A Web-Based Data Source for Metabolomics

Ali Cakmak[1] cakmak@case.edu
Arun Dsouza[1] dsouza@case.edu
Richard Hanson[2] rwh@case.edu
Gultekin Ozsoyoglu[1] tekin@case.edu
Meral Ozsoyoglu[1] meral@case.edu

[1]Department of Electrical Engineering and Computer Science
[2]Department of Biochemistry
Case Western Reserve University
Cleveland, OH 44106, USA

Abstract—**With the development of improved and cost-effective technologies, it is now possible to detect thousands of metabolites in biofluids or specific organs, and reliably quantify their amounts. Metabolomics focuses on studying the concentrations of metabolites in a cell or a tissue. In this paper, we describe a prototype web-based metabolomics data analysis system, PathCase^{MAW} (PathCase Metabolomics Analysis Workbench), which features (1) A web-accessible metabolic pathway database that supports online browsing and querying, and is novel in that it includes tissue and subcompartment information for pathways, and models transport processes, (2) Tissue-aware visualization support for viewing processes, pathways, or groups of pathways in PathCase^{MAW} database, and (3) An online metabolomics data analysis tool, called Automated Consequence Prediction Tool, which allows users to upload their own observed/measured metabolite level changes, and computationally invalidates or M(aybe)-validates those biological mechanisms that produce the observed metabolite changes.**

Keywords – metabolomics; biological web databases; metabolic networks; metabolism; biochemical pathways; bioinformatics.

I. INTRODUCTION

Metabolomics is the systematic study of the distributions (profiles, concentrations) of small-molecular-weight substances in cells, tissues and/or whole organisms as influenced by multiple factors including genetics, diet, lifestyle, and pharmaceutical interventions. The metabolome refers to the complete set of small-molecule metabolites in a cell or a tissue. Metabolites are ideal for monitoring dynamic behavior and cellular mechanisms in a biological system [1], as metabolite concentrations are highly sensitive to changes at the gene expression [2] level. Hence, metabolome analysis is performed in order to take "screenshots" of an organism under different conditions for a differential study.

Metabolomics analysis relies on interpreting the biological importances of measured values of (significant numbers of) identified chemicals within bio-fluid samples. The ability to understand the data in a biochemical context can and does yield insights into the mechanisms and biological functions involved in any experimental condition. For example, with metabolomics analysis, it is possible to understand the defect of a target enzyme, receptor, or signaling system through biochemical pathways analysis of the precursors and products of measured metabolites. In this context, biochemical reactions operate as a network of changes rather than a linear set of reactions. The biochemical approach to data interpretation has the potential to yield highly relevant hypotheses that can be directly tested. Another very powerful outcome is that such a technology can make use of more than a century's worth of detailed biochemical experience. When metabolomics data is viewed in biochemical context, the interpretation is often immediate, likely to be highly relevant and - in hindsight- obvious [3].

With recent advances in experimental technologies, it is now possible to measure large numbers of metabolites in body fluids. However, such measurements are only useful to the extent that they can be interpreted with projections onto the underlying cellular mechanisms described by the biochemical networks. Manually going through such networks with the goal of deriving consistent conclusions suggested by the measurements requires expert knowledge, and is costly in terms of the required time and effort. Furthermore, manual interpretation efforts usually focus on a limited number of commonly used biomarkers, and cannot take advantage of the large numbers of available measurements due to the breath and complexity of metabolic networks. Thus, computational platforms that can help researchers interpret metabolomics measurements, and automatically eliminate the unlikely hypotheses, are highly desirable and useful.

