RESEARCH STRATEGY

A. BACKGROUND

1. Complimentary Analytical Platforms Generate Vast Metabolomics Data: A deep understanding of network and pathway regulation in health, disease and response to treatment, requires an ability to identify and quantify hundreds to thousands of small molecules, which are the building blocks of such networks. A variety of analytical platforms have been developed over the past decade that enable measurements of small molecules from primary and intermediary metabolism, of exogenous compounds, such as drugs and other xenobiotics, and of metabolites produced by the gut Microbiome (Kaddurah-Daouk 2008,2009). No one platform can capture the biochemical complexity and heterogeneity of all molecules and hence the need to use a variety of analytical platforms in a complimentary way. GC-MS, LC-MS, and NMR-based metabolomics platforms are suited for mapping global biochemical changes in nontargeted ways; LC-electrochemistry array metabolomics platform (LCECA) is an example of targeted metabolomics platform excellent for mapping neurotransmitter pathways and pathways related to oxidative stress (Kristal 2007). Some analytical platforms can provide quantitative measures of metabolites others provide semi-quantitative or relative concentrations. All of these issues make horizontal integration of data rather difficult but yet an important step towards providing biological insights.

2. Bottlenecks in Data Interpretation: aiming at studying the entire metabolic network, metabolomics possesses a unique potential to address major scientific questions from a variety of biomedical research areas. Examples are (a) what is the impact of an environmental stimulus, e.g., drug treatment, to our metabolism; (b) can we define prognostic, diagnostic, and surrogate markers for a phenotype, e.g., a disease status or drug response; (c) can we make accurate prediction of the phenotype based on these biomarkers; (d) can essential information be derived about underlying mechanisms of the phenotype? Some of these questions, to a certain degree, can be translated into classical bioinformatics and biostatistics problems to be addressed by well-established methods (Kaddurah-Daouk et al., 2008; Lavine and Workman, 2010; Madsen et al., 2010). Steady progress has been made in the improvement of these classical methods (Lavine and Workman, 2008, 2010) mostly through adopting different mathematical and statistical techniques for the sake of better visualization, prediction performance, and/or interpretation of results. Additionally, the evolution of other theoretical and applied areas, such as network modeling, casual inference, and the analysis of other types of omics data, can also benefit this area.

Although the integration of existing powerful computational tools with metabolomic data has generated a large number of successful stories, bottlenecks still exist. For instance, 1) Can we integrate metabolomics data derived from different metabolomics platforms (horizontal integration) so that we can get a comprehensive view of how metabolism was modified in disease or in response to treatment? 2) Realizing the tremendous amount of information hidden in a metabolomics dataset, can we develop an integrated strategy based on available numerical tools that can guide a comprehensive mining of metabolomic data and extract the maximum information contained within a given dataset? 3) What are ways to improve this strategy, e.g., how to shift the current common practice of examining individual metabolite effects to pathways-centric analyses? 4) How to integrate numerical approaches with our knowledge on the metabolic network to improve the accuracy of estimation, and power of hypothesis testing; 5) Pursuing 4 further relies on how to represent such knowledge; 6) How to expend such knowledge for a better understanding of the underlying biological processes of a certain phenotype; 7) Can we de-convolute stable isotope labeled metabolomics data and derive more detailed insights about mechanistic questions? 8) How do we learn kinetic and dynamic perspective on the metabolomics states; 9) Can we integrate metabolomics data with other layers of data such as genetics data (vertical integration) to provide us with more information regarding a particular biological mechanism and/or system? Our proposal attempts to address some of these challenges to enable progress in this important area.

B. PRELIMINARY DATA

Highly Experienced Team Working Collaboratively and Tackling Bottlenecks in Metabolomics Research: 1. Metabolomics Research Network brings relevant experience and rich datasets: During the past five years support from NIGMS (Dr. Kaddurah-Daouk- PI) enabled the creation of the Pharmacometabolomics Research Network (http://www.pharmacometabolomics.org). The network has pioneers in the field of metabolomics**,** experts in clinical research, informatics, pathway analysis, biochemical modeling, as well as experts in building metabolomics databases. The network has pioneered application of metabolomics and bioinformatics in the medical field and provided novel insights about pathways implicated in disease and in variation of response to treatment. Examples include 1. Establishing a new concept that a patients metabolic profile (metabotype) defines trajectories of response to treatment with antidepressant and placebo and enables sub classification of complex disease like depression (Kaddurah-Daouk et al., 2011; Ji et al., 2011); 2. The gut microbiome is implicated in mechanism of variation of response to statin (Kaddurah-Daouk et al., 2011); 3. Metabolic profiles highlight novel pathways implicated in schizophrenia and its treatment (Kaddurah-Daouk et al., 2007; Yao et al., 2010, 2011) and insights into pathogenesis of Alzheimer's Disease (Kaddurah-Daouk et al., 2011). Rich datasets exist from the profiling of human samples (plasma, serum, CSF) some include labeling with a stable isotope, such as ^{13}C glucose. Many of the samples are profiled on different metabolomics platforms generating vast biochemical data and an opportunity to explore development of tools for horizontal data integration. Additionally, for many of the subjects studied there is vast additional data including genome-wide association data, which presents an excellent opportunity to explore vertical data integration. Data also exist from model systems such as cell culture lines and animal models. For example "Human Variation Panel" (LCLs) were used to generate 1.3 million genome-wide SNPs, 54,000 expression array probe sets, whole genome methylation data, and genome-wide microRNA data as well as multiple drug response phenotypes for each cell line. Metabolomics data is being added to this rich layer of data to address mechanistic questions related to therapeutic effects of drugs. **2. Rich experience in mining metabolomics data and developing novel statistical tools:** In the last five years, the statistics team from NC State (lead by Dr. Zeng) and the Duke team (Dr. Zhu) have been working with the investigators in the Pharmacometabolomics Research Network on various projects in metabolomics data analysis and were responsible for identifying new insights about mechanisms of variation of response to drugs such as SSRIs (Ji et al., 2011; Kaddurah-Daouk et al., 2011) and Statins (Kaddurah-Daouk et al., 2010, 2011) and for defining signatures that highlight heterogeneity of disease (unpublished results) Their work leads to the totally novel concept "Metabolomics Informs Pharmacogenomics" a major contribution to effort in personalized therapy (Ji et al., 2011). Through these data analyses, they have also customized a number of routine statistical tools and developed new methods for metabolomics analyses. More importantly, they have further developed a series of protocols for the data analysis based on tools available and our rich knowledge on the various types of information that can be extracted from a metabolomics dataset. This lays a solid foundation to develop a comprehensive data mining strategy to tackle the major bottleneck in data interpretation encountered by the metabolomics research community. Moreover, the team is conducting active research on developing novel statistical methods to enrich the arsenal of metabolomics data mining tools. For example, they have developed a framework of dimension reduction methods (Zhu and Li, 2011), which are capable of capturing complex nonlinear effects of metabolites within a metabolic pathway for a clinical phenotype of interest. This development will facilitate an enhanced understanding of underlying biological mechanisms. **3. Leading effort in Genome-Scale Metabolic Network Reconstruction:** The Icelandic group (Thiele and Fleming) have lead a community effort to generate the most comprehensive network reconstruction available for human metabolism (Recon2). Recon2 consists of 6852 biochemical and transport reactions distributed over eight intracellular compartments, 2469 unique metabolites, and enzymes encoded by 1766 genes. A high quality, manual curated reconstruction is an essential prerequisite for successful *in silico* prediction of *in vivo* physiological states. Recon2 is a fundamental tool for the study of the systems biology of human metabolism as it represents a generic model representing the metabolic capabilities of any human cell, and thus requires further tailoring with data, e.g., metabolomics data, in order to compute hypothesis for a specific organ or tissue. Particularly relevant to the current proposal is that Recon2 is saturated with unique metabolite identifiers, allowing accurate bioinformatic integration of metabolomic data. **4. The Netherlands Metabolomics Center** (NMC): Dr. Hankemeier, brings a very comprehensive metabolomics analytical capability, with largest investment in metabolomics research in Europe (over fifty million Euros invested to date). They provide excellent capabilities to tackle issues related to integration of data derived from different metabolomics platforms (horizontal data integration). **5. UCSD:** His laboratory has developed the biology workbench (Subramaniam, et al., 1998), the signaling gateway (Dinasarapu, 2011), and the lipidmaps gateway (Subramaniam, et al., 2011). In addition, they have developed several sophisticated computational strategies for vertical integration and dynamical modeling of "omics" data (Asadi et al., 2010; Dennis et al., 2010; Dinasarapu et al., 2011; Gupta et al., 2007, 2009, 2010, 2011; Hsiao et al, 2009; Papin et al., 2005; Sburamaniam, et al., 2011).

