploratory Analysis

KEGG Atlas (http://www.kegg.jp/kegg/atlas/)[13] is a web application that enables users to explore metabolic maps from KEGG. The user can select from a large set of available maps for standard subsets and pathways in metabolism. It is not possible to view maps for custom subsets of metabolism. Rather it is necessary to search through multiple separate maps for reactions of interest. Maps in the atlas use a static, spatial layout that arranges reactions by process or pathway. This static layout distorts relations between metabolites. Two metabolites may be far apart on the map although they relate by a single reaction. It is not possible to change the map's layout to represent other properties such as cellular compartment. The map only represents the priority metabolites of each reaction, leaving out reactants and products that might be of interest. Also it is not obvious where the map represents a single metabolite in multiple places. It is only possible to search the map by identities of metabolites or reactions. It is not possible to search by relations or properties. It is possible to change the colors of individual icons for metabolites. It is not possible to vary color saturation of icons to represent quantitative attributes. This design renders KEGG Atlas inappropriate for visual exploration of **custom subsets of the metabolic network that traverse standard pathways**.

Escher[23] (https://escher.github.io/) is a web application that enables users to draw metabolic maps and visualize experimental data on these maps. It imports information from metabolic models. The user selects individual reactions from these models to include in the maps. It is not possible to query the metabolic model by relations or properties, so the user must know the identities of metabolites and reactions of interest. The user manually positions the individual reactions and their metabolites to create a static map. The user can arrange reactions and metabolites however she/he wants to represent sets and create a map that is readable. The user can change the color of individual icons for metabolites or reactions to represent properties or quantitative attributes. This design renders Escher inappropriate for **efficient, dynamic exploration of custom portions of the metabolic system**.

Features

The **metabolic system is a network** of relations between biological entities. In reality, this network has multiple dimensions or layers[24,25] since there are many types of relations between many types of entities (**Table 1**). The user is familiar with a simpler model, in which **metabolites are entities, and reactions are relations between them**. This project will use a model that is conceptually similar but more accurate and capable. Rather than treating reactions as relations, this project will treat reactions as special entities since they involve the functions of multiple other entities (**Table 1**). Hence, **metabolites are entities, and reactions are entities that facilitate relations between metabolites**. In terms of network theory, **metabolites and reactions are distinct types of nodes**, and **reactant and product metabolites have links to their reactions**. This network is **sparse**, and the relations between metabolites and reactions are **many-to-many** (such that the count of links is much greater than the counts of metabolites and reactions). This project will assume this model of metabolism and will apply principles of network and graph theory to support its study.

Multiple categories of information are relevant to this study of the metabolic network. The **topology or connectivity** of the network describes the **metabolites that participate as reactants or products in**

Source	Category	Example
Community	**Topology** (connectivity of a network)	Count and identities of metabolites as **reactants** in a reaction
		Count and identities of metabolites as **products** in a reaction
		Degree of a metabolite (count of reactions in which it participates)
		Continuous **paths** of multiple metabolites and reactions
		Betweenness centrality of a metabolite (count of shortest paths that pass through it between all other metabolites in network)
	Property of Metabolite	**Identity**
		Name
		Chemical formula
		Charge
		Cellular compartment
	Property of Reaction	**Identity**
		Name
		Rate
		Reversibility
		Type (chemical reaction or transport event)
		Cellular compartment(s)
		Metabolic pathway(s)
		Identities and properties of **constituent genes, transcripts, and proteins**
		Cofactors of proteins
User	**Attribute** (experimental measurement)	**Fold change** in study condition relative to a control condition
		P-Value for comparison between study condition and control condition

Table 3: Categories of information about metabolic network.

reactions as well as **paths across multiple reactions**. The topology of the metabolic network is complex because **some metabolites participate in many reactions**, and **path loops are also common**. In addition, both metabolites and reactions in the metabolic network have many **properties**. Metabolites and reactions that share the same properties belong to **sets**, such as those in the same cellular compartment. Together, this **information about topology and properties describes the metabolic network** and is available in metabolic models from the systems biology community. For development and demonstration of the tool, this project will use the **most recent model of general human metabolism, Recon 2.2**[10,11] (**Table 2**). The tool will also maintain compatibility for use with metabolic models for other species, tissues, or conditions. If the user desires, it will also be possible to supplement the properties in these models with information from databases. Another category of information includes **attributes of entities from experimental measurements**, which are often specific to experimental conditions. This information about attributes comes from the user. **Table 3** lists these categories of information along with examples of each.