This paper describes PathCase^{MAW} (PathCase Metabolomics Analysis Workbench -- available [1] as a prototype at this point) which has the following features:

1. *A web accessible metabolic pathways database* which (i) features biological compartments (i.e., tissue/organ, cell, etc) for each pathway, and (ii) models transport processes that carry metabolites from bio-fluids to organs and vice versa. Currently our database is populated by manually collecting and entering major pathways from the literature. While there many are web-based metabolic network data sources e.g., KEGG [4], Reactome [5], MetaCyc [6], PathCase [7], PATIKA [14], etc., all such data sources lack location information for individual pathways. And, without location information, different parts of the metabolism are incorrectly inter-connected.
2. *A tissue-aware visualization framework* that can be used to visualize processes, pathways, or groups of pathways. PathCase^{MAW} extends the PathCase [7] pathway visualization framework, in order to incorporate the tissue location information associated with pathways so as to create a more accurate view of the relevant biological mechanisms.
3. *An Observed Metabolite Analysis (OMA) Tool* with a web interface for automated prediction of biological mechanisms. OMA automatically extracts sets of likelihood scenarios which we call hypotheses, via "incremental change models" [8, 9] from (i) the latest metabolic network knowledge captured in its database, and (ii) observed metabolite concentration levels in a given bio-fluid data. The tool helps with physiological condition

[1] http://dblab.case.edu/PathwaysMetabolomics/Web

978-1-4244-5023-7/09/$25.00 ©2009 IEEE

ISCIS September 14-16, 2009
METU
Northern Cyprus Campus

prediction/analysis, and contains multiple metabolic assessment models where, given a set of observations in a bio-fluid data, possible scenarios as hypotheses are enumerated in either complete or summarized form, and ranked based on their likelihood.

Given a set of observations on metabolite concentration level changes, the OMA tool enables researchers to automatically invalidate some of the possible hypotheses as causes for the observed relative metabolite concentration changes. To this end, the OMA tool utilizes the latest metabolic network knowledge and rules that capture the causal effects of metabolic changes on other selected metabolites and on physiological changes.

The OMA tool of PathCase^{MAW} also lays the framework for a broader goal: Computationally combining the known metabolic network and the domain expert knowledge to automatically and accurately identify a small set of highly probable metabolic mechanisms that may have led to the given set of observed metabolite concentration changes. This paper describes the software architecture, data model, OMA tool capabilities and interfaces, and the performance evaluation of the OMA tool of PathCase^{MAW}. It does not present the key concepts and algorithms defining the context of our system; for such information, please see [8, 9].

This paper is organized as follows. Section 2 provides an overview of the system architecture. In Section 3, we present the data model. Section 4 describes the features of metabolomics analysis workbench. Section 5 presents running time performance study results. In Section 6, we discuss the related work, and Section 7 concludes.

II. PATHCASE^{MAW} ARCHITECTURE

In this section, we present a high level architecture of the PathCase^{MAW} system. Since the PathCase^{MAW} System is an extension of the existing PathCase [7] system, its architecture derives from that of the PathCase system. The subsections from Figure 1 that are shaded/colored are developed from the ground up and/or extended for this system, while the uncolored (white) subsections are unmodified from the PathCase system.

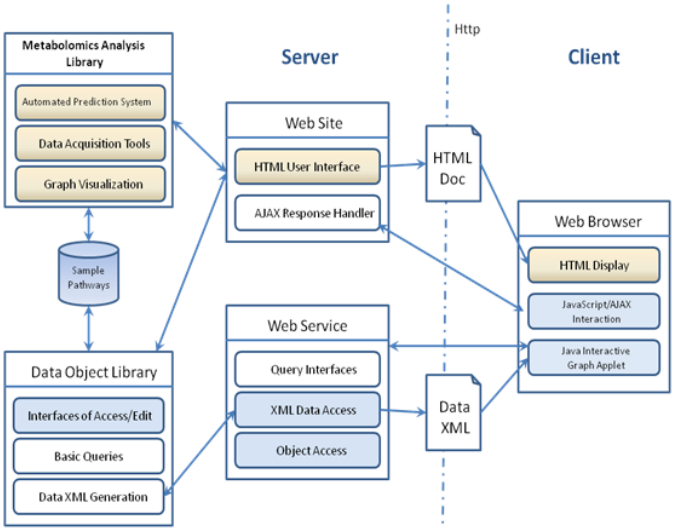

Figure 1. PathCase^{MAW} Architecture

The PathCase^{MAW} architecture has five distinct subsystems.