C. SIGNIFICANCE

While progress has happened in analytical chemistry that enabled the creation of powerful metabolomics platforms that can yield measurements of vast number of metabolites, the ability to generate biological knowledge out of this complex data has lagged behind. This is one of the biggest bottlenecks in metabolomics research today. No one analytical platform can provide full coverage of the metabolome and hence the need to use complimentary analytical platforms with the need for tools to enable horizontal integration of metabolomics data to derive biological insights. In addition, large metabolomics data sets now exist where there is added genomic, transcriptomic and phenotypic data that needs to be validated, analyzed and integrated to provide biological knowledge that will further serve as important hypotheses for experimental investigations and provide systems models of diseases. There is pressing need to develop bioinformatics infrastructure and tools to achieve these goals. We bring pioneers in the field of bioinformatics and metabolomics who have been collaborating closely for several years and bring a complimentary set of skills to work closely with clinical groups to address bottlenecks in the field and create bioinformatics tools that can help lead to rapid advances in the field. We propose to develop an analysis pipeline that will address issues of horizontal integration of metabolomics data derived from complementary analytical platforms; development and utilization of statistical validation and analysis of data tools, genomics-scale reconstruction of human metabolism, dynamic and kinetic metabolic network modeling, and vertical integration with other “omics” data. The analysis pipeline will be biologist-friendly and will be made accessible to the metabolomics community, and the larger biomedical research community. We believe these developed tools will enable rapid advances in the metabolomics field towards applications that can impact medical practice.

D. INNOVATION

In terms of innovation this proposal: 1. integrates existing methods for rigorous and quantitative estimation of metabolites measured in “omics” experiments; 2. Buids an easily-accessible guideline for a systematic and in-depth mining of metabolomics data; 3. Builds novel statistical methods for genome-scale analysis of metabolomics data to enrich the guideline; 4. Develops new software for genome-scale reconstruction of human metabolic network; 5. Develops methods for analysis of isotopomeric data, toward validation of existing mechanisms and identifying novel mechanisms in pathways; 6. Develops sophisticated statistical methods for integrative analysis of “omics” data; 7. Develops methods for dynamic modeling of integrated metabolomics data; 8. Provides a workflow for reconstruction, analysis and modeling of mechanisms and pathways that lead to defined phenotypes involving metabolic changes.

E. APPROACH

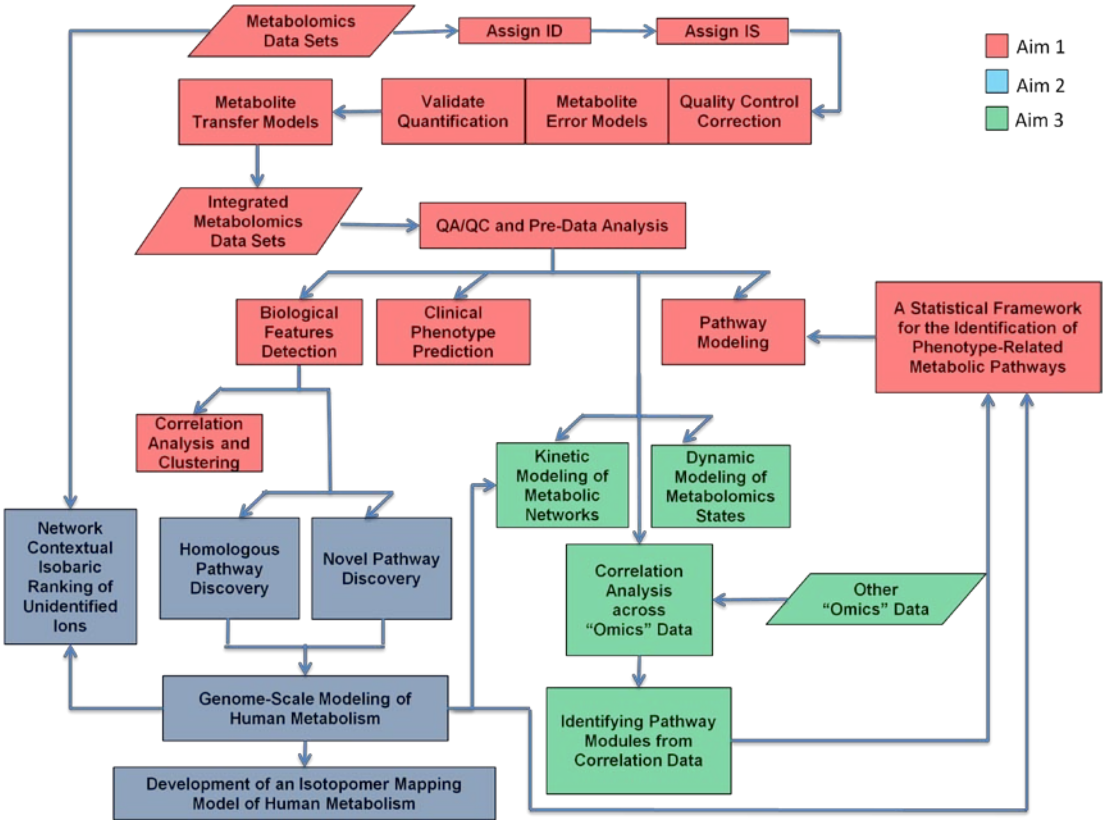
Metabolomics Data Sets
Assign ID
Assign IS
Aim 1
Aim 2
Aim 3
Metabolite Transfer Models
Validate Quantification
Metabolite Error Models
Quality Control Correction
Integrated Metabolomics Data Sets
QA/QC and Pre-Data Analysis
Biological Features Detection
Clinical Phenotype Prediction
Pathway Modeling
A Statistical Framework for the Identification of Phenotype-Related Metabolic Pathways
Correlation Analysis and Clustering
Kinetic Modeling of Metabolic Networks
Dynamic Modeling of Metabolomics States
Network Contextual Isobaric Ranking of Unidentified Ions
Homologous Pathway Discovery
Novel Pathway Discovery
Correlation Analysis across “Omics” Data
Other “Omics” Data
Genome-Scale Modeling of Human Metabolism
Identifying Pathway Modules from Correlation Data
Development of an Isotopomer Mapping Model of Human Metabolism

