Analyzing Metabolite Measurements for Automated Prediction of Underlying Biological Mechanisms

Ali Cakmak[1]
cakmak@case.edu

Arun Dsouza[1]
dsouza@case.edu

Richard W. Hanson[2]
rwh@case.edu

Meral Ozsoyoglu[1]
meral@case.edu

[1]Department of Electrical Eng. and Computer Science
[2]Department of Biochemistry
Case Western Reserve University

ABSTRACT

The emerging field of metabolomics enables researchers to measure concentrations of large numbers of metabolites in biofluids, and to interpret them in connection with the underlying metabolic network, which poses a significant challenge for manual analysis. Given a set of observations on metabolite concentration changes, our goal in this study is to employ automated reasoning, and provide researchers with possible metabolic action scenarios that may have occurred in the body to produce the observed metabolite changes. Our proposed methodology, called the Observed Metabolite Analysis, is to (1) computationally chase the implications of a given a set of metabolite concentration change observations in body fluids, relative to a control subject, (2) eliminate metabolic action scenarios, called hypothesis, that are invalid (i.e., those scenarios that could not have happened) (e.g., increased protein turnover), and (3) rank possibly valid metabolic action scenarios on the basis of pre-defined flux ratio information. We computationally evaluate the proposed methodology with typical metabolomics data, and demonstrate that (a) through consistency analysis against a small number of measured metabolite concentration changes, over 90% of the automatically generated hypotheses are invalidated with no manual analysis, (b) using summarization techniques, the entire hypothesis set is represented by a much smaller (2% of the original) hypothesis set, and (c) performing hypothesis generation and consistency checking in an interleaved manner leads to over 95% improvement in running time.

1. INTRODUCTION

Metabolites are intermediates and products of metabolism; the metabolome refers to the complete set of metabolites in a cell, a tissue, or an organism; and, metabolomics is the study of the distributions (profiles, concentrations) of the metabolites in the metabolome [1, 2]. Metabolites in the metabolome have low-molecular weight [3], and are ideal for sensitively analyzing changes in a biological system [4].

With the recent advances in experimental technologies, such as gas chromatography and mass spectrometry, the number of metabolites that can be measured in biofluids has rapidly increased. In order to identify the metabolic mechanisms that lead to changes in the concentrations of given metabolites, one needs to interpret the metabolic significance of the observed changes. This necessitates a time consuming, extensive and manual cross-referencing of metabolic pathways, in order to critically evaluate the data. The large number and breath of the metabolites represents a challenge to an informed interpretation of the results, when the goal is to determine the biochemical mechanisms that are responsible for the observed changes. There is thus a need for computational tools to help biologists and clinical researchers to derive meaningful interpretations of metabolomics data. This paper proposes and evaluates techniques for automated interpretation and analysis of metabolomics data via existing metabolic networks.

We make the following assumptions about our model.

(a) A complete metabolic network is pre-captured and available in a metabolic network database.

(b) The metabolic network database captures tissue-level compartmentalization.

(c) The only input that the user provides to the system is a set of bio-fluid metabolite level observations (referred to as “observed metabolite changes”) in the form of “increase”, “decrease”, or “no change”, with respect to a control subject (i.e., metabolomics data); and

(d) The system is studied in a stable steady state; that is, the rate of formation of every metabolite is equal to its rate of degradation.

This research has two goals.

1. *Goal 1:* Specify those metabolic action scenarios that are invalid (i.e., could not have happened--given the set of observed metabolite concentrations), which results in a set of “maybe-valid” metabolic action scenarios (that are then to be manually evaluated by biologists). In the rest of the paper, we refer to such metabolic action scenarios as *(low-level) hypothesis.*
2. *Goal 2:* Map, as much as possible, the maybe-valid metabolic action scenarios to disease and/or physiological conditions, giving users possible clues on the implications of the set of observed metabolites. We also refer to such mappings as *(high-level) hypothesis*.

ACM-BCB 2010, Niagara Falls, NY, USA
Copyright © 2010 ACM ISBN 978-1-4503-0192-3... $10.00.

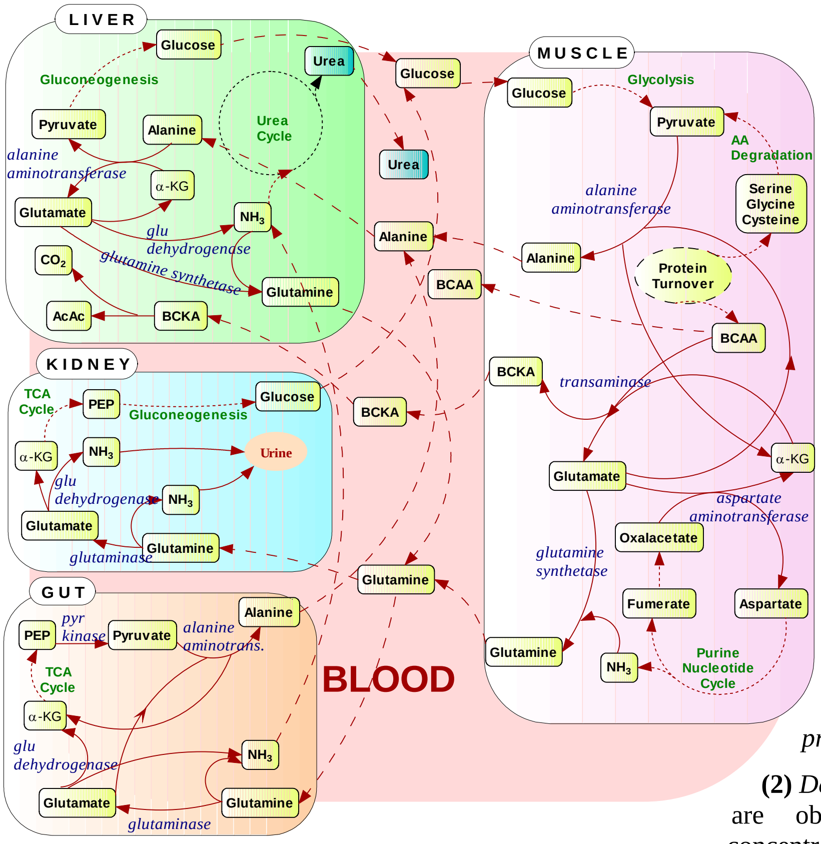


Figure 1. Metabolism of BCAAs

In this paper, we concentrate only on goal 1, though we briefly describe how to achieve goal 2. We start with a simple example (that illustrates a case involving goal 2).