This project will develop computational methods and a tool to help the user study the metabolic network. The tool will access information for a **model of the entire metabolic network** of the species of interest. This model will not be comprehensive and depends on the availability of information. Still the model describes a vast and complex network. The tool will enable the user to **select a portion of the metabolic network that is of interest**. The tool will then **provide information to the user about this selection from the model**. This project will apply principles of **data visualization**[26,27,28] to communicate this information to the user in an **interactive, visual interface**. Information about the topology and properties of the metabolic network will give the user context to **design experiments effectively**. Associating information about attributes with this metabolic network will give the user context to **interpret experimental results accurately**.

Selection Features

The user will **select portions (subnetworks) of the metabolic network (model) according to her/his interest**. The tool will provide an **interactive, graphical interface** to help the user to assemble **queries by topology and properties**. The tool will also assemble queries from **tables of entities with attributes** that the user provides. Queries against the metabolic network will return a **list of subnetworks that satisfy the selection criteria**. The user will then select between items in this list of subnetworks to **explore either separately or simultaneously**. The user will also refine the subnetwork interactively by selecting individual metabolites and reactions. **Figure X** illustrates the design for the **Query Interface** that supports this functionality.

1. **Select metabolites and reactions by topology.**
   This type of selection depends only on the **topology** of the metabolic network around or between entities of interest. It is independent of their properties. The user will specify a set of entities of interest. These can be metabolites or reactions, or they can be the genes, transcripts, or proteins that contribute to reactions. As the user adds each entity to the set, the tool will suggest matching entities from the metabolic network dynamically.

   Selections by topology will have the option to ignore certain categories of metabolites. Some metabolites are extremely pervasive in metabolism and participate in many more reactions than other metabolites. Examples include water, proton (H+), carbon dioxide, nicotinamide adenine dinucleotides (NAD+, NADH, NADP+, NADPH), flavin mononucleotides (FMN, FMNH2), and adenosine phosphates (AMP, ADP, or ATP). Nodes for these metabolites have **very high degrees in the metabolic network, imparting excessive connectivity**. If these metabolites are less interesting to the user, then the user can opt to ignore them in selections by topology.

   - **Select metabolites and reactions within a specific proximity to a single entity or to multiple entities.**
     As **Figure X** illustrates, **proximal selections** return all metabolites and reactions that are within

a specific **path length** in all directions from a specific metabolite or reaction. These selections to multiple entities (or to a single entity in multiple instances, such as the same protein in multiple reactions) are likely to be multiple discontinuous subnetworks. Proximal selections give the user information about the **local context of individual metabolites and reactions**.
Example: Select all metabolites and reactions within 1 reaction of pyruvate.
Example: Select all metabolites and reactions within 1 reaction of gene 1, gene 2, and gene 3.

- **Select metabolites and reactions in the shortest paths between multiple entities.**
  As **Figure X** illustrates, selections by **shortest path** return all metabolites and reactions that are part of the shortest path(s) between multiple metabolites or reactions. Shortest path selections give the user information about **how multiple metabolites or reactions relate to each other**.
  Example: Select all metabolites and reactions in the shortest path(s) between glucose and pyruvate.

2. **Select metabolites and reactions by property or attribute.**
   This type of selection depends only on the properties or attributes of entities in the metabolic network. It is independent of topology. The user will specify a set of properties to include or exclude from the selection. The tool will inform the user of the properties and attributes that are available for the metabolic network.

   - **Include in selection metabolites and reactions that have some property or attribute.**
     Selections by inclusion criteria include only entities that have some property or attribute.
     Example: Select all transport reactions and their metabolites.
     Example: Select all transport reactions and their metabolites between the cytoplasm compartment and the mitochondrion compartment.
   - **Exclude from selection metabolites and reactions that have some property or attribute.**
     Selections by exclusion criteria include all entities except those that have some property or attribute.
     Example: Select all metabolites and reactions that are not in the cytoplasm compartment.