Figure 1: A schema showing the integrated approach of this proposal.

1. Overall Approach: Figure 1 captures our integrated approach. It is important to realize that no one individual or one group can address bioinformatics bottlenecks for metabolomics research. It is more efficient to integrate the existing expertise of this group, than to attempt to establish their combined expertise in another manner. We propose to build an integrated platform that includes integration of data from multiple complimentary platforms, a comprehensive strategy for systematic mining of metabolomic data that can lead to mechanistic insights; use of both parts list of significant metabolites derived from last step and extended knowledge and annotation of genes to develop genome-scale network reconstruction; validation of the developed networks through the use of isotope analysis, which will develop context specific metabolic maps; and finally, analysis of built networks and kinetic and dynamic network modeling in order to understand the regulation of metabolic networks, which will in turn provide further insights into experimental design. **2. Overall Coordination: Two Co-PIs bring metabolomics and bioinformatics experiences:** The Co-PI’s Dr. Kaddurah-Daouk (Duke Medical Center) and Dr. Zeng (NC

State University) have complimentary backgrounds in metabolomics and bioinformatics. They have been collaborating closely for more than four years through the Metabolomics Research Network, funded by the NIH. They have gained significant experience in managing large networks. They will jointly be responsible for oversight of the entire program and development and implementation of all policies, procedures and processes. *Dr. Kaddurah-Daouk* trained at Johns Hopkins (where she worked with Nobel Laureate Hamilton Smith), followed by appointments at the Massachusetts General Hospital and MIT. She is currently Associate Professor at the Duke Medical Center and has been a seminal force in the development and evolution of the metabolomics field. Dr. Kaddurah-Daouk has extensive experience in assembling teams of researchers to work collaboratively on large scientific projects and has led scientific programs from an early stage of discovery through clinical trials. With funding from NIGMS she established the Metabolomics Research Network with a mission to integrate metabolomics in clinical pharmacology and use of metabolomics data in efforts to personalize treatment (www.pharmacometabolomics.org). The network has over ten academic centers involved and includes co investigators in this application. *Dr. Zeng* is the William Neal Reynolds Distinguished Professor and director of Bioinformatics Research Center, North Carolina State University. He has worked on various research problems in quantitative and statistical genetics over 20 years.. With the progress of other “omics” technologies, he also worked on statistical methods for gene expression QTL analysis that connects DNA polymorphisms to gene expressions and to phenotypes. Over the past four years he has taken an active role within the Metabolomics Research Network and worked closely with its PI Dr. Kaddurah-Daouk to develop methods to mine metabolomics data to define metabolomic signatures and establish metabolomic pathways implicated in disease and response to treatment. **3. Management of Project - Overall Strategy:** The two Co-PIs will closely coordinate the management of the project as they have done over the last four years through the Metabolomics Research Network. Dr. Kaddurah-Daouk will coordinate all activities related to metabolomics including metabolomics datasets for the bioinformatics team and ensuring close interactions between bioinformatics team and the metabolomics groups generating profiling data as well as clinical and basic researchers who are providing biological samples. Dr. Zeng will coordinate activities among bioinformatics teams and help integrate approaches to optimize development of tools that can most effectively help the metabolomics community**. 4. Process for Making Decisions on Scientific Direction:** We are fortunate that several of the team members involved in this application including the metabolomics, and the informatics groups have been working closely with Drs. Kaddurah-Daouk and Zeng through the NIH-funded Metabolomics Research Network a network that is multidisciplinary and multi-institutional. A lesson already learned during “pilot” studies conducted with support from the NIH that have sustained the formative stages of the proposed collaboration is that for the successful implementation of a project, like the one proposed here, it is critical to value the contributions of and actively engage scientific team members, since capabilities they bring are highly complementary and together enable achievement of goals proposed. Creating and sustaining ongoing scientific dialog among all contributing members within the project will be essential for the success of each project. These close interactions will be ensured via regular monthly conference calls, and two annual meetings to review progress and integrate knowledge gained. A process required to establish consensus with regard to research directions will be established through healthy debates to derive goals by all participating investigators.

5. Approaches to Address Specific Aims

Aim 1: Development of Statistical Methods and Software to Systematically Mine Metabolomics Data and Enable Horizontal Data Integration

In this section we propose to create an integrated capability and tools for metabolomics data analysis. We aim to a) develop approaches for horizontal integration of metabolomics data derived from different analytical platforms (such as GC-MS, LC-MS, NMR or LCECA) to enable researchers who derive metabolomics data from different analytical platforms to integrate and maximize biochemical information derived from use of complimentary platforms; b) develop a systematic and comprehensive metabolomic data mining strategy and software platform with a user-friendly graphical interface to enable metabolomics researchers derive biological insights, c) develop statistical framework for the identification of metabolic pathways implicated in the mechanism of disease or drug response. The NC State and Duke team have already taken major steps to tackle many of these issues and have made initial contributions to demonstrate feasibility and impact (see preliminary data section). Below we describe our approach to address this specific aim.