Example 1.1. Consider five changes in the concentration of metabolites that were observed in the blood of patients: (Glutamine, "Increase by 4-fold"), (Alanine, "Increase by 2-fold"), (Urea, "Decrease by 0.5-fold"), (BCAA, "no change"), (Glucose, "Increase by 1.3-fold"), where BCAA refers to branched-chain aminoacids (i.e., valine, leucine, isoleucine). The metabolic fate of glutamine is as follows:

Observation 1: *The relative concentration of glutamine may increase due to (i) an increase in its production by muscle as a result of increased protein turnover, and/or (ii) and increased production by liver, where its synthesis serves as a sink for ammonia, due to a dysfunctional urea cycle, and/or (iii) decreased uptake by kidney due to the decreased level of activity of glutaminase, and/or (iv) the decreased rate of glutamine uptake by the gut due to a lowered activity level of glutaminase.*

We turn the above observations into four separate (high-level) hypotheses (which are maintained in the metabolic network database): The cause of increase in the concentration of glutamine is due to: **H$_1$:** an increase in its rate of production by *muscle*, as a result of possible increase in protein turnover, **H$_2$:** increased hepatic production, as a result of dysfunction of a key enzyme in the urea cycle. **H$_3$:** decreased glutaminase activity in kidney. **H$_4$:** decreased glutaminase activity in gut.

With respect to these four hypotheses and the five metabolite concentration change observations, our goal is to employ organ-/tissue-based metabolic pathway knowledge to choose between the four hypotheses listed above. We start by redefining the notion of *hypothesis* in our context, as a set of statements on the concentration changes of metabolites, created to explain a possible biological mechanism leading to the observed metabolite concentration changes. Hypothesis generation has four steps, with the first two being:

(1) *Chase Process* generates hypotheses by chasing the observed metabolite concentration changes within the human metabolic network, and employs (multiple variations of) the metabolic reasoning: if the concentration of a metabolite m is observed to *decrease*, then either it is *consumed more* and/or *produced less* than it was before a perturbation. And, if the relative concentration of a metabolite m is observed to *increase*, then either it is *consumed less* and/or *produced more* than it was before a perturbation.

(2) *Derived changes in the concentration of a metabolite* are obtained from (a) the observed metabolite concentration changes and (b) the whole metabolic network. We give an example for steps 1 and 2.

Example 1.2. Consider the change in the concentration of 5 metabolites in Example 1.1.

Goal. In the light of the observed events (i.e., blood test results), by employing computational techniques, interpret the increase in the concentration of glutamine in blood, and determine which of the four possible alternative mechanisms (hypotheses) may have led to this increase.

Figure 1 shows (part of) a human metabolic network related to the metabolism of branched chain amino acids.

Verifying that hypothesis H$_1$ is invalid: Hypothesis H$_1$ implies an increased concentration of BCAAs due to elevated protein turnover; but the measured concentration for BCAAs indicates no such major changes. Hence, based on the metabolic network of Figure 1, we conclude that *H$_1$ is not valid*, i.e., an increase in glutamine concentration is not due to increased muscle protein turnover.

Verifying that hypothesis H$_2$ may be valid (M-Valid): The cause of increase in glutamine may be due to an increased production by the liver. Next, on the basis of the five observed metabolite changes, we show that hypothesis H$_2$ is indeed M-Valid. The path representing hypothesis H$_2$ is marked with label "H2" in Figure 2. This scenario, from the root to the leaf node, basically states that: (a) *Gln_Blood*↑: Glutamine increases, because (b) *Gln_Lvr*↑: it is produced more by the liver, which is due to (c) *NH$_3$-Lvr*↑: accumulation of ammonia (NH$_3$) in liver, due to a decreased synthesis of urea possibly caused by dysfunction

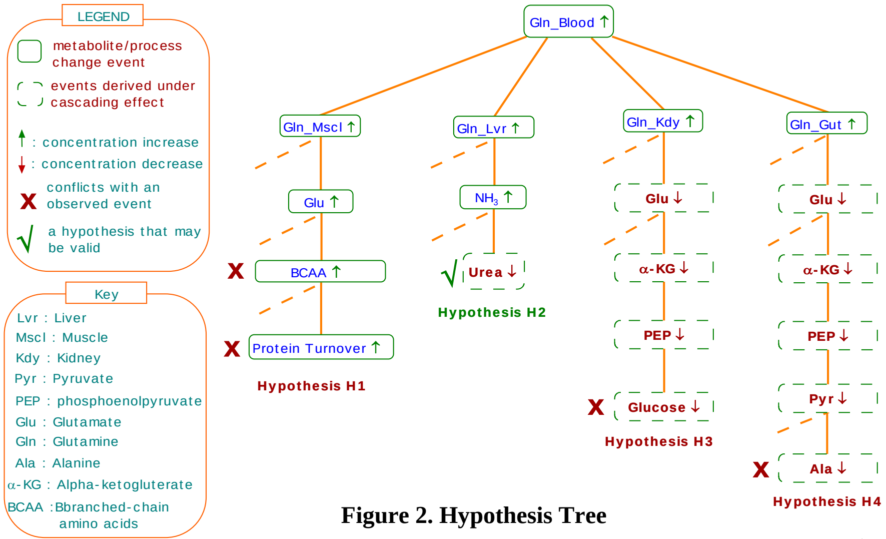


Figure 2. Hypothesis Tree

of urea cycle leading to *Urea-Lvr*↓, and to (d) *Urea*↓: a decreased urea in the blood. Thus, we conclude that hypothesis H_2 is validated and may have occurred, i.e., "H_2 is M-Valid".

Similar to H_1, *H_3 is not valid* due to the conflict with the observed concentration change of glucose, and *H_4 is not valid* due to the conflict with the observed concentrate change of Alanine. ■ **End of Example 1.2.**

Examples 1.1 and 1.2 illustrate the use of automated ways of eliminating hypotheses from among a list of likely alternatives known *a priori* as possible. A broader way of using our approach is to evaluate *all* possible hypotheses (either in the whole metabolic network or its sub-network), eliminate the ones that are invalid, and rank and list the ones that are M(aybe)-Valid. In the rest of this paper, we formalize and present our within this latter alternative.

In addition steps 1 and 2 above, we use two more steps.

(3) *Hypothesis likelihood* is evaluated based on "expected flux ratios" of metabolites in reactions/pathways.

(4) *Physiological condition mapping* is done based on the overlap between known biomarkers for physiological conditions and the set of metabolites in a hypothesis.

Section 3 presents an evaluation of our approach using a typical metabolomics data set. First, we empirically show that the majority (over 90%) of possible hypotheses can be eliminated via employing a reasonably small number of observations. Second, our proposed hypothesis summarization methods allow for the representation of the whole hypothesis set with size as small as 2% of the original hypothesis set. Third, employing an early termination strategy during the hypothesis generation improves running time up to 97%. Thus, our automated interpretation approach is effective and useful.