1. *Database*: This subsystem contains the actual pathways information that is collected from literature, and allows for efficient querying. The database adopts the relational model, and is hosted on a Microsoft SQLServer 2005 platform.
2. *The Data Object Library*: This subsystem provides a data access layer, which manages all data retrieval, creation, and update tasks. This library is developed in Microsoft C# 2005.
3. *Metabolomics Analysis Library*: This subsystem contains libraries to perform *Automated Prediction* and also includes tools for *Data Acquisition* and *Standalone graph visualization*. This Library is written in Microsoft C# 2005.
4. *The Web Server*: This subsystem includes the PathCase^{MAW} web site and the web services, both of which are written in C# and ASP.NET. It generates standard HTML pages and XML data. This allows the site to be accessed by users from a standard web browser on any operating system.
5. *Data Presentation:* The final subsystem contains the components of PathCase^{MAW} that run on the user's web browser. This includes the basic HTML that renders the main site interface to the user, the JavaScript with AJAX that makes the site highly responsive to the user, and, finally, the graph viewer Java applet used for interactive pathway graph, and hypothesis (*Closure Tree*) visualization. The graph viewer applet makes use of the web service subsystem in order to request additional data as needed, and to enhance the graph visualization without requiring the user to reload the web page. All graph manipulations such as zooming in and out, panning, and application of different layouts are carried out on the client side with no server side requests, which makes the system highly scalable.

III. DATA MODEL

A metabolic pathway is a series of metabolic reactions occurring within an organism. Each reaction in a pathway is a biochemical step from specific substrates (input molecules) to products (output molecules) that are chemically modified substrates. Each step may also use various combinations of molecules as cofactors, activators, inhibitors, and regulators, and usually involves at least one genetically unique gene product that catalyzes the reaction step.

The PathCase^{MAW} database is designed to provide a standard and consistent means of representing all the relevant information about metabolic pathways which then serves as a foundation for the tools created to work with this information. The database design is based on the PathCase database model[2], and extends the PathCase model to enable following new features

1. Capturing tissue location information,
2. Storing rules that capture the causal effects of dietary conditions or certain physiological processes on metabolite concentrations. The database is modeled around three basic entities: molecular entities, processes and pathways.
3. Representing a set of reactions as a single higher level reaction to provide simplified views of metabolism.

In this section, we describe representations of these entities and various other data items that complete the PathCase^{MAW} database. Some of the relations are the same as those in

[2] http://nashua.case.edu/PathwaysWeb/DataModel.aspx

PathCase, while some have been extended from relations with the same name in the PathCase system, and some relations are new and have been added specifically for this system. For brevity, in the context this paper, we only describe extensions that are unique to the PathCase^{MAW} database. The full database contains 48 relations. Complete data model discussion is available in [8].

Processes: A process is any interaction between molecular entities (also called as "reaction"). A process is almost entirely defined by its molecular entities and what role they play in that interaction. The molecular entities involved in a process and the role they play in the process are stored in separate relations. In order to allow for the representation of multiple reactions to be grouped into a single higher level reaction, a *parent_process* field has been included the PathCase^{MAW} database. This field holds the id of the higher level process (e.g., a pathway represented as a single process) to which the current process belongs. The *is_reaction* attribute stores a bit value that indicates if the current process is a single leaf level process or is a higher level process that abstractly represents multiple processes (1=leaf level process, 0=Higher level process). Moreover, it is necessary to distinguish a transport process from a regular process because special automated prediction and visualization rules apply to this class of processes. The *is_transport* attribute stores a bit value that indicates whether the current process is a transport process.

Process Entities: This relation is used to identify the molecular entities involved in a process. This relation is extended with the new *tissue_id* field that is used to specify the tissue in which the molecular entity must exist to take part in the process.

Tissues: This is a new relation created for this system, and stores the biological compartments at different levels of compartment hierarchy.

Pathway links: Pathway links relation stores the interconnection points (i.e., links) among pathways in the metabolic network. While links between pathways can be computationally identified (i.e., by looking for common molecular entity and tissue combinations amongst the processes in different pathways), the completely computational approach sometimes leads to false positives, as some pathways selectively work together with certain other pathways (e.g., *Beta-oxidation* and *Ketone Synthesis* pathways). Furthermore, such connections may differ from one tissue to another. Hence, linked pathways are explicitly and manually identified with connecting molecular entity and specific tissue information. Moreover, a *flux score* which indicates the likelihood (probability) of the link is stored in this table to allow for hypothesis ranking.