Aim 1a: Horizontal Integration of Metabolomics Data Derived from Different Analytical Platforms

In data integration or fusion, often three types of combination of data from a common set of objects are considered: high-level fusion, which is the combination of results of statistical data analyses obtained on sets of different variables, low-level fusion, or the concatenation of data matrices in such a way that the objects are the shared mode, and mid-level fusion, a term used to describe the combination of variables selected from different data sets. Combination at low level allows maximal flexibility in the choice of the subsequently applied (multivariate) data analysis methods yielding results for the combined data sets and allows deposition of the data into data repositories (e.g. Reijmers et al, 2005). In metabolomics low-level fusion is possible when the data sets to be combined all contain quantitative data. However, currently obtaining quantitative data from metabolomics experiments is still rather difficult, because due to the absence of reference standards for all detected compounds it is impossible to create a complete calibration model per compound. Expertise gained within the Netherlands Metabolomics Centre (the *Data Fusion* and the *Platform Independent Quantification* projects) will be essential to ultimately combine the strengths of the different platforms so the different readouts of the same biological process are integrated. In the coming 4 years we will tackle all the issues listed below to facilitate proper horizontal integration. Mainly low-level fusion approaches will be investigated for integrating metabolomics data derived from different analytical platforms. To enable low-level data fusion, the quality and/or quantification level of the individual metabolomics data sets need to be improved. Depending on the type of metabolomics platform used (targeted vs. non-targeted, quantitative vs. semi-quantitative) different issues will be addressed.

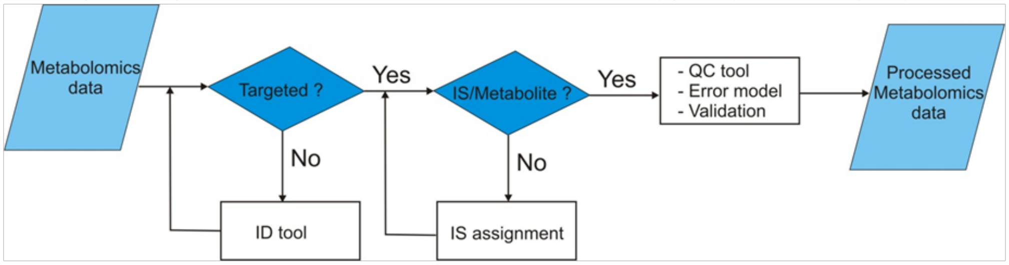

Figure 2: Flowchart of improving the quantification and quality level of metabolomics data.

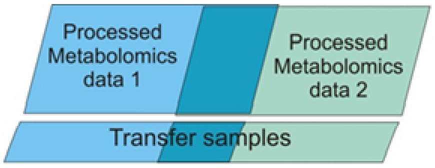

Figure 3: Low level metabolomics data integration using transfer samples.

Targeted Quantitative Platforms: Targeted quantitative platforms typically have for each metabolite measured a dedicated internal standard (IS) available that is used for quantification. *QC, replicates & reference samples:* The first approach that will be examined is the removal of as much as possible measurement variation in an effort to increase the reliability of the obtained signal for each metabolite in the individual metabolomics data sets (improve intra-platform accuracy or analytical variation) prior to merging them into one data set. Addition of extra samples at specific positions in the measurement series (quality control (QC) samples, different replicates) and/or adding compounds to the individual samples are possible ways to monitor and also correct for unwanted variation in the data (Hendriks et al, 2011). Per measurement batch, QC samples are analysed between the real biological samples to correct within-batch and batch-to-batch variation (van der Greef et al, 2007; van der Kloet et al, 2009). Research needs to be done what the most suited QC samples are and if there is a need for multiple different types of QC samples and QC samples having different concentration levels. To allow comparison between studies reference samples need to be included in the measurement design (Draisma et al, 2010). If reference samples are not available re-measuring cross-study samples would be an option but the question is here which samples to re-measure and how to analyse the results to enable comparison between studies? We will deliver an optimal QC, replicate and reference sample strategy that allows monitoring and correcting the unwanted variation present in the different metabolite data sets. *Error model per metabolite:* The second approach that will enable horizontal integration is to estimate the measurement error per metabolite per analytical platform and use this information in the subsequent data analysis of the low level, concatenated data matrices. Estimation of an error model per metabolite is possible using the classical QC approach (van der Kloet et al, 2009) or the method recently developed within NMC (van Batenburg et al, 2011) but the ultimate use of these error models for horizontal integration has not been addressed yet. Insufficient chromatographical separation and ion suppression are two ways in which concentration changes in one compound can cause variance in the measurement of a different compound. To monitor and eventually correct for ion suppression, post column addition experiments are envisioned and will be analyzed leading to improved estimation of the error models. *Validate quantitation:* In addition, in normal practice, calibration lines are included per batch to monitor the overall performance of the analytical method, and possibly to control for that. Different ways of generation calibration models will be examined, either via adding a non-endogenous (often isotope-labeled metabolite) internal standard (IS) or via the standard addition method (adding different concentrations of non-labeled endogenous metabolites). To check the quantitative performance of the metabolomics platforms they will be validated against quantitative results obtained from clinical measurements. **Targeted (Semi-) Quantitative Platforms with a limited number of internal standards:** In addition to all issues raised for the platform above the following issue will be addressed. *IS* *assignment:* We will examine the use of a set of internal standards, ranging from one isotopically labeled metabolite for each metabolite to one internal standard per metabolite class. Additionally uniformly metabolically labeled microorganism cell extracts will be tested as potential quantification candidates. Based on statistical analysis of QC and replicate samples a strategy is developed which IS to use for quantifying which metabolite. In addition, the feasibility of estimating response factors via NMR analysis of modeling of response within a metabolite class will be studied. **Non-Targeted Semi-Quantitative Platforms with a limited number of internal standards:** In addition to all issues raised for the platforms above the following issue will be addressed. *ID pipeline:* In the non-targeted profiling approach all 'peaks' are detected as features and reported as a combination of retention time and mass (preferably elemental composition). With a hybrid ion trap-Orbitrap MS system we are able to get very detailed and reproducible fragmentation information using MS^n experiments. Algorithms have been developed within NMC (Rojas-Cherto et al, 2011 & Peironcely et al, 2011) to compare fragmentation spectra of unidentified metabolites detected by different platforms and/or labs which ultimately are part of the NMC identification pipeline. Using this pipeline we will characterize a number of selected unknown compounds to facilitate further quantification and as such horizontal data integration. **Horizontal metabolomics data integration using transfer samples:** The last topic we will work on is not really connected to a certain type of platform. It deals with the use of reference samples to build transfer models for overlapping compounds between different analytical platforms (not to be confused with the use of reference samples for comparing the results of different studies measured on one single analytical platform). NIST offers one blood sample as reference, but for each sample type (serum, EDTA plasma, citric plasma, tissue extract) a reference sample is needed to build transfer models to combine different analytical platforms. Repeated measurement on different analytical platforms of a set of reference samples differing in metabolite concentrations followed by subsequent multivariate data analysis with e.g. MLPCA (maximum likelihood Principal Components Analysis) allows possible estimating of the transfer models for the non-overlapping correlating metabolites (e.g. metabolites connected via biological pathways). We will study both options. Especially the way how to handle this kind of data is relatively new in the field of metabolomics and needs further attention.