3. **Select metabolites and reactions by combinations of criteria for topology and property or attribute.**
   This type of selection may depend on both topology and property or attribute. It has the most ability to select specific subnetworks. The sequence or hierarchy of selection criteria matters. This type of selection is appropriate to define queries from a table of entities with attributes that the user provides.

   - **Reduce a set of entities by property or attribute before selecting metabolites and reactions by topology.**
     This combination of selection strategies reduces or filters a set of entities by inclusion or exclusion criteria of properties and attributes. Then it returns metabolites and reactions by topology without further consideration of properties or attributes.
     Example: Select all metabolites and reactions that are within 1 reaction of all metabolites in the set that have a fold change greater than 1.5 and a p-value less than 0.01.
     Example: Select all metabolites and reactions that are in the shortest paths between all metabolites in the set that have a fold change greater than 1.5 and a p-value less than 0.01.
     Example: Select all metabolites and reactions that are in the shortest paths between all metabolites in the set that have a fold change greater than 1.5 and a p-value less than 0.01 and are in the mitochondrion compartment.
   - **Reduce a set of metabolites and reactions by property or attribute after selecting these metabolites and reactions by topology or by property or attribute.**
     This combination of selection strategies selects metabolites and reactions by topology or by property or attribute. Then it reduces these by inclusion or exclusion criteria of properties and attributes.
     Example: Select all metabolites and reactions that are within 1 reaction of pyruvate and are in the

mitochondrion compartment.
Example: Select all reactions and their metabolites that are in the mitochondrion compartment and do not involve proteins with iron-sulfur cofactors.

Exploration Features

The user will explore information about the metabolic network (model) and the subnetwork (selection) in different ways according to her/his interest. **Table 3** lists categories of information about the metabolic network. The tool will provide an **interactive, graphical interface** to communicate this information to the user visually. **Figures X-Z** illustrate the design for the **Detail, Navigation, and Exploration Interfaces** that will support this functionality.

1. **Display summary information for the model and for the selection.**
   Summary information orients the user to a **concise context** of both the model and the selection. It also allows the user to compare the two networks. Multiple different types of information about the network contribute to this summary. **Figure X** illustrates the design for the **Detail Interface** that will support this functionality.

   - **Display basic, descriptive information about the model and the selection.**
     This information includes the name of the model and the counts of metabolites, reactions, and compartments in both the model and the selection.
   - **Display the distribution of degrees of metabolites in the model and in the selection.**
     In a network, the degree of a node is a measure of the local relevance of the node in the network. It is the count of links in which that node participates directly. Every node in the network has a degree. The distribution of degrees in a network is the proportion of nodes with each value. This distribution conveys some idea for the connectivity of the network.
   - **Display the distribution of betweenness centralities of metabolites in the model and in the selection.**
     In a network, the betweenness centrality of a node is a measure of the global relevance or centrality of the node in the network. It is the count of shortest paths between all other nodes in the network that pass through the node. Every node in the network has a betweenness centrality. The distribution of betwenness centralities in a network is the proportion of nodes with each value. This distribution conveys some idea for the connectivity of the network.
   - **Display the distribution of entities between sets in the model and in the selection.**
     If multiple entities have the same property, then that property defines a set of those entities. Any property can define a set. Likewise the set for any property might be interesting to the user. The distribution of entities between sets is the proportion of entities in each set or combinations of sets for a specific type of property. This distribution conveys some idea for the structure of the network with respect to a specific property.
     Example: Display the proportion of metabolites in the model or selection that occur in each compartment (cytoplasm, mitochondrion, peroxisome, etc).
     Example: Display the proportion of metabolites in the model or selection that occur in each combination of multiple compartments (cytosol and mitochondrion, mitochondrion and peroxisome, etc).
   - **Display relations between sets in the model and in the selection.**
     Sets can have relations between themselves. These relations can be members that multiple sets share, or they can be relations between the members of separate sets.
     Example: Display the count of metabolites that occur simultaneously in the cytoplasm and mitochondrion compartments.

Example: Display the count of transport reactions between the cytoplasm and mitochondrion compartments.

2. **Represent topology of the metabolic network.**
   The topology of relations between entities in the metabolic network is a priority type of information, and the tool will communicate this information to the user. **Figure X** illustrates the design for the **Exploration Interface** that will support this functionality.