This paper is organized as follows. In section 2, we discuss related work. Section 3 formalizes our model (referred as OMA) for metabolomics analysis. In section 4, we experimentally evaluate OMA, and section 5 concludes.

2. RELATED WORK

For related work to goal 1 (eliminating invalid metabolic scenarios, and locating possible (M-valid) ones), our proposed OMA technique can be considered in the general category of metabolic analysis techniques which include metabolic control analysis (MCA) [5], flux balance analysis (FBA) [6], metabolic flux analysis [7], and metabolic pathway analysis (i.e., elementary flux modes and extreme pathways) [8]. For related work to goal 2, we are not aware of any computational or algorithmic techniques. Next, we briefly summarize related studies and compare to our proposed framework.

Metabolic control analysis (MCA) aims to characterize the sensitivity of metabolic responses against changes in enzyme activities or parameters [5]. MCA, through the summation and the connectivity theorems, allows one to relate the global (systemic) properties of the pathway to the (local) properties of individual enzymes.

Flux balance analysis (FBA) computes stationary fluxes in metabolic networks. It is based on convex analysis imposing an objective function subject to several constraints, to determine the metabolic flux vector. Usually, the fluxes are determined to maximize a specific network output, e.g., the biomass which is a reasonable objective for primitive cells such as bacteria, but not necessarily for complex eukaryotic cells. Further critiques of FBA include: (a) it identifies only one optimal solution (while there may be other optimal/suboptimal solutions), (b) flux distributions predicted by FBA are hypothetical (as they depend on the choice of the flux criteria) [9].

Metabolic pathway analysis (MPA) identifies the topology of cellular mechanism based on the stoichiometry and thermodynamic constraints of reactions. Two main techniques in MPA are elementary flux mode analysis (EMA) [10], and extreme pathways analysis (EPA) [8]. In comparison with FBA, MPA can identify all metabolic flux vectors; but it also has high computational complexity. (See [8] for an excellent review of EMA.) Free applications that compute elementary flux modes include COPASI [11], Metatool [12], SNA [13], FluxAnalyzer [14], YANA [15].

Comparison. Next we briefly list the differences between the MCA, FBA, EMA, and OMA approaches:

►*Different goals.* The four approaches are useful in different contexts and have different goals. (a) MCA focuses on "control as a property of the whole system". One can measure the effect of single enzyme perturbations on the system. (b) EMA can be used for tasks like the recognition of operational elementary modes, finding all optimal paths, analysis of network flexibility [14]. However, identifying the weighting factors to determine the contributions of each elementary mode is difficult, if not impossible [16]. (c) OMA, working with the whole (and possibly large) metabolic network within a multi-tissue environment (i.e., not within a cell), returns to users a list of *possible* metabolic action scenarios (i.e., M-valid paths) as well as their visualizations, allowing users to quickly concentrate on locating possibly activated paths for a given set of observed metabolite concentration changes. ►*Different underlying fundamentals.* OMA is rule-based, and employs graph search algorithms across the whole metabolic network. In comparison, MCA and FBA involve solving a set of underconstrained differential equations corresponding to a possibly smaller metabolic network. EMA determines elementary fluxes via a linear combination of "null space basis vectors" of the stoichiometry matrix [17]. ►*Ease of use.* MCA (or FBA), even with the easiest-to-use software tools (such as COPASI), requires setup and usage expertise, for biologists to use them. The EMA tools and YANA do provide user-friendly elementary flux derivations and their visualizations. In comparison, OMA uses a metabolic pathways database, which already contains the metabolic network so that all that a user is expected to provide is a set of observed metabolite changes. ►*Modeling-related restrictions/assumptions.* MCA and EMA have a number of assumptions (e.g., connected pathway network) [20] which are not needed for OMA. ►*Computational Complexity.* Computational complexity of MCA is exponential in the number of reactions involved, forcing users to use various compaction, aggregation, and clustering techniques. Computational complexity of EMA is also exponential [19], and various approaches to tackle the high complexity are proposed such as parallel computing [20]. In its worst case, OMA is also exponential in the number of paths between the root node of the closure tree and other nodes. However, metabolic networks form sparse graphs, and, for the prototype metabolic network used in Section 4, the worst-case complexity has not been a limiting factor.

Relationship to Artificial Intelligence studies. In general, our computational modeling and analysis technique can also be viewed in the class of qualitative reasoning [21], qualitative simulation [22], and qualitative process theory [23] in artificial intelligence. In particular, our hypothesis formation framework may be viewed as a specialized case of hypothesis formation design methods, as developed by Karp [24]. There are also some other metabolomics studies, which we do not discuss here due to space limitations. Please see Wishart [25] for an excellent survey of existing computational approaches in metabolomics data generation.

3. MODEL FOR METABOLIC NETWORK-BASED OBSERVATION ANALYSIS

This section presents our model for metabolic network-based observation analysis, and defines chase process rules.

Def'n *(Reaction)*: A reaction RN(E, S, P) consists of a set E of enzymes which collectively consume the metabolite set S (i.e., substrates), and produces the metabolite set P (i.e., products). The metabolites in S are called *co-substrates* of each other.

Def'n *(Metabolic Network)*: A metabolic network is a graph G(V, E) of a vertex set V of reactions and metabolites, and a directed edge set E such that there is an edge from node u to node v if (i) v is a reaction, and u is a substrate of v, or (ii) u is a reaction, and v is a product of u.

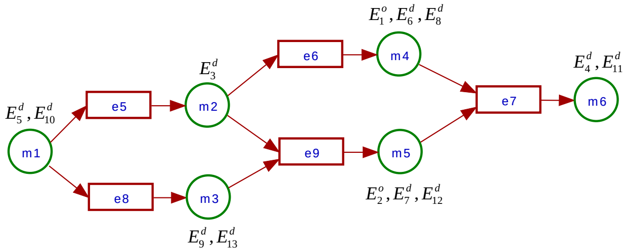


Figure 3. A hypothetical metabolic network

Def'n *(Upstream/Downstream)*: Given a metabolic network M and two metabolites m_i and m_j where m_i, $m_j \in V_M$, we say that *m_i is located downstream of m_j* or *m_j is located upstream of m_i*, if there is a path from m_j to m_i. If m_i and m_j belong to the same reaction as substrates and products, respectively, then, we say that *m_j is located immediately upstream of m_i* or *m_i is located immediately downstream of m_j*.

Example 3.1. Figure 3 illustrates a hypothetical metabolic network. In the reaction with enzyme e_9, the metabolites m_2 and m_3 are co-substrates. m_2 is located upstream of m_6, while m_4 is located immediately upstream of m_6.