Pathway Co-Occurrence: When constructing the entire metabolic network it is not possible or unlikely that the reactions of a particular pathway *co-occur* together with the reactions of some other pathway as part of the same valid metabolic path in a particular tissue. The *Pathway Co-occurrence* relation is newly added, and represents such kind of relationships between pathways. This information is mainly used in Automated Prediction operations. Each row in the table represents a co-occurrence rule. The Id field is an auto generated unique identifier for this table. *pathway_id_1* and *pathway_id_2* stores the ids of the pathways and *tissue_id* attribute is used to store the tissue Id over which this rule exists. The *cooccurrenceLikelihood* attribute is used to store the numerical value of the likelihood that the reactions of the two pathways can occur together in a particular tissue.

IV. METABOLOMICS ANALYSIS WORKBENCH

The web interfaces for browsing pathways, processes, molecular entities, organisms, and the Built-in queries remain the same as in the PathCase System [7], and, hence, are not discussed here.

A. *Consequence Prediction Web Interface*

In this section, we introduce the web front-end for the OMA tool, and walk through a sample run of "hypotheses" generation, where hypotheses are those metabolic pathway fragments that are "activated" [8, 9]. Figure 2 shows the main screen of the Automated Prediction web front-end. Next, we describe specific parts of this interface.

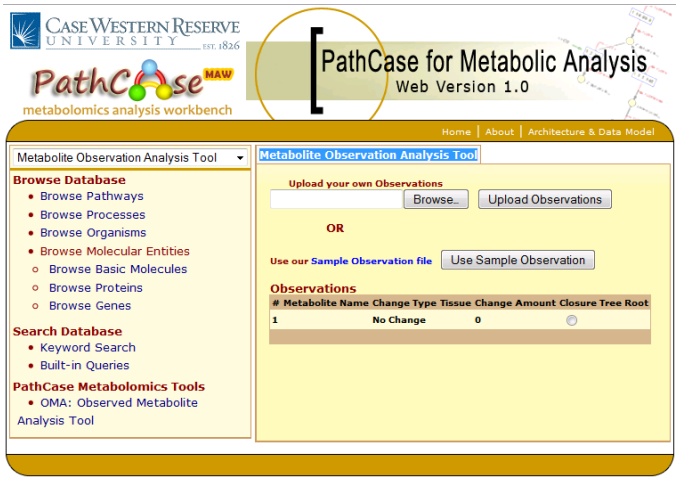

Figure 2. Automated Prediction web front-end

1) *Uploading Observations:*

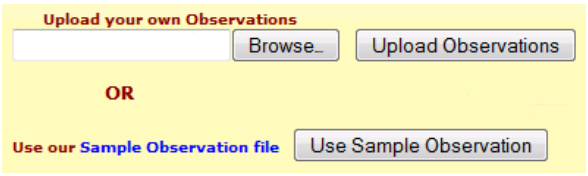

Figure 3. Options for uploading observations

The user initiates the entire automated prediction process by uploading the "fold" changes of metabolites, as observed from the bio-fluid samples. We currently allow three ways in which observations can be uploaded.

a. *XML File Upload*: Users can upload their observations in the form of an XML formatted file. Figure 4 shows an example of an XML file which contains 2 observations. To upload observations, the user first clicks the Browse button, and navigates to the location where the observations XML file is saved on the user's computer. The user then clicks the "Upload Observations" button to upload the observations from the selected XML file.

```
<Events>
    <Event entityName="Glycocholate" tissue="liver"
           changeType="increase" changeQuantity="3"/>
    <Event entityName="Glutamate" tissue="liver"
           changeType="increase" changeQuantity="2"/>
</Events>
```

Figure 4. XML file of sample user observations

b. *Using sample observation set*: Users have an option of using our sample observations file instead of uploading

their own observations. This can be done by clicking the "Use Sample Observation" button (Figure 3).

c. *Manually Entering Observations on the Web Form*: Alternatively, users can manually insert their own observations. Figure 5 shows a screenshot illustrating how the user manually adds observations. To assist users, when the user attempts to enter a metabolite name, an *AJAX Autocomplete* list is generated with valid metabolite names from the database. The user selects the change type and the tissue from which the observation was taken using dropdown lists. The user then enters the observed fold change, and clicks the "Add Observation" button.