   - **Represent metabolites and reactions in an explicit diagram of nodes and links.**
     An explicit visual diagram of nodes and links is the most appropriate representation for the metabolic network. The versatility of this representation accommodates the complexity of the network and represents the relevant information effectively. It provides distinct nodes and links for individual metabolites and reactions that the user can select and manipulate. Nodes for metabolites will be boxes with labels, nodes for reactions will be arrows with labels, and links will be linear connections between nodes. The relative positions of these nodes and links depend on their relations to each other such that the topology of the network is intuitive. It portrays both near and distant relations between entities clearly, making path tracing intuitive and loops clear.
   - **Use an automatic and robust algorithm for the layout of the node-link diagram.**
     A force-directed layout is appropriate to represent the node-link diagram of the metabolic network. This layout algorithm simulates repulsive and attractive forces between nodes and links to assign their positions and orientations in a clear layout. This layout algorithm is sufficiently automatic and robust to accommodate many different custom selections from the model of the metabolic network.
   - **Support dynamic interaction with the representation of the metabolic network.**
     Through dynamic interaction, the user will explore the topology of the metabolic network to understand it better. The user will select and interact with the nodes and links for individual metabolites and reactions in the diagram. In this way it will be possible to **refine the selection from the model of the metabolic network**. The user may choose to exclude individual metabolites or reactions from the diagram. The user may also choose to include additional metabolites or reactions in the diagram that relate by proximity to existing metabolites or reactions. It will also be possible to **refine the visual representation of the metabolic network in the diagram**. Adjusting the positions of nodes and links might improve readability. Also, nodes from some metabolites have very high degrees. If these metabolites are less interesting to the user, the user can replicate their nodes to simplify the network and improve readability. As **Figure X** illustrates, the **Navigation Interface** also displays a list of metabolites with high degrees from which the user can control node replication.

3. **Represent properties and attributes of the metabolic network in context of the network's topology.**
   It will be useful to analyze properties and attributes of the metabolic network in their true context of the network's topology. It is possible to represent some of this information directly on the node-link diagram. This explicit representation of the metabolic network offers favorable visual channels to represent multiple types of information. Using these visual channels, it will be possible to represent different types of information on the network diagram, keeping the information in proper context of the network's topology. **Table 3** lists some examples of properties and attributes of metabolites and reactions. Nodes of metabolites will represent properties or attributes for metabolites. Nodes of reactions will represent properties or attributes for reactions or their constituent genes, transcripts, and proteins. For the sake of clarity, there is a limit to the count of different types of properties and attributes that the diagram can represent simultaneously. The user will interactively select the properties and attributes to represent on the diagram as well as how to represent them. As **Figure X** illustrates, the **Navigation Interface** will

support this functionality of controlling the representation of properties and attributes in the node-link diagram.

- **Represent categorical properties and attributes of metabolites and reactions.**
  Categorical properties and attributes have a nominal measurement scale. They are discrete and are neither quantitative nor continuous. If multiple metabolites or reactions possess the same property or attribute, then this property or attribute defines a set or group. It is useful to recognize the metabolites and reactions that belong to each categorical set on the network diagram. **Positional groups** of nodes and links represent these sets clearly. **Color hue** (red, blue, green, etc) of nodes also represents these sets clearly.
  Example: Arrange nodes and links for metabolites and reactions in separate positional groups according to their compartments.
  Example: Set the colors of nodes for reactions according to their metabolic pathways.
  Example: Set the color of nodes for reactions to red if they involve the function of a protein with an iron-sulfur cluster.
  Example: Set the color of nodes for reactions to red if they transport metabolites between the cytoplasm and mitochondrion compartments.
  Example: Set the color of nodes for reactions to red if they involve the function of a transcript with a fold change in abundance greater than 1.5 and a p-value less than 0.01.
- **Represent quantitative properties and attributes of metabolites and reactions.**
  Quantitative properties and attributes have either an ordinal, interval, or ratio scale. **Color luminance or saturation** of nodes represents these properties or attributes clearly.
  Example: Set the color saturation of nodes for metabolites according to their fold change in abundance (red for increase, blue for decrease).
  Example: Set the color saturation of nodes for metabolites according to their p-value.