Aim 1b: Development of a Systematic Metabolomics Data Analysis Strategy Implemented in Software for the Mining of Metabolomics Data

The primary goal of Aim 1b, will be the development of a systematic and comprehensive metabolomic data mining strategy, that will be implemented into an extendable software platform with a user-friendly graphical interface. In the last few years, we have been working with the investigators in the Pharmacometabolomics Research Network on various projects in metabolomics data analysis (Ji, et al., 2010; Kaddurah-Daouk, et al., 2010, Kaddurah-Daouk, et al., 2011a, 2011b). Through these data analyses, we have also customized a number of routine statistical tools and developed new methods for metabolomics analyses. More importantly, we have further developed a series of protocols for the data analysis based on tools available and our rich knowledge on the various types of information that can be extracted from a metabolomics dataset. By integrating these existing protocols with our persistent efforts in developing and implementing novel tools for the mining of metabolomics data from all angles, the strategy will eventually mature to a guideline for metabolomics data mining and interpretation. Below we provide major components of the proposed strategy.

Initial Raw Data Processing, Quality Control and Pre-Data Analysis: The first analysis we will perform is to process the raw data for a proper data quality assurance and quality control (QA/QC). For instance, we will compare and select statistical approaches, e.g., unsupervised dimension reduction methods, to search for subject outliers and possible miss-annotations. Strategies will also be taken on missing data at this time, depending on the missing mechanisms. We will also perform a number of analyses to learn the impact of the distributions of metabolites on subsequent data analysis to determine whether there is a need for data transformation. **Biological Features Detection**: An important initial analysis for most metabolomics studies is to identify biological features (metabolites or clinical variables) that are significantly changed by a particular environmental stimulus and/or associated with a particular biological or clinical phenotype of interest. A number of statistical analyses, e.g., t test and regression analysis, will be performed for this purpose in a systematic way. The information will be used to compare with other more sophisticated analysis results to help to understand and interpret the study results. For the measure of test significance, we will report both p-value and q-value (which accounts for the simultaneous multiple testing issue) of a statistical test. Our strategy will also consider non-linear marginal effect of metabolites, which will be inspected and modeled by some statistical techniques, e.g., smoothing. Biological signatures can be quite different among sub-populations defined by one or more clinical variables, e.g., male and females, or African Americans and Caucasians. The various signals can negate the other if the sub-populations are merged and their difference is not properly treated. We will

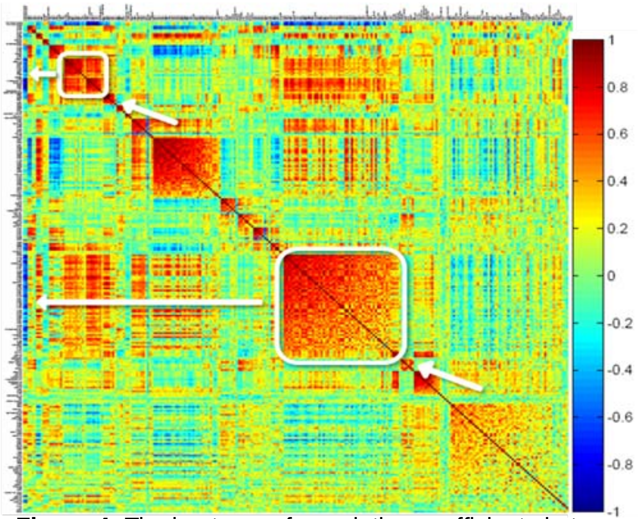

Figure 4: The heatmap of correlation coefficients between metabolites and a clinical phenotype (in the first row and first column) and among metabolites (the remaining rows and columns). The metabolites are grouped according to a cluster algorithm of Stone and Ayrole (2009). The two blocks circled in white show the metabolite clusters that are highly correlated among themselves and are significantly correlated with the phenotype.

determine the differences either from prior knowledge or from data analysis by using appropriate approaches, e.g., analysis of variance (ANOVA). Potential biological features are not limited to the absolute concentration of individual metabolites. For instance, changes in ratios (Yao et al., 2010) and correlations among metabolites can potentially reveal alternations in activities of underlying enzymatic and regulatory pathway activities. Statistical approaches will be shaped in our strategy for the various types of biological features according to their statistical properties. **Study of the Relationship among Biological Features:** We typically present the pair-wise correlation structure among change in metabolites in heat-map diagrams (figure 4). In such a heat-map, the order of metabolites can be arranged either according to their pathway relationship and distance or according to some clustering algorithms. In the former case, one can visualize whether metabolites in the same pathway tend to be more correlated in the study sample and/or respond similarly to an environmental stimulus; and in the later case, one can identify which metabolites are highly correlated in blocks, regardless of pathway information. Clusters of metabolites will be also compared among different subjects groups. **Prediction of a Biological Phenotype Using Subjects' Metabotype:** One of the most important objectives of many metabolomics studies is to determine how the metabotype can be used to predict a particular phenotype under study. What is also of keen interest is to learn which part of the metabolome the predictive power comes from, which may shed light on the hidden mechanism of the phenotype of interest. Our strategy will select and combine approaches depending on several factors: the number of metabolites compared to sample size; whether we want to capture non-linear effects of metabolites; the type of the phenotype to be predicted. The predictive power of different models will be assessed and compared by an iterative sampling and cross-validation method. **Test of Association of Metabolic Pathways with a Biological Phenotype**: An imperative act after the aforementioned analysis is to shift from examining individual metabolite effects to pathways-centric analyses. As clinical phenotypes, e.g., disease symptoms and drug response, tend to involve heterogeneous behavior of metabolic pathways that are composed of complex interactions among metabolites, it is more efficient and effective to examine the effect of entire metabolic pathways on phenotypes. For this purpose, we have developed a two-step procedure (Zhu and Li, 2011) based on a nonlinear dimension reduction framework to aid us to identify phenotype-related metabolite pathways for metabolomics studies. In aim 1c, we will continue to develop novel approaches for pathway selection with better properties. **Statistical Tool Development:** We will develop a comprehensive package for metabolomics data analysis in R, and deposit our package in the R project for general use. The reason to use R for package development is that R is a free software environment for statistical computing and graphics, and has been used extensively for genomics tool development by the community. There are many R routines and resources available and the community of R literate scientists is growing. The package will utilize a number of statistical tools generally available in R, and will include a number of new methods and procedures specifically for metabolomics data analysis, such as cluster algorithms, heat map construction and display, dimension reduction analysis, pathway selection, model construction, and pathway assessment and test. The package will be developed for a variety of metabolomics studies and data structures, and will be structured as a metabolomics data analysis pipeline, rather than just a collection of individual routines. As we proceed with more metabolomics studies and encounter with more problems, we will develop more procedures and tools for the pipeline. This metabolomics R package will provide a platform for the community both for data analysis and for software development.