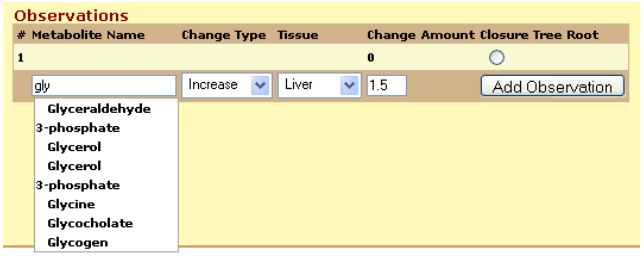

Figure 5. Manually entering observations

When users enter (or upload) observations, they are presented with a screen shown in Figure 6, which displays several different options that can be used while performing Automated Hypotheses generation. Each presented option has a default value; however, the user can change these default values to customize the automated prediction process.

Figure 6. Consequence prediction screen of the users uploaded observations

(1) *Model to be used*: Users have the option of performing automated prediction using either level 1 or level 2 models [8], which can be selected using a radio button list.

(2) *Levels of Expansion*: Users can control the depth of the hypotheses (closure) tree [9] to be generated for the purposes of automated prediction. The depth of the closure tree used in a particular automated prediction run determines the running time, and also the number of hypotheses that are generated.

(3) *Pathways*: Users have a choice of generating hypotheses based on the entire metabolic network or on a subset of the pathways that make up the metabolic network. The "Pathways" list box is populated with the names of all the pathways currently stored in the database, and users can select pathways from this list to be used for automated prediction. By default, the entire metabolic network is used for hypotheses generation.

(4) *Tissues*: Users can "restrict" the metabolic network to be used to those reactions in a particular tissue or a subset of tissues of interest. The "Tissues" list box is populated with the names of all the tissues for which reactions exist in our database. The user can use this list box to select one or more tissues. By default, all tissues are selected.

Once a user sets up the automated prediction options according to her satisfaction, she clicks the "Generate Observation Supported Hypotheses button" to initiate the Automated Prediction Process. Then, the user is presented with two additional collapsible panels which contain the generated hypotheses (in tabular form), as well as the hypothesis tree visualization.

2) The Hypotheses Display Panel:

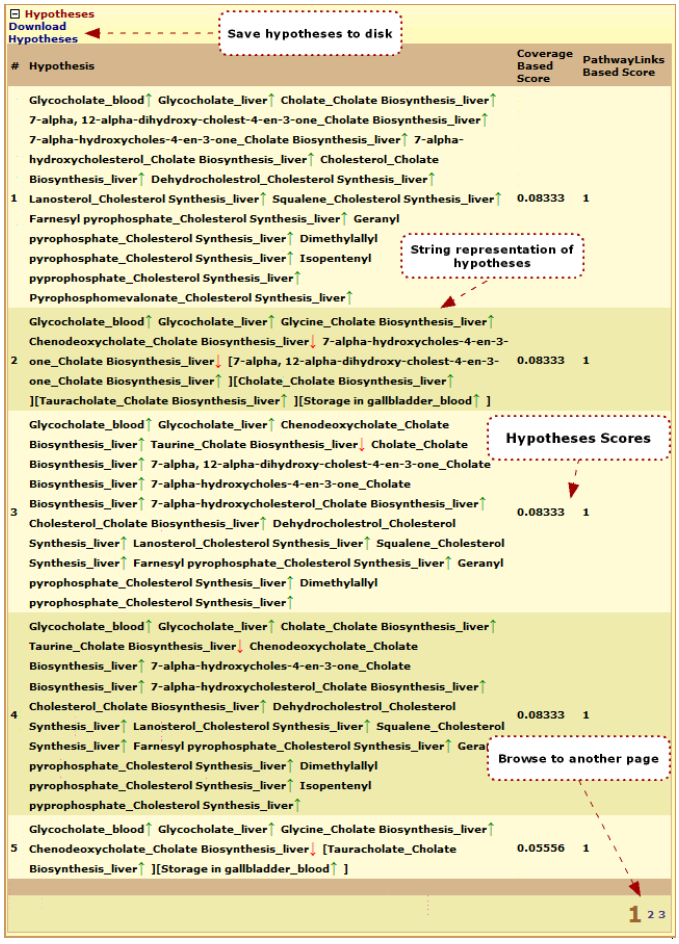

Figure 7. The Hypotheses Display Panel

The Hypotheses Display panel shown in Figure 7 lists all the M(aybe)-Valid Hypotheses that were generated by the

automated prediction process. Each hypothesis in the list has the following format: E_1 E_2 E_3 E_4……E_n where E_i represents an individual event [9] in the hypothesis. Each event has the following format:

{Metabolite/Process Name]_[Tissue Name]_[Change symbol]

The Change symbol can be one of the following:

↑: increase in concentration (for metabolites); increase in the activity level (for process).

↓: decrease in concentration (for metabolites); decrease in the activity level (for process).

+: the corresponding process is activated.

-: the corresponding process is inhibited.

In addition, each hypothesis is also associated with a Coverage, Implication and Pathway Links Score [8, 9]. Moreover, users are also allowed to save the generated hypotheses onto their own system by clicking on the "Download Hypotheses" link. Hypotheses are displayed in a "paged" manner, and users can navigate through the pages selecting the desired page on the bottom left of this panel.

3) The Hypotheses (Closure) Tree Display Panel

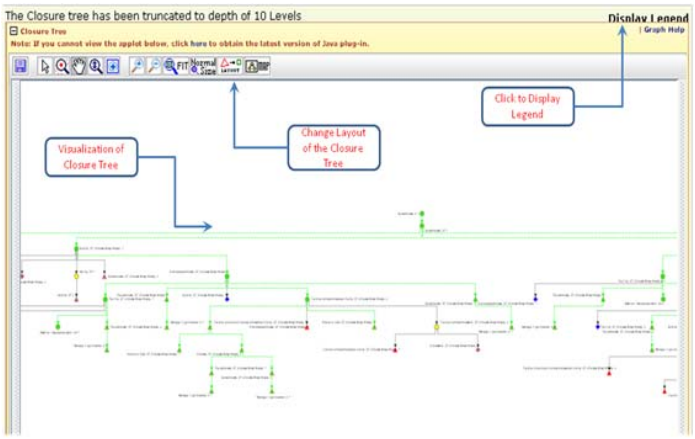

Figure 8. Hypotheses Tree display panel

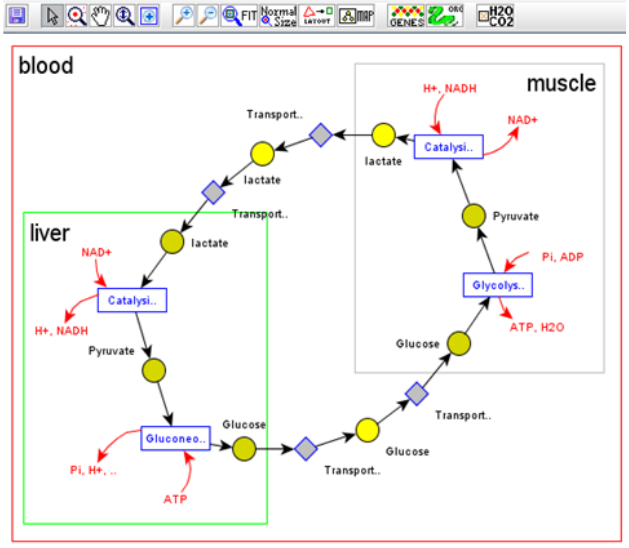

Figure 9. Tissue-aware visualization

The panel shown in Figure 8 displays the generated Hypotheses (Closure) Tree [9]. The Closure tree can become very large if the user selects a large depth. And, because the Java Applet renders the tree on the client's machine, displaying the complete closure tree can be extremely demanding on the Client machine. To avoid this, the depth of the closure tree is truncated to 10 levels. The closure tree depth restriction is placed only for visualization of the closure tree and not for the purpose of automated hypotheses generation. Users can change the layout of the closure tree visualization by using the layout button on the applet. A legend is displayed if the user clicks the "Display Legend" link.