Def'n *(Concentration C of a Metabolite):* A metabolite m has the *concentration C_m* before a perturbation and the *concentration $\hat{C}_m$* after the perturbation.

Def'n *(Observed Event):* An observed event E^o (m, c) is a pair of a metabolite m and an observation c on the concentration of m where c represents a concentration change from C_m to $\hat{C}_m$, and is one of "increase [by X fold]", (ii) "decrease [by X fold]", or (iii) "no change".

Please see example 1.1 for sample observed events.

Two types of events are possible, namely, *observed events* E^o and *derived events* E^d. All events that are directly stated in a biofluid test result are called observed events, while those that are derived based on the observed events and the structure of a metabolic network are called derived events. When there is no need to distinguish between observed and derived events, we drop the superscripts from E^o and E^d.

3.1 Derived Event Characterization

Derived events are defined using the following metabolic biochemistry reasoning.

Remark 3.1. If the concentration of a metabolite m is observed to *decrease* after a perturbation, then either it is *consumed more* and/or *produced less* than before the perturbation. Likewise, if the concentration of a metabolite m is observed to *increase* after a perturbation, then either it is *consumed less* and/or *produced more* than before the perturbation.

Also, reaction rates are controlled by many factors, referred to here as *reaction-rate-control (RRC) events*, involving (allosteric or competitive) inhibitors/activators, enzyme (or gene) expression rate changes [26]. While modeling each of these factors separately is a more precise approach, in this paper, as a first step and for simplicity in rules and algorithms, we model them all as having only one type, i.e., the RRC event. As an example, in hypothesis H_2 of Example 1.1., *Urea*↓ is an an RRC-induced event.

Assume metabolite m is a product of reaction R_p, and a substrate of reaction R_s. Then,

- Increase in concentration of m is caused by:
  - More production due to increase in concentrations of substrates in R_p (*substrate-induced causality*), or
  - Less consumption due to decreased concentrations of a co-substrate of m in R_s (*co-substrate-induced causality*), or
  - Less consumption due to an RRC event such as increase in the concentration of an inhibitor for R_s, or gene/enzyme expression changes for R_s [26]. The RRC event also causes “decreased production of product(s) of R_s” (*RRC-induced causality*)
- Decrease in concentration of m is caused by:
  - Less production due to decreased concentrations of substrates in R_p (*substrate-induced)*, and/or
  - More consumption due to increased concentration of a co-substrate of m in R_s (*co-substrate-induced causality)*, and/or
  - More consumption due to an RRC event, which causes an increase in products of R_s (*RRC-induced causality).*

Formalizing the above remark and the RRC event, next we present the derived event notion.

Def’n *(Negation of a concentration change):* Given a concentration change c_i in an event $E_i(m_i, c_i)$, the negation of c_i, denoted as $\neg$ c_i, represents the concentration change in the opposite direction of c_i, e.g., if c_i involves an “increase”, then $\neg$ c_i = “decrease”.

Def’n *(Derived Event via Remark 3.1)*: Given a metabolic network M, two metabolites m_i and m_j in V(M), and a (derived or observed) event $E_i(m_i, c_i)$, $E_j^d(m_j, c_j)$ is a derived event induced by $E_i(m_i, c_i)$ where c_j is determined based on c_i using one of the three causality rules:

$$c_j = \begin{cases} c_i & \text{if } m_j \text{ located immediately } \textit{upstream} \text{ of } m_i \text{ (Rule1)} \\ \neg\, c_i & \text{if } m_j \text{ is a } \textit{co-substrate} \text{ of } m_i \text{ (Rule2)} \\ \neg\, c_i & \text{if } m_j \text{ located immediately } \textit{downstream} \text{ of } m_i \text{ (Rule3)} \end{cases}$$

Example 3.2. Consider Figure 3 and the event $E_1^o(m_4$, “increase by 2 fold”), which may be either due to an increase in m_2, decrease in m_5, or decrease in m_6. Hence, one can create three derived events which are induced by E_1: $E_3^d(m_2$, “increase”), $E_4^d(m_5$, “decrease”), as well as $E_5^d(m_6$, “decrease”) due to RRCeffect.

So far, we have modeled “caused-by” relationships on metabolite concentration changes. Next, we model the “causes” relationship as creating a (forward) “cascading effect”.

Remark 3.2. *(Forward Cascading Effect):* If a metabolite m is produced less or produced more during a perturbation, in the absence of downstream RRC events, it triggers the same effect (i.e., increase/ decrease) on the concentrations of metabolites that follow m within the metabolic network.

Finally, increases/decreases in the concentration of some metabolites may be caused by two additional factors: (i) dietary intake, and (ii) certain physiological processes (e.g., muscle breakdown). In order to model such factors, we employ the notion of *external process*, and register them as producer/consumers of metabolites. Then, such external producers/consumers are treated as regular reactions to derive new events. Please see [29] for more details.

3.2 Hypothesis/Closure Tree

Using the above event derivation models, an event E_j indirectly induces a larger set $S^{i(ndirect)}$ of derived events than the set $S^{d(irect)}$ of derived events that are directly induced by E_j. We give an example.

Example 3.3. Consider Figure 3 and the event $E_1(m_4$, “increase by 2 folds”). The set S^d of derived events that are directly induced by E_1 are $S^d(E_1) = \{E_3(m_2$, “increase”), $E_4(m_5$, “decrease”), $E_5(m_6$, “decrease”)}. Then, E_3 in turn induces $S^d(E_3) = \{E_6(m_1$, “increase”), $E_7(m_3$, “decrease”), $E_8(m_4$, “decrease”), $E_9(m_5$, “decrease”)}. Similarly, E_4 induces $S^d(E_4) = \{E_{10}(m_2$, “decrease”), $E_{11}(m_3$, “decrease”), $E_{12}(m_4$, “increase”), $E_{13}(m_6$, “increase”)}.

Note that some newly derived events may conflict with other derived or observed events, and an accurate analysis should not include any conflicting events. In example 3.3, $E_8(m_4$, “decrease”) which is induced by $E_3(m_2$, “decrease”) is *in conflict with* $E_1^o(m_4$, “increase by 2-fold”).

Def’n *(Conflicting Events):* Given two events $E_i(m_i, c_i)$ and $E_j(m_j, c_j)$, E_i is *conflicst with* E_j if $m_i = m_j$ and $c_i \neq c_j$.