4. **Display information in detail for specific metabolites and reactions that the user selects.**
   There is a limit to the count of properties or attributes that the node-link diagram can represent simultaneously. There is also a limit to the clarity of this representation on the node-link diagram. For example, small differences in quantitative properties or attributes will not be very clear by color saturation, especially when the relevant nodes are far apart in the node-link diagram. Also, **individual entities have properties and attributes in the context of the entire model**. It might be most clear to represent these properties and attributes separate from the node-link diagram for the subnetwork selection. The user will select individual metabolites or reactions to view their properties and attributes in more detail. The user can select multiple metabolites or reactions to view their properties and attributes simultaneously. **Figure X** illustrates the design for the **Detail Interface** that will support this functionality.

   - **Represent properties and attributes of metabolites and reactions.**
     It will be more practical or clear to represent some properties or attributes separate from the node-link diagram. Properties or attributes of individual entities within the entire model will be more clear in this context. For example, it might be interesting to the user to compare the degrees or betweenness centralities of a specific metabolite in different cellular compartments. This comparison could convey the respective relevance of the metabolite in each compartment.
     Example: Display the betweenness centralities of pyruvate in the entire metabolic network and in the subnetwork of the mitochondrion compartment.
   - **Represent properties and attributes of constituent entities.**
     Reactions involve the functions of multiple entities including genes, transcripts, proteins, and metabolites. Each of these entities can have properties and attributes.
     Example: Display the fold change in abundance of a specific transcript in multiple samples.

- **Represent relations between individual entities and other entities.**
  It might be useful to represent the relations between entities in a way other than the node-link diagram. This representation might especially be useful to represent relations to individual entities in the entire model.
  Example: Display the count of reactions in which pyruvate participates along with each of its most common metabolite neighbors in the model.

- **Display relations between individual entities and the sets to which they belong.**
  Relations between individual entities and multiple sets to which they belong might be interesting. These set relations might be in the context of the entire model.
  Example: Display the count of reactions in which pyruvate participates in each cellular compartment.

Practical Features

These features have to do with the implementation of the tool.

1. **Derive information from custom metabolic models.**
   The tool will be able to use information for **custom metabolic models**. The metabolic system varies for different species, tissues, and conditions. The user will need to study a model for the metabolic system of her/his interest. The user might also want to customize the model to include additional information about entities, such as properties of proteins. This project will maintain compatibility with the standard type and structure of information in metabolic models. In this way the tool will accommodate custom models and revisions to existing models.

   For development and demonstration, this project will use information from the most recent model of human metabolism, Recon 2.2[10,11] (**Table 2**). As the product of extensive community contribution, this model is publicly available and does not have a license. The original version of Recon 2.2 is in SBML format, and is 24 Megabytes in size. I will use libSBML[21] and COBRApy[22] to convert information from the model to JSON format. In this more concise format, I estimate that the model will only be 4 Megabytes in size. The JSON format will also be convenient for use in the JavaScript program.

   In the tool, the user will choose a metabolic model to use. I will make Recon 2.2 in JSON format available for download. I will also provide instructions for the user to convert her/his own metabolic model to JSON format to use in the tool.

2. **Implement the tool as an entirely client-side web application.**
   The tool will run entirely through the client's internet browser on the client's computer. The user will not need to install any custom programs on her/his machine. Also, it will not be necessary to establish and maintain a server to host the tool. The client-side web application without a server back-end will be much simpler to implement and will require much less maintenance.

   The tool will access information for the metabolic model. The model for development and demonstration in this project is Recon 2.2, which will be about 4 Megabytes in size in JSON format. Including this information as part of the web application would require it to download every time the user refreshes her/his browser. If the metabolic model is too large then it will cause the browser to refresh slowly. If necessary, then I will include a smaller subset of the metabolic network from Recon 2.2 as a demonstration model that comes with the web application. I will allow the user to download the complete Recon 2.2 model in JSON format to her/his computer. Then the user will select whether to run the tool using the model in the web application or a model (including her/his own custom models) from her/his computer.

Implementation Schedule

This section gives some detail on the steps that will be necessary to implement this project. These steps are very much not comprehensive. Many additional steps will become obvious during implementation. It is difficult to anticipate the duration of each step, since many problems might arise. Still, this project has a main goal. This project will implement a functional prototype of the tool that is suitable for trials with users by June 2017 (6 months time from preparation of this document).