4) Tissue-Aware Visualization

As described briefly in the above sections, PathCase^{MAW} provides a tissue-aware visualization. As an example, Figure 9 illustrates Cori Cycle pathway, which spans over two tissues namely, liver and muscle.

V. Running Time Performance Study

In this section, we evaluate the running time behavior of our metabolomics analysis framework. In particular, we study how the length of the hypotheses affects the running time of the algorithm, and also the effect of several enhancements that we have developed during our implementations. These sets of experiments were run on the entire network with glutamate as the root of the closure tree with a sample metabolomics data set having 34 observations [8].

The database is populated by manually entering major *pathways of aminoacid, lipid,* and *carbohydrate metabolisms* from the literature (mostly, from a biochemistry textbook [11] and an atlas of human metabolism [12]. Currently, our database contains a total of 50 pathways in 9 tissues, 241 reactions, and 205 metabolites. Please see Table 1 for statistics about the current database content.

Table 1. Database Content

	Amino Acid Metabolism	Carbohydrate Metabolism	Lipid Metabolism	Whole Database
Num. of pathways	28	11	11	50
Num. of processes	118	68	55	241
Num. of metabolites	145	52	70	205
Num. of tissues	5	9	5	9
Num. of graph nodes	476	426	219	980
Num. of pathway links	42	31	5	123

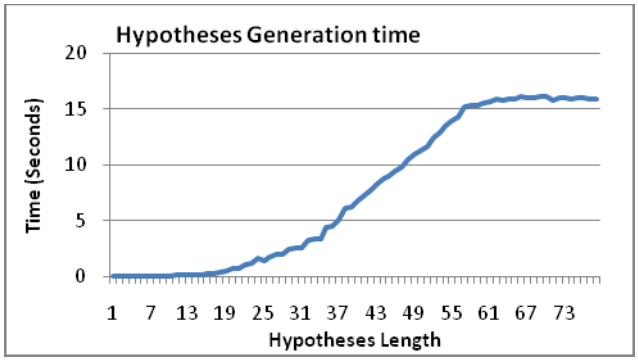

Figure 10. Hypotheses length vs. runtime

Observation: Running time of the algorithm progressively increases with an increase in the hypotheses length (Figure 10). However, after a particular length (72 in our case), it stabilizes.

The runtime of the algorithm for a particular closure tree depth depends on the size of the closure tree that is generated,

and as a consequence on the number of hypotheses ("M-valid" and "invalid" [9]) generated. The plateau seen in the runtime corresponds to that seen in the number of generated hypotheses.

In order to improve the efficiency of our approach, we have developed a simple early termination strategy for closure tree generation. Our hypotheses generation approach involves two main stages: (a) *the hypotheses (closure) tree generation*, and (b) *the identification of observation-supported hypotheses* [9]. We had initially implemented these two stages independently, and run them one after another as separate procedures. We call this implementation *baseline* approach. In order to improve the performance, we later modified the baseline approach and implemented the "Closure tree generation" step and the "Identification of observation supported hypotheses" in an interleaved manner. More specifically, closure tree generation at each step performs a check against the observed events. During the generation of the closure tree, when a new event is generated and added to the tree, we check if it conflicts with any of the observed events in addition to checking for duplicates and conflicts on its own path to the root of the closure tree. And, if it conflicts with any other observed events, the event is marked as "conflicting with observed event", and the expansion along that path of the tree is terminated. The motivation here is to avoid the generation of hypotheses that will later be pruned in the "Identification of observation supported hypotheses" step. We refer to this improved version as *early termination* approach. Figure 11 presents the running time for the baseline and early termination approaches.