Def’n *(Event Closure Set):* Given an event E_i, let $S^d(E_i)$ be the set of events that are directly induced by E_i. *The event closure set* $S^+(E_i)$ of E_i is the set of all events that are either

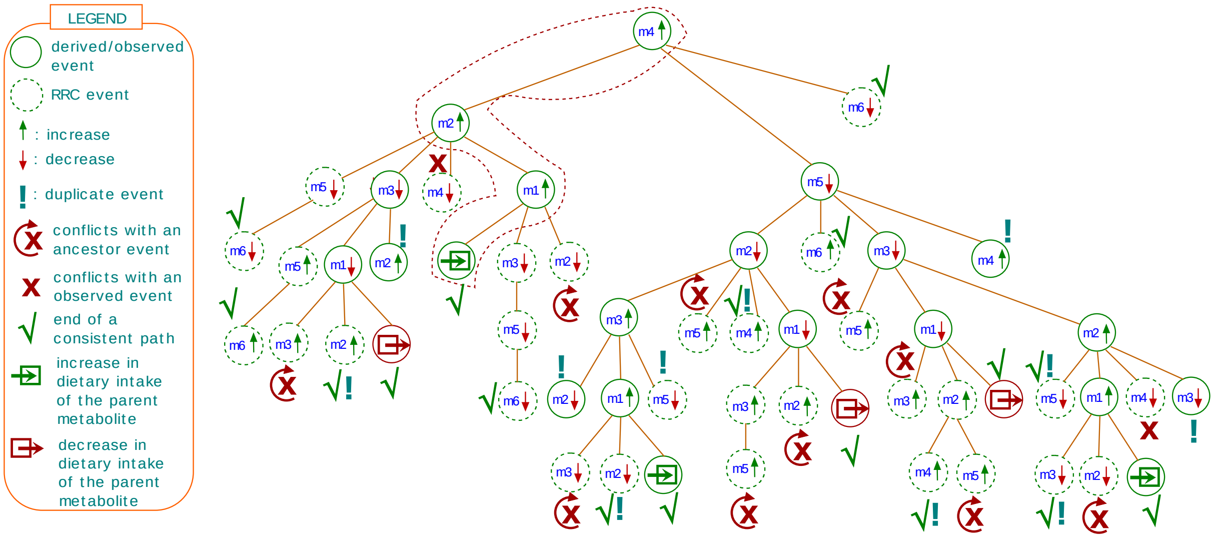

LEGEND
derived/observed event
RRC event
↑ : increase
↓ : decrease
! : duplicate event
conflicts with an ancestor event
X conflicts with an observed event
√ end of a consistent path
increase in dietary intake of the parent metabolite
decrease in dietary intake of the parent metabolite
m4↑
m2↑
m6↓
m5↓
m3↓
m4↓
m1↑
m2↓
m6↑
m3↑
m4↑
m1↓
m2↑
m5↑
m5↓

Figure 4. Closure Tree T+(E_1^o) for Event E_1^o(m4, "increase by 2 folds")

included in $S^d(E_i)$ or in the event closure set of any event E_j ∈ $S^d(E_i)$, that is, $S^+(E_i) = S^d(E_i) \cup [\cup_{E_j \in S^d(E_i)} S^+(E_j)]$ where, for any $S^d(E_k)$, there is no event $E_m \in S^d(E_k)$ such that E_m conflicts with E_k.

Example 3.4. Consider the observed event E_1^o(m_4, "increase by 2-fold") and the metabolic network of Figure 3. Then, the closure set of the event E_1^o is $S^+(E_1^o)$ = {E_3(m_2, "increase"), E_4(m_5, "decrease"), E_5(m_6, "decrease"), E_6(m_1, "increase"), E_7(m_3, "decrease"), E_{12}(m_4, "increase"), E_{13}(m_6, "increase")}.

Next we define a tree data structure that enumerates all derived events from a given observed event.

Def'n *(Closure Tree)*: Given an event E_i(m_i, c_i), the events in $S^+(E_i)$ can be enumerated and represented as a tree $T^+(E_i)$, called *closure tree*, such that (i) the root is E_i(m_i, c_i), (ii) each event in $S^+(E_i)$ corresponds to a node in the closure tree, (iii) given two events E_k(m_k, c_k), E_j(m_j, c_j) ∈ $S^+(E_i)$, E_k(m_k, c_k) is a child of E_j(m_j, c_j) if E_k(m_k, c_k) is a derived event "induced" by E_j(m_j, c_j), and (iv) an event E_j(m_j, c_j) is a leaf node if (a) no other event can be derived from E_j, and/or (b) E_j has an ancestor event E_k(m_j, c_k) defined on the same metabolite m_j where $c_j = c_k$ (duplicate, marked with "!"), and/or (c) E_j has an ancestor event E_k(m_j, c_k) defined on the same metabolite m_j where $c_j = \neg c_k$ (*conflict*, marked with "X"), and/or (d) E_j has an ancestor event E_k(m_j, c_k) defined on the same metabolite m_j where $c_j = c_k$ (*duplicate*, marked with "!").

Figure 4 shows an example closure tree.

3.3 Hypothesis Generation

In this section, we introduce the notion of (low-level) *hypothesis* and define its properties in terms of consistency and minimality.

Def'n *(Consistent Event Set)*: A set S of events said to be *consistent* if there are no two pair of events E_i(m_i, c_i), E_j(m_j, c_j) ∈ S such that E_i(m_i, c_i) conflicts with E_j(m_j, c_j).

Def'n *(Minimal Event Set)*: Given a set S of events, S is said to be *minimal* if there are no two pair of events E_i(m_i, c_i), E_j(m_j, c_j) ∈ S such that E_i(m_i, c_i) = E_j(m_j, c_j).

Def'n *(Consistent Minimal Path)*: Given a root-to-leaf path P and the set S of events on P in a closure tree, P is a *consistent* and *minimal* path if S is consistent and minimal.

Example 3.5. In Figure 4, the path P1 = {m_4↑, m_2↑, m_1↑, ⇥} is a consistent and minimal path, while the path P2 = {m_4↑, m_2↑, m_4↓} is an inconsistent path since the first and the last events in the path are in conflict.

Definition *(Hypothesis):* Given an observed event E_i^o(m_i, c_i) and its closure tree $T^+(E_i^o)$, a root-to-leaf path P in $T^+(E_i^o)$ represents a hypothesis $H(E_i^o)$ if P is consistent and minimal.

Example 3.6. In Figure 4, consider the consistent path {m_4↑, m_2↑, m_1↑, ⇥} that is enclosed in dashed-line borders. P represents one of the several alternative hypotheses for the observed event E_1^o(m_4, "increase by 2-fold"). The hypothesis explains the increase in the concentration of m_4 (i.e., event E_1^o) as follows:

m_4↑ : m_4 increased, since it was produced more, because:

m_2↑: m_2 has increased, i.e., more production due to:

m_1↑: m_1, which used to produce m_2, has increased, since:

⇥ : dietary intake of m_1 has increased.