1. **Organize information from the metabolic model.**
   It will be necessary to convert the information of the metabolic model from SBML format to JSON format in order to access it conveniently with a JavaScript program.

2. **Develop methods to interpret the metabolic model as a network.**
   The metabolic model provides collections of metabolites and reactions with properties. These are the nodes of the network. The reactions also have properties that designate their reactant and product metabolites. This information describes the links of the network. It will be necessary to develop custom methods to interpret this information from the model as a network. For example, methods need to navigate between metabolites and reactions. It will also be necessary to develop methods to **derive additional properties** from the network. Counts of nodes and links and degrees and betweenness centralities of metabolite nodes are examples. It will also be necessary to derive a property to distinguish between transport reactions and chemical reactions. Some tools may already be available to provide some of this functionality for analyzing the network. In order to be useful, tools will need to be packages or libraries for JavaScript.

3. **Develop methods to query the network.**
   It will be necessary to query the network by topology and properties. It will be necessary to develop appropriate methods for these queries. Some tools may already be available to provide this functionality. In order to be useful, tools will need to be packages or libraries for JavaScript.

4. **Design and develop the interface for queries against the network.**
   The **Query Interface** will enable the user to construct and execute queries and to select between subnetworks that these queries return. The program's module for the **Query Interface** will support this functionality. It will assemble and complete queries against the network of the metabolic model. These queries will return subsets of the metabolic model. The interface will display a list of these to the user. When the user selects an element from the list, the module will pass the subset of the metabolic model to the program's modules for the **Detail Interface**, the **Navigation Interface**, and the **Exploration Interface**. These modules may modify copies of the subsets from the **Query Interface**. They will not modify the original subsets of the metabolic model from the module for the **Query Interface**. These original subsets will persist so that the user can revert.

5. **Design and develop the interface and visualizations for summary and detail information about the network and individual entities.**
   The **Detail Interface** will display to the user various properties and attributes about the entire model of the network, selections of subnetworks, or individual entities. The program's module for the **Detail Interface** will support this functionality. It will derive additional properties from the network or subnetwork (counts of nodes and links, degrees, betweenness centralities). It will also construct the appropriate visual representations.

6. **Design and develop the interface for changing the visual representation of properties and attributes on the network.**
   The **Navigation Interface** will enable the user to control the visual representation of the subnetwork.

It will also enable the user to control the representation of properties and attributes on the visual representation of the subnetwork. The program's module for the **Navigation Interface** will support this functionality.

7. **Develop and refine the visual representation of the network.**
   The **Exploration Interface** will represent the subnetwork along with any properties or attributes visually. The program's module for the **Exploration Interface** will support this functionality.

References

[1] Agatha A. van der Klaauw and I. Sadaf Farooqi. "The hunger genes: pathways to obesity". In: *Cell* 161.1 (Mar. 26, 2015), pp. 119–132. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.03.008.

[2] David W. Haslam and W. Philip T. James. "Obesity". In: *Lancet (London, England)* 366.9492 (Oct. 1, 2005), pp. 1197–1209. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(05)67483-1.

[3] Rui-Sheng Wang, Bradley A. Maron, and Joseph Loscalzo. "Systems medicine: evolution of systems biology from bench to bedside". In: *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 7.4 (Aug. 2015), pp. 141–161. ISSN: 1939-005X. DOI: 10.1002/wsbm.1297.

[4] Xiang Zhang, Jan A. Kuivenhoven, and Albert K. Groen. "Forward Individualized Medicine from Personal Genomes to Interactomes". In: *Frontiers in Physiology* 6 (2015), p. 364. DOI: 10.3389/fphys.2015.00364.

[5] Kirk Beebe and Adam D. Kennedy. "Sharpening Precision Medicine by a Thorough Interrogation of Metabolic Individuality". In: *Computational and Structural Biotechnology Journal* 14 (2016), pp. 97–105. DOI: 10.1016/j.csbj.2016.01.001.

[6] M. Benson. "Clinical implications of omics and systems medicine: focus on predictive and individualized treatment". In: *Journal of Internal Medicine* 279.3 (Mar. 2016), pp. 229–240. ISSN: 1365-2796. DOI: 10.1111/joim.12412.