Aim 1c: Development of a statistical framework for the identification of metabolic pathways implicated in the mechanism of disease or drug response

We plan to develop a class of Bayesian supervised dimension reduction approaches to perform pathway-based analysis. The proposed solutions have several features that are expected to greatly facilitate the mining of metabolomic data. First, supervised dimension reduction (Cook, 1998) is employed for aggregation of multivariate metabolites, which produces a smallest number of combinations of metabolites that retain maximum amount of information between the response and the high-dimensional, potentially correlated metabolites. Second, a Bayesian framework provides a flexible and systematic platform to incorporate prior biological information, such as the pathway structure and correlations among metabolites (Monni and Li, 2010; Tai et al, 2011). Utilization of such knowledge can not only improve estimation accuracy but also facilitate results interpretation. Our approach takes metabolite measurements, clinical response, and metabolite network and pathway information as input, and outputs a posterior probability for each metabolic pathway/individual metabolite that evaluates how likely a particular pathway/metabolite is associated with the clinical phenotype of interest.

Our work is built upon the Bayesian dimension reduction (BDR) framework proposed by Reich, Bondell and Li (2012), whereas our main mission here is to integrate BDR into the context of metabolomic analyses, and the specific tasks include (1) to permit effect modeling of multiple pathways, (2) to design appropriate priors for embedding biological knowledge, and (3) to allow selection of important pathways and metabolites. To fix the idea, we illustrate the model via the single-index scenario where one linear combination of metabolites x, $\lambda = \beta^{\mathrm{t}} x$ captures all relevant information about the phenotypes. The extension to multi-index case is straightforward. The BDR framework assumes that, for individual i, the conditional distribution of the phenotype y_i given the sufficient predictor $\lambda_t = \beta^{\mathrm{t}} x_i$ follows a finite mixture distribution $p(y_i | \lambda_i) = \sum_{k=1}^{K} p_k (\lambda_i) N(\mu_k, \sigma_y^2)$. In the model, the mixture weights $p_k(\lambda_i)$ satisfy that $\sum_{k=1}^{K} p_k (\lambda_i) = 1$ for all $\lambda_i \epsilon R^1$, and follow a probit model $p_k (\lambda_i) = \Phi\left(\frac{\phi_{k+1}-\lambda_i}{\sigma_z}\right) - \Phi\left(\frac{\phi_k-\lambda_i}{\sigma_z}\right)$ where Φ is the standard normal distribution function and $\phi_1 < \phi_2 < \cdots < \phi_{k+1}$ are cut points with $\phi_1 = -\infty$ and $\phi_{k+1} = \infty$. Besides, the Gaussian mixture can also be replaced by a discrete mixture distribution, e.g., binomial or Poisson. This setup results in a fully-conjugate model and facilitates rapid MCMC sampling and convergence. It is also shown to span a wide class of conditional densities and thus is really flexible.

Based on the above framework, we propose our framework of pathway selection that models the posterior distribution of the phenotype using data from *G* metabolic pathways with the conditional density, $p(y_i | \lambda_i) = \sum_{g=1}^{G} \sum_{k=1}^{K} p_k (\lambda_{gi}) N(\mu_{gk}, \sigma_y^2)$ where pathway g is characterized by a single linear combination $\lambda_{gi} = \beta_g^T x_{gi}$, and $\mu_{gk} \sim N\left(0, \sigma_{\mu gk}^2\right)$. Next we briefly list our designs of priors for various purposes: (a) selection of important pathways, (b) selection of important metabolite within a pathway, and (c) incorporation of metabolite and pathways relationship induced by network structure. For (a) selection of important pathways, we assume a prior normal distribution $N\left(0, \sigma_{\mu g}^2\right)$ for μ_{gk}'s, and adopt an stochastic search variable selection (SSVS; George and McCulloch, 1993) idea through the two component mixture prior, $\sigma_{\mu g}^2 = \delta_g + c(1-\delta_g)$, where $0 < c < 1$ is a small constant, $\delta_g \sim Bernoulli(\pi_0)$ is an indicator of whether pathway g is selected, and π_0 is a prior pathway inclusion probability. For (b) selection of important metabolites within a pathway, the similar SSVS prior can be applied, except that it is imposed on metabolites instead of pathways. That is, we introduce an indicator γ_{gm} that indicates whether the mth metabolite in pathway g is important. Finally for (c) incorporation of network structure, we specify priors on γ_{gm} that impose dependence induced by the network relationship among metabolites through Gaussian Markov random field (MRF, Tai et al., 2011). The advantage of such priors is that it imposes a less stringent assumption on individual metabolite contributions. That is, it enforces smoothness on the *significance* of the metabolites that are connected in the network, rather than enforcing these metabolites to have similar effect size.

Aim 2: Metabolomic Driven Genome-Scale Modeling of Human Metabolism

Aim 2a: Pathway Discovery and Metabolite Identification via Flux Balance Analysis

We propose to develop an open source, user-friendly, graphical user interface based, software suite that integrates metabolomic data with existing cutting-edge, command line based, genome-scale computational modelling algorithms. With computational biology expertise, Recon2 and other data (see Table 1), can be combined with untargeted metabolomic data to generate testable hypotheses about novel human metabolic pathways. However, this technology is currently not presented in an accessible form to individuals carrying out metabolomic analyses.