Note that, in the above hypothesis definition, in addition to consistency, we enforce minimality of a path in order for it to be considered as a hypothesis, mainly, because a hypothesis is a transitive causality relationship between the events that constitute the hypothesis. Hence, having duplicate events in the same hypothesis would lead to the inference that an event is caused by itself, which is not possible without an external cause.

Finally, a candidate hypothesis that is generated to explain an observation should be consistent with the other observations that are included in the same biofluid test.

Definition (*Supporting Experiment*): Given a hypothesis H(E_i^o) for an observed event E_i^o, and a set OE of observed events, H(E_i^o) is said to be *supported* by OE, if there is no pair of events E_j^o(m_j, c_j) ∈ OE and E_k^o(m_k, c_k) ∈ H(E_i^o) such that E_j^o(m_j, c_j) conflicts with E_k^o(m_k, c_k).

Example 3.7. Consider the observed event set OE = {E_1^o(m_4, "increase by 2 folds"), E_2^o(m_1, "increase by 3 folds")}. The hypothesis in Example 3.6 is supported by OE. Nevertheless, the hypothesis that is represented by the path {m_4↑, m_2↑, m_3↓, m_1↓, m_2↑} in Figure 4 is not supported by OE, since the event m_1↓ conflicts with E_2^o.

Problem Statement *(Metabolic Network-based Observation Analysis Problem)*: Given a metabolic network M, and a set OE of observed events obtained from a metabolomics study, the metabolic network-based analysis problem is to compute the set *P* of all (*completeness*) distinct (*minimality*) hypotheses for the observed events in OE such that each hypothesis H ∈ *P* is supported by OE.

Time/Space Efficiency of the Chase Process. Each event in a closure tree may lead to multiple new events (i.e., branches in event closure tree). Therefore, the number of generated hypotheses grows exponentially in terms of the average number of reactions per metabolite in the metabolic network. In more detail, let each metabolite in each tissue/bio-fluid be a distinct node in the graph G (V, E) representing the metabolic network. Then the worst-case time complexity of our approach is the number of paths between the bio-fluid metabolite chosen as the root node of the closure tree and all other nodes in G. In other words, the worst-case time complexity of the OMA method, while exponential, is directly related to the sparseness of the metabolic network. Note that metabolic networks are usually sparse; i.e., the number of edges from a node n represents the number of reactions that the metabolite n participates as a substrate/product which is usually (but not always) a small number ranging from 2 to 5, much less than the maximal number of edges |V|.

3.4 Enhancements on OMA

We develop a number enhancements on the OMA framework to increase its effectiveness. The enhanced features include (i) hypothesis ranking based on expected flux ratio information, (ii) linking hypotheses to known physiological conditions via the overlaps between biomarkers (e.g., diabetes and glucose, or cardiovascular disease and cholesterol) and metabolites in a hypothesis, and (iii) hypotheses set summarization for a more manageable view of the generated hypotheses. Due to the lack of space, we do not discuss these features here. For more details, please see the extended version [29].

4. EXPERIMENTS AND RESULTS

In this section, we study the computational aspects of our metabolomics analysis framework by presenting results on an empirical study of hypothesis generation, elimination, and summarization. We have also applied OMA on the Non-alcoholic Fatty Liver Disease, and successfully produced hypotheses that are consistent with (manually performed) expert analysis. We could not include our results here due to space limitations, but make it available as a technical report [31]. We next describe our data set, and then present the experimental results.

4.1 Testbed: PathCase^{MAW}

Currently, there are many web-based metabolic network data sources, e.g., KEGG [32], Reactome [33], MetaCyc [34], or our own PathCase [35]. However, all of these data sources (with perhaps some exceptions for Reactome) lack location (i.e., tissue/organ, cell, etc.) information for individual pathways. For this study, we have built our own prototype database (*PathCase^{MAW}* [28]) with organ information by manually entering major pathways (mostly, from a biochemistry textbook [26] and an atlas of human metabolism [27]. Please see table 1 for database content.

Table 1. Experimental Database Content

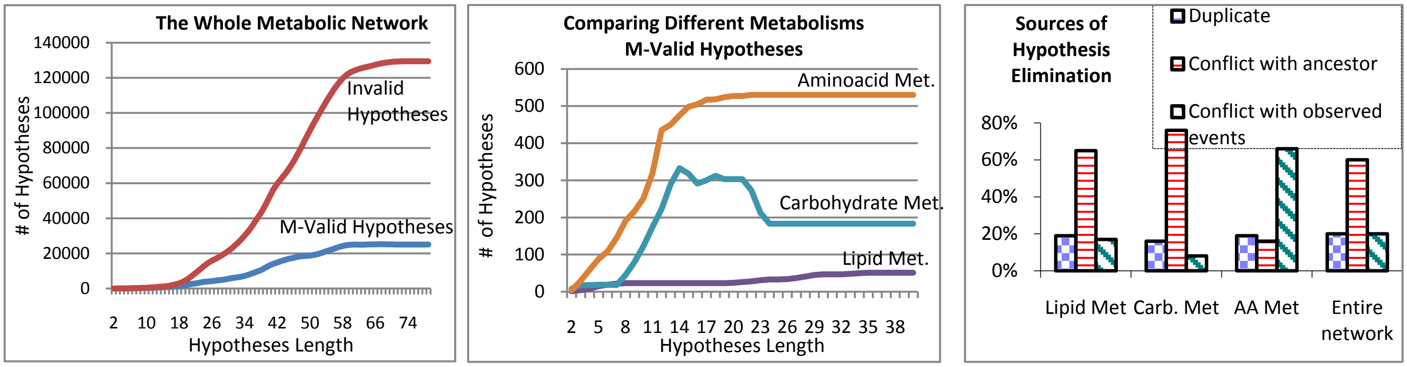
	Amino Acid Metabolism	Carbohydrate Metabolism	Lipid Metabolism	Whole Database
Num. of pathways	28	11	11	50
Num. of processes	118	68	55	241
Num. of metabolites	145	52	70	205
Num. of tissues	5	9	5	9
Num. of graph nodes	476	426	219	980
Num. of pathway links	42	31	5	123

As the metabolomics data set, we have used a subset of the sample metabolomics dataset from [36], extended by ten additional metabolite measurements. More specifically, this dataset contains concentration changes on 34 metabolites:

(a) Amino Acid metabolism: {glutamate↑, pyruvate↑, Isoleucine↑, valine↑, Leucine↑, thyroxine↑, alanine↑, kynurenine↑, Tyrosine↑, lysine↑, glutamine↑, trans-4-hydroxyproline↑, alpha-ketoglutarate↑, Threonine↑, serine↑, creatine↑, phenylalanine↑, citrulline↑, proline↑, histidine↑, glycine↑, Methionine↑, 5-hydroxytryptophan↓},

(b) Lipid metabolism:{glycocholate↑, cholate↑, glycerol↑, palmitate↑, 2-hydroxybutyrate↑, glucose↑, cholesterol↑, linoliate↑, glycine↑ }, and

(c) Carbohydrate metabolism: {pyruvate↑, isocitrate↑, lactate↑, alpha-ketoglutarate ↑, glucose↑}.