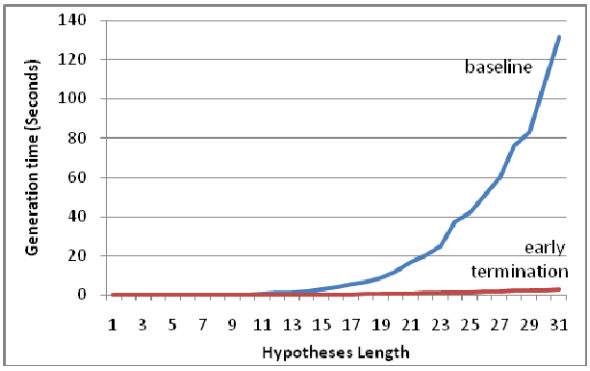

Figure 11. Baseline vs. Early Termination Approach

Observation: Early termination approach significantly improves the running time.

The above observation is mainly because of the fact that the size of the in-memory tree generated by the early termination approach will be considerably smaller than the baseline approach.

VI. Related Work

Currently, there are many web-based metabolic network data sources, e.g., KEGG [4], Reactome [5], MetaCyc [6], PATIKA [10], PathCase [7], etc. that contain metabolic network information for humans as well as other model organisms. PathCase, which is in a way our parent project, does not aim to compete with, but to complete, these systems by providing an alternative user interface with additional capabilities. However, none of the above data sources (with perhaps some exceptions for Reactome) contain tissue information for pathways, or model transport processes. To the best of our knowledge, there is no other web-based tool that provides metabolic assessment capabilities.

VII. Conclusion

Metabolomics measurements provide opportunities for non-invasive detection of physiological conditions. However, manual analysis of the measurements is time-consuming and requires expert knowledge. In this paper, we have described a metabolomics analysis workbench called PathCase^{MAW} which provides a web accessible metabolic pathways database with location information for each pathway, as well as transport processes. Moreover, we have also developed a tissue-aware metabolic pathway visualization framework that incorporates our stored tissue information for the visualization of pathways, processes, and groups of pathways, and thus creates a more accurate view of the relevant biological mechanisms.

Finally we have presented a web-based front-end to our metabolic analysis framework, which clinical users can use to upload metabolite level changes, as observed from biofluid measurements, and use our framework to computationally identify a list of mechanisms that produce the observed/ measured metabolite changes.

ACKNOWLEDGEMENTS

This research is supported in part by the NSF award DBI-0218061, a grant from the Charles B. Wang Foundation, and equipment grants from NSF and Microsoft.

References

[1] Harrigan, G.G., Goodacre, R. (Eds), *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Plubishers, 2003.

[2] Kell, DB, and Westerhoff, HV. 1986. Metabolic Control Theory—its Role in Microbiology and Biotechnology. *FEMS Microbiol. Rev.*, Vol. 39, pp. 305-320.

[3] Milburn, Michael. Metabolomics for Biomarker Discovery:, IPTOnline Journal, 12-2006 , page 40.

[4] Kanehisa M et al. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34 (Database issue):D354-7.

[5] Joshi-Tope G et al. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33 (Database issue):D428–32.

[6] Caspi, R et al. 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 34 (Database issue):D511–16.

[7] Elliott, B., Kirac, M., Cakmak, A. et al. 2008. PathCase Pathways Database System. Bioinformatics 24(21): 2526-2533, November 2008.

[8] D'Souza, Arun. PathCase^{MAW}: A Workbench for Metabolomic Analysis. MS Thesis, Case Western Reserve University, EECS - Computer and Information Sciences, 2009. Available online at http://cakmak.case.edu/TechReports/PathCaseMAW.pdf

[9] Cakmak, A., Dsouza, A., Hanson, R., Ozsoyoglu, ZM, Ozsoyoglu, G. Analyzing Metabolomics Data for Automated Prediction of Underlying Biological Mechanisms, 2009 (*Technical Report).* Available online at http://cakmak.case.edu/TechReports/Metabolomics.pdf

[10] Demir, E., Babur, O., Dogrusöz, U. et al. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics 18(7): 996-1003 (2002).

[11] Devlin, TM. 2006. Textbook of Biochemistry with Clinical Correlations, Sixth Edition. Hoboken, NJ, *John Wiley & Sons*.

[12] Salway, JG. 1999. Metabolism at a Glance, 2nd Edition. *Blackwell Science.*