Table 1: Four algorithms will underlie the proposed software package. The input data for each algorithm is indicated. Except for metabolomic data, the software will automatically access the most recent datasets from a remote database hosted by the applicants.

Algorithm: **Input data:**	Homologous pathway discovery	Novel pathway discovery	Network contextual iso-baric ranking
Unidentified ion		√	√
Orphan metabolite (not in Recon2)	√		
MDL mol file		√	√
Recon2	√	√	√
Universal reaction DB	√		√
Chemistry rules		√	√

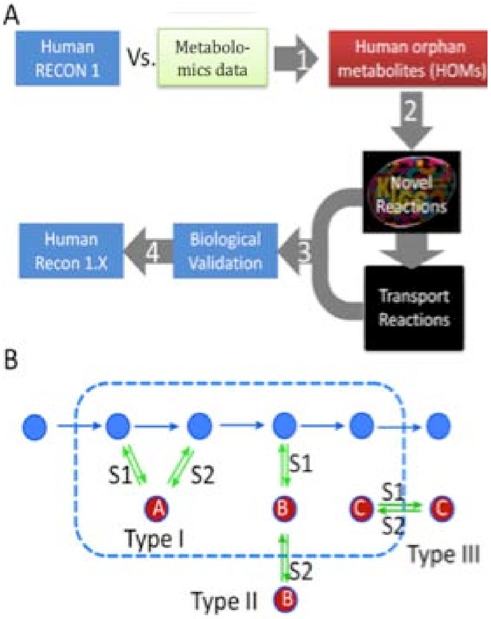

Figure 5: Identification of reactions capable of explaining the fate of orphan metabolites. **A**) Metabolites detected by metabolomic analysis and that are not contained in Recon 2 are orphan metabolites (1). The SMILEY algorithm proposed reactions from a non-organism specific metabolic reaction database or a transport reaction (2-3). Following validation of the biological relevance of proposed solutions through manual inspection (3), orphan metabolites can be incorporated into an updated metabolic reconstruction (4). **B**) A toy network showing the solution pathways predicted by SMILEY. In solution type I, metabolite *A* is incorporated into Recon 2 with two new pathways (S1 and S2, green) which couple the production and consumption of *A* to metabolites already accounted for in the network (blue). A type II solution involves transport into the cell and incorporation into the network as shown for *B*. A type III solution involves transport into and out of the cell as shown for *C*. Note that solution routes S1 and/or S2 can be composed of multiple reactions.

Pathway Discovery: Given a candidate human metabolite (identified or suggested by metabolomic analysis), which is not currently contained in Recon2, we will develop an accessible interface to two published algorithms (Reed et al., 2006; Hatzimanikatis et al., 2005) that generate hypotheses about novel mass balanced pathways that can connect candidate metabolites with the rest of the human metabolic network (Rolfsson et al., 2011). *Homologous Pathway Discovery:* The first algorithm, SMIELY (Reed et al., 2006), takes as input a list of metabolites, Recon2, a comprehensive reaction database (e.g., KEGG (Kanehisa et al, 2008)), and a transport database (e.g., Transport DB (Ren et al., 2004)). For each orphan metabolite we will create a mixed integer linear programming (MILP) problem to identify the minimal number of reactions that need to be added to Recon2 from the reaction database and/or the transport database, such that the orphan metabolite is connected to Recon2. Note that transport reactions will be only added from the extracellular space to the cytosol if the metabolite has been identified in the spent medium or in plasma. At its end, the algorithm will result in a list of candidate reactions to be added to Recon2, which will require detailed manual inspection. We have recently shown that this semi-automated method of metabolic network gap-filling not only helps to reconstruct known biochemical pathways, but is also capable of generating biologically plausible reaction hypotheses (Rolfsson et al., 2007). These reaction hypotheses can be subsequently experimentally validated by a biochemist and thus will lead to the expansion of our knowledge about human cellular metabolism. *Novel Pathway Discovery:* The second algorithm will be implemented similarly to a computational approach designated Biochemical Network Integrated Computational Explorer (BNICE , Hatzimanikatis et al., 2005) that computes every possible hypothetical biochemical reaction between a Recon2 metabolite and a candidate metabolite using a predefined set of template enzyme reaction rules (e.g., based on standard organic chemistry and a target compound). The candidate solutions will then again be manually inspected for known enzymes to catalyze such reactions in the relevant organism. **Network Contextual Isobaric Ranking of Unidentified Ions:** Another challenge in untargeted metabolomics is that many detected ions cannot be uniquely identified due to insufficient mass accuracy or identical empirical formulas, as it is the case for isomers. We will employ Recon2, the list of candidate identities for each predicted metabolite, and the aforementioned algorithms to propose reactions to be added to the network. By inspecting the nature and length of the solutions (i.e., the number of necessary reactions to be added to Recon2) we will assign a confidence score to each potential match for an unidentified ion.

Aim 2b: Development of an isotopomer mapping model of human metabolism

First, we shall generate of a draft isotopomer mapping model of Recon 2.0 using the existing algorithm (Ravikirthi, et al., 2011). Then, to the extent permitted by published biochemical literature, manual curation will be used to remove as many chemically infeasible atom mappings as possible. Next, we shall develop algorithms to mine metabolomic data from tracer studies to refine the atom mappings. In tracer studies using ^{13}C, ^{15}N or D-isotopic labeling of metabolites, isotopic tracer atoms propagates through metabolism into a wide range of structural subgroups within molecules of contiguous reactions. This data has the potential to reveal information on the atom transitions between metabolites in reactions along many pathways of metabolism. We shall develop algorithms to deconvolute raw LC-MS data from tracer studies to calculate the structural subgroup of a metabolite that contain isotope labeled atoms and thereby refine the draft atom mapping model. The deconvolution will use algorithms developed to identify metabolites using MS^n trees for the assignment of the position of tracer atoms within a molecule based on algorithms developed in NMC (Rojas-Chertó , et al., 2011). These data are being acquired after fractionation, or on-the-flight. The fragmentation reduces the number of possibilities for the position of an isotopic atom. NMC is establishing a generic data acquisition and data processing pipeline.

Aim 3: Network analysis, integration with other "omics" data and dynamic modeling

Metabolomic analysis of biological samples normally provides a single point perspective on the "metabolomic state" of the cell under diseased and drug-treatment conditions. While this is an important piece of data, it should be complimented by temporal sequence experiments that will provide a more kinetic and dynamic perspective on the metabolomics states. Isotope labeling measurements are essential to provide validations of the mechanisms and models. For comprehensive modeling, it is important to integrate the metabolomics data with transcriptomic and other phenotypic measurements on the same systems. At the first level, this data integration will be achieved through correlation analysis. While correlations do not imply causality, in combination with other phenotypic data, we can derive causal dynamic connectivity.