(a) (b) (c)

Figure 5. Number of Hypotheses vs. Hypothesis Length

4.2 Results and Discussion

In this section, we present a computational analysis of the proposed framework.

(a) Experiment 1: Number of Hypotheses vs. Length

In this experiment, we study the relationship between the maximum allowed hypotheses length (in terms of the number of events) and the total number of automatically generated hypotheses by the system. In each metabolism, the highest changing metabolite (glutamate in aminoacid metabolism, glycocholate in lipid metabolism, pyruvate in carbohydrate metabolism, glutamate for the whole network) was used as the root of the closure tree. Figure 5 (a, b) depicts the change in the number of hypotheses as we increase the maximum hypothesis length.**Observation 1:** *As the maximum allowed hypothesis length increases, the number of hypotheses generated initially increases exponentially until the maximum hypothesis length reaches a certain value, and from that point on, the total number of hypotheses does not change.*

Since our hypotheses correspond to paths in the whole metabolic network, the number of paths gets exponentially larger as we span over a larger fraction of the metabolic network. However, due to the highly connected nature of the metabolic network, after a certain point (at length 70), the increase in the number of hypotheses slows down, and finally becomes stable as the probability of encountering a metabolite which is already included in a hypothesis increases (due to the stopping criteria at duplicate or inconsistent events during closure tree construction).

Observation 2: *The amino acid metabolism has the highest number of m-valid hypotheses, which is followed by the carbohydrate metabolism, and, lastly, the lipid metabolism has the smallest number of hypotheses*.

The above observation is well-correlated with the size of each particular metabolism in our database. Amino acid metabolism has the largest number of pathways among the three metabolisms. Although the lipid and the carbohydrate metabolisms have the same number of pathways, the carbohydrate metabolism is significantly more interconnected (31 pathway interconnections vs. 5 pathway interconnections) than the lipid metabolism.

Observation 3: *In terms of the ratio of invalidated hypotheses, the carbohydrate metabolism has the highest invalidation rate (90%), which is followed by the lipid metabolism (84%), and lastly, the amino acid metabolism has the lowest hypothesis invalidation rate (61%).*

Having significantly less number of metabolites (i.e., 52) in the lipid metabolism, the probability of creating a duplicate event (causing the elimination of a hypothesis) on the same metabolite during the closure tree generation is much higher in comparison to the amino acid metabolism (i.e., 145). In fact, for carbohydrate metabolism, 92% of hypothesis elimination was due to encountering an already visited metabolite in the network (Fig. 5.c). In contrast, the same ratio in amino acid metabolism was only 35%.

(b) Experiment 2: Invalidated Hypotheses vs. the Number of Observed Events

Eliminating some of the possible hypotheses by utilizing all of the existing observed events through an integrative approach is one of the essential promises of the proposed framework. In this experiment, we investigate the contributions of observed events to invalidate a fraction of possible hypotheses. Figure 6.a shows the total number hypotheses as the number observed events increases.

Observation 4: *Using measurements on 60 metabolites in the database reduces the hypotheses set by 99.9%.*

Observation 5: *As the number of observations gets larger, the number of invalidated hypotheses that we can invalidate increases dramatically, which results in over 94% reduction in the total number of hypotheses when measurements on 30 metabolites are employed.*

We have also repeated the same experiment by creating a random observed event set, where each metabolite in the database had an equal likelihood to be included in the observed event set. Similarly, decrease and increase events for each metabolite were assumed to occur equally likely.

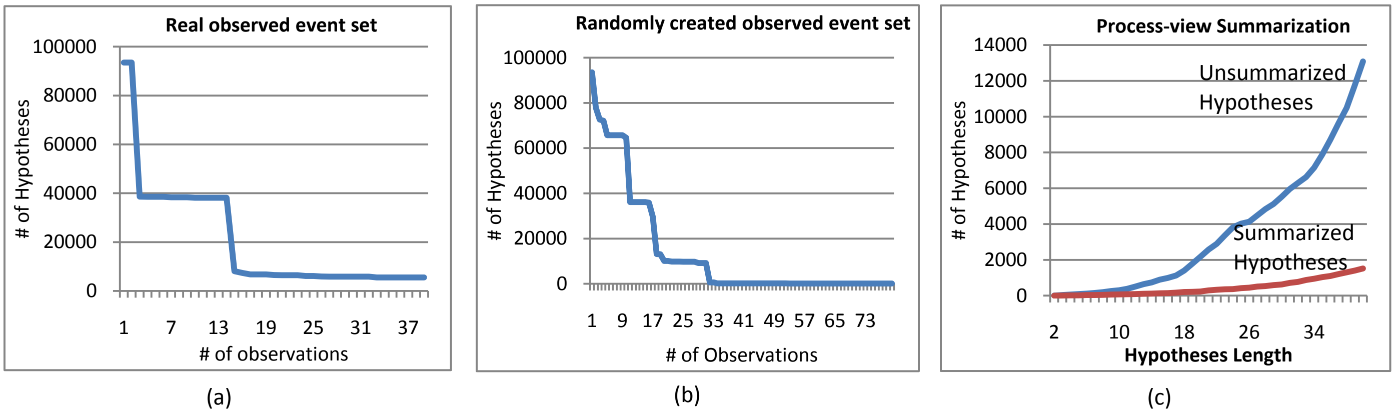


Figure 6. Invalidated Hypotheses vs. the Number of Observed Events, and Summarization

Figure 6.b presents the results for a random event set.

The usefulness of our framework (on a large scale) depends on the number of metabolites that can be measured and made available to the system. Even with 30 measurements from a real data set, we were able to automatically eliminate the majority (95%) of hypotheses that are not consistent with the measurements. Fortunately, the measurement technologies are rapidly getting more sensitive and less expensive, which will result in the elimination of more and more hypotheses (99.9% with 60 events for a random event set) before any manual expert review is performed. Hence, the above observations prove that the proposed metabolic analysis system is promising in terms of helping and supporting researchers in their quests for interpretations of metabolic observations.

(c) Experiment 3: Effect of Summarization

In this experiment, we study the impact of hypothesis summarization. For the lack of space, we only present our observations. Please see [29] for more details.