[7] Abdellah Tebani et al. "Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations". In: *International Journal of Molecular Sciences* 17.9 (Sept. 14, 2016). ISSN: 1422-0067. DOI: 10.3390/ijms17091555.

[8] Francis S. Collins and Harold Varmus. "A new initiative on precision medicine". In: *The New England Journal of Medicine* 372.9 (Feb. 26, 2015), pp. 793–795. ISSN: 1533-4406. DOI: 10.1056/NEJMp1500523.

[9] Eric J. Topol. "Individualized medicine from prewomb to tomb". In: *Cell* 157.1 (Mar. 27, 2014), pp. 241–253. ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.02.012.

[10] Ines Thiele et al. "A community-driven global reconstruction of human metabolism". In: *Nature Biotechnology* 31.5 (May 2013), pp. 419–425. ISSN: 1546-1696. DOI: 10.1038/nbt.2488.

[11] Neil Swainston et al. "Recon 2.2: from reconstruction to model of human metabolism". In: *Metabolomics: Official Journal of the Metabolomic Society* 12 (2016), p. 109. ISSN: 1573-3882. DOI: 10.1007/s11306-016-1051-4.

[12] UniProt Consortium. "UniProt: a hub for protein information". In: *Nucleic Acids Research* 43 (Database issue Jan. 2015), pp. D204–212. ISSN: 1362-4962. DOI: 10.1093/nar/gku989.

[13] Minoru Kanehisa et al. "KEGG as a reference resource for gene and protein annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D457–462. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1070.

[14] Ron Caspi et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D471–480. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1164.

[15] Ines Thiele and Bernhard Ø Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction". In: *Nature Protocols* 5.1 (Jan. 2010), pp. 93–121. ISSN: 1750-2799. DOI: 10.1038/nprot.2009.203.

[16] Edward J. O'Brien, Jonathan M. Monk, and Bernhard O. Palsson. "Using Genome-scale Models to Predict Biological Capabilities". In: *Cell* 161.5 (May 21, 2015), pp. 971–987. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.05.019.

[17] Aarash Bordbar et al. "Constraint-based models predict metabolic and associated cellular functions". In: *Nature Reviews. Genetics* 15.2 (Feb. 2014), pp. 107–120. ISSN: 1471-0064. DOI: 10.1038/nrg3643.

[18] Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. "Human metabolic atlas: an online resource for human metabolism". In: *Database: The Journal of Biological Databases and Curation* 2015 (2015), bav068. ISSN: 1758-0463. DOI: 10.1093/database/bav068.

[19] Sébastien Moretti et al. "MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D523–526. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1117.

[20] Zachary A. King et al. "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D515–522. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1049.

[21] Benjamin J. Bornstein et al. "LibSBML: an API library for SBML". In: *Bioinformatics (Oxford, England)* 24.6 (Mar. 15, 2008), pp. 880–881. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn051.

[22] Ali Ebrahim et al. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". In: *BMC systems biology* 7 (Aug. 8, 2013), p. 74. ISSN: 1752-0509. DOI: 10.1186/1752-0509-7-74.

[23] Zachary A. King et al. "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". In: *PLoS computational biology* 11.8 (Aug. 2015), e1004321. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004321.

[24] S. Boccaletti et al. "The structure and dynamics of multilayer networks". In: *Physics Reports* 544.1 (Nov. 2014), pp. 1–122. ISSN: 03701573. DOI: 10.1016/j.physrep.2014.07.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S0370157314002105 (visited on 10/27/2016).

[25] Manlio De Domenico et al. "Structural reducibility of multilayer networks". In: *Nature Communications* 6 (Apr. 23, 2015), p. 6864. ISSN: 2041-1723. DOI: 10.1038/ncomms7864.

[26] Nils Gehlenborg et al. "Visualization of omics data for systems biology". In: *Nature Methods* 7.3 (Mar. 2010), S56–68. ISSN: 1548-7105. DOI: 10.1038/nmeth.1436.

[27] Corinna Vehlow et al. "Visual analysis of biological data-knowledge networks". In: *BMC bioinformatics* 16 (Apr. 29, 2015), p. 135. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0550-z.

[28] Georgios A. Pavlopoulos et al. "Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future". In: *GigaScience* 4 (2015), p. 38. DOI: 10.1186/s13742-015-0077-2.