Observation 6: *Process-view summarization provides a high level condensed view of the whole hypotheses set by reducing the total number of hypotheses by 88%.*

Observation 7: *As the support threshold increases, the size of the summary set dramatically decreases. At support thresholds of 0.1 and 0.2, the summary set is 91% and 98%, respectively, smaller than the original hypotheses set.*

(d) Experiment 4: Running Time Performance Study

In this experiment, we study the running time behavior of our metabolic analysis framework, and the effect of several efficiency enhancements that we have developed during our implementations. For brevity, we only include our observations. Please refer to [29] for the corresponding charts and explanations.

Observation 8: *As the maximum hypotheses length increases, the running time our system initially increases exponentially, and later stabilizes after length 70.*

Observation 9: *The early termination approach is significantly (97%) more efficient than the baseline approach.*

5. CONCLUSIONS

In this study, we have presented models for computationally identifying the mechanisms that produce the observed/ measured metabolite changes. To this end, biologically motivated event derivation rules are discussed to estimate possible concentration changes of metabolites which are not measured. We have proposed a data structure, called the Closure Tree, to derive new event and identify candidate "hypotheses" as explanations for observed concentration changes. Moreover, we have defined notions of consistency and minimality to eliminate hypotheses that are not conforming to the observed events.

We have evaluated our metabolic analysis framework through an empirical study of computational hypothesis generation. Our results show that a majority of generated hypotheses can be invalidated automatically using the provided set of observed concentration changes. Furthermore, summarization greatly helps to create a manageable, yet effective, view of a large hypothesis set.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grants DBI-0849956, DBI-0743705, CRI-0551603, and CCF-0820217.

REFERENCES

[1] Daviss, B. 2005. Growing pains for metabolomics. *The Scientist*, 19[8]:25-28.

[2] Wikipedia Entry (*Metabolomics*), http://en.wikipedia.org/wiki/ Metabolite. (Retrieved on May 13, 2008)

[3] Oliver, SG et al. 1998. Systematic Functional Analysis of the Yeast genome. *Trends BioTechnol.*, Vol. 16, 1998, pp. 373-378.

[4] Harrigan, GG., and Goodacre, R. (Eds). 2003. Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. *Kluwer Academic Publishers*, Boston, USA.

[5] Fell, D.A. (1997). *Understanding the control of metabolism.* (Portland Press, London, UK).

[6] Schilling, C.H., Schuster, S., Palsson, B.O., Heinrich, R. (1999). Metabolic Pathway Analysis: basic concepts and scientific applications in the post-genomic era. Biotechnol. Prog., 15, 296-303.

[7] Stephanopoulas, G., Aristidou, A., Nielsen, J.H. (1998). *Metabolic Engineering: principles and methodologies*. (Academic Press, Maryland Hts, MO).

[8] Trinh, C.T., Wlaschin, A., Srienc, F. (2009). Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism, Appl. Microbiol. Biotech., 81, 813-826.

[9] Hoppe, A. et al. 2007. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol.* 1:23.

[10] Schuster, S., Higetag, S. (1994). On elementary flux modes in biochemical reaction systems at steady state", J. Biol. Syst., 2, 165-182.

[11] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N.et al. (2006). COPASI - a COmplex PAthway Simulator, Bioinformatics 22, 3067-74.

[12] Pfeiffer, T., Sánchez-valdenebro, I., Nuño, J.C. et al. (1999). METATOOL: for studying metabolic networks, Bioinformatics, 15: 251-257.

[13] Urbanczik, R. (2006). SNA-A toolbox for the stoichiometric analysis of metabolic networks. BMC Bioinformatics, 7: 129.

[14] Klamt, S., Stelling J, Ginkel, M., Gilles, E.D. (2003). FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. Bioinformatics, 19, 261-269.

[15] Schwartz, J.M., Gaugain, C., Nacher, J.C., De Daruvar, A., Kanehisa, M. (2007b). Observing metabolic functions at the genome scale. Genome Biol, 8, R213.

[16] Wlaschin, A.P., Trinh, C.T., Carlson, R., Srienc, F. (2006). The fractional contributions of elementary modes to the metabolism of Escherichia coli and their estimation from reaction entropies, Metabolic Eng., 8, 338-352.

[17] Urbanczik, R., Wagner, C. (2005). An improved algorithm for stoichiometric network analysis: theory and applications. Bioinformatics, 21, 1203-1210.

[18] Cakmak, A., Ozsoyoglu, G., Hanson, RW. Querying Metabolism under Different Physiological Constraints. *Journal of Bioinformatics and Computational Biology,* 8:(2) pp. 247-293, April 2010.

[19]Klamt, S., Stelling, J. (2003). Two approaches for metabolic pathway analysis? Trends in Biotech., 21, 2, 64-69.

[20] Glykys, D.J., Banta, S. (2009). Metabolic Control Analysis of an enzymatic biofuel cell. Biotechnology and Bioengineering. 102 (6), 1625-1635.

[21] Kuipers, B.J. 1993. Reasoning with Qualitative Models. *Artificial Intelligence* 59, 125-132.

[22] Kuipers B.J. 2001. Qualitative Simulation. In *Encyclopedia of Physical Science and Technology,* 3rd edn. *Academic Press*, 287-300.

[23] Forbus, K.D. 1984. Qualitative Process Theory. *Artificial Intelligence* 24. 85-168.

[24] Karp, P.D. 1993a. Design Methods for Scientific Hypothesis Formation and Their Application to Molecular Biology. *Machine Learning,* 12, 89-116.

[25] Wishart, DS. 2007. Current progress in computational metabolomics. *Briefings in Bioinformatics* (8) 5: 279-293.

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

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

[28] Cakmak, A., Dsouza, A., Hanson, R.W., Ozsoyoglu, M., Ozsoyoglu, G. A Web-Based Data Source for Metabolomics. *ISCIS*, 2009. Available online at: http://dblab.case.edu/PathwaysMetabolomics/Web/

[29] Cakmak, A. et al. (2010). Analyzing Metabolite Measurements for Automated Prediction of Underlying Biological Mechanisms. Technical Report. Dept. of EECS, CWRU, Cleveland, OH.

[30] Wishart, DS et al. (2008). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research,* doi:10.1093/nar/gkn810, October 25, 2008.

[31] Cakmak, A. et al. (2010). Automated Metabolomics Data Interpretation: A Case Study on Non-Alcoholic Fatty Liver Disease. Technical Report. Dept. of EECS, CWRU, Cleveland, OH.

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

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

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

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

[36] Lawton, K.A., Berger, A., Mitchell, M. et al. (2008). Analysis of the adult human plasma metabolome. Pharmacogenomics Apr;9(4):383-9.