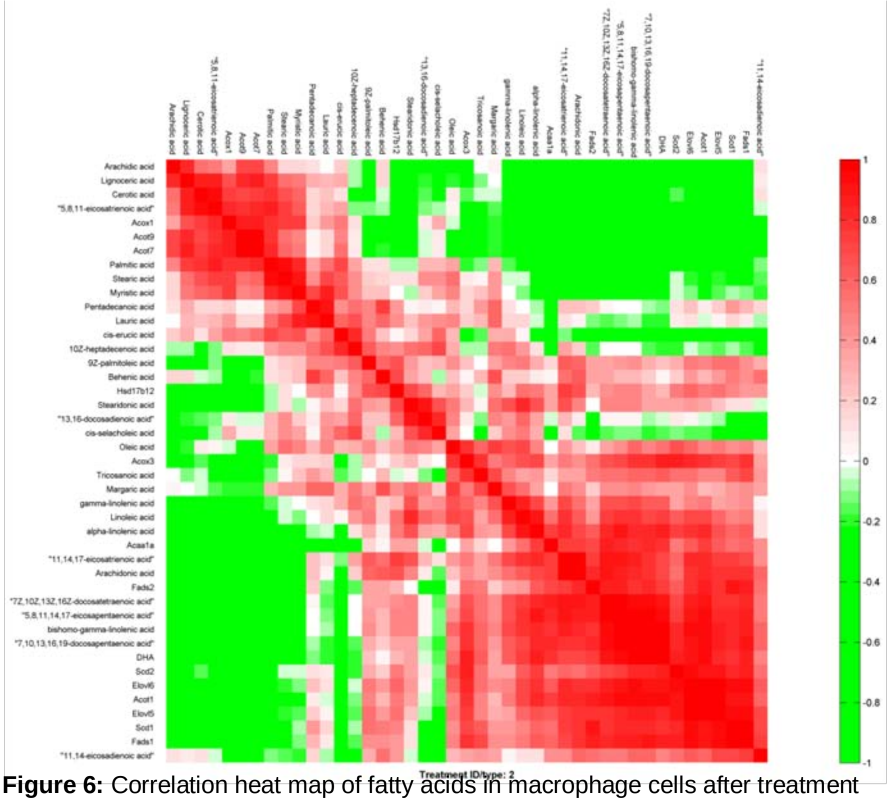

Figure 6: Correlation heat map of fatty acids in macrophage cells after treatment with LPS. For e.g. Acox1, Acot9 and Acot7 (genes) clustered with Arachdic acid, lignoceric acid, cerotic acid, palmitic acid, stearic acid, myristic acid (lipids) [top-left corner] show high correlation.

Correlation Analysis across "Omics" Data: To find out how similar two time courses (temporal responses) are, Pearson correlation between them can be computed. Correlation value can be thought of as the cosine of the angle between the normalized time-course curves (z-scores). To illustrate this, let us consider two time courses, X and Y (both are vectors):

$$X = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8]; \quad (1)$$
$$Y = [y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8]$$

We normalize the two vectors and the resultant is the z-score given by,

$$X_z = (X - mean(X)) / std(X) \quad (2)$$
$$Y_z = (Y - mean(Y)) / std(Y)$$

Then the correlation is simply the dot-product of the two z-scores.

$$\text{Correlation}: \quad r = X_z * Y_z^T \quad (3)$$

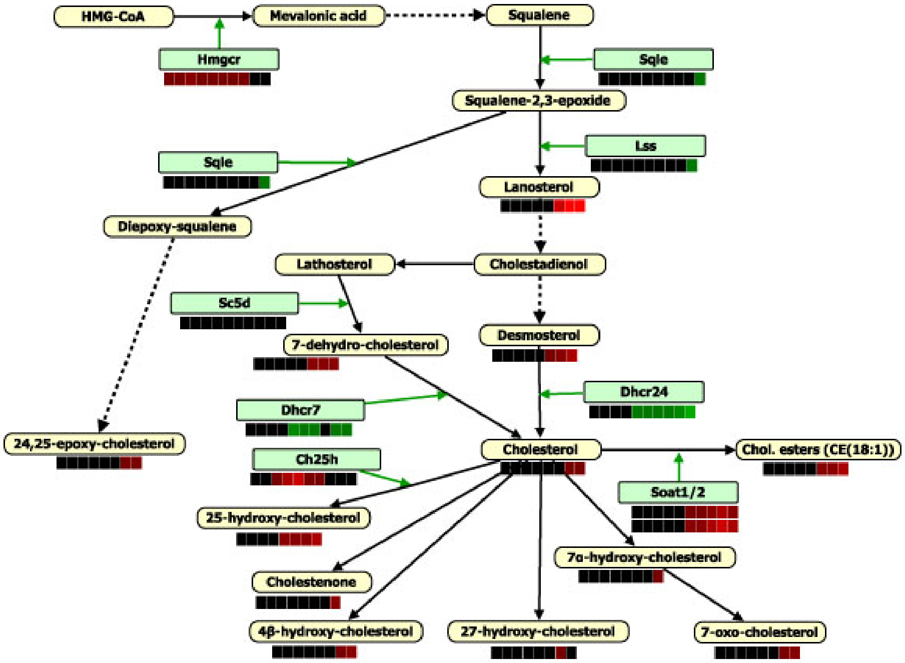

Figure 7: Reconstructed pathway map of Sterol metabolism in macrophages. Heat maps show time course measurements of genes and lipids. Red indicating fold increase and green fold decrease.

In the LIPID MAPS project, we have carried out extensive analysis of transcriptome-lipidome correlations (Subramaniam et al., 2011). In macrophage cell experiments, the time-course for gene data or lipid data consists of 8 time points (include the value at t = 0 hr), we display the data and the correlation using hierarchical clustering to layout the variables showing high correlations near each other. For example in the eicosanoid metabolic pathway changes associated with treating macrophage cells with lipopolysaccharide, we show the cross-correlations between genes and lipids in Figure 6. For example, genes Acox1, Acot9 and Acot7 (coding for enzymes associated with arachidic acid metabolism) are clustered with arachidic acid, lignoceric acid, cerotic acid, palmitic acid, stearic acid and myristic acid (top left corner of the correlation map). Such correlation analysis aids reconstruction of dynamic pathway maps in conjunction with legacy maps (e.g., from Recon1). We have developed metabolic pathway maps in macrophage cells treated with lipopolysacchride and with a statin drug, compactin. A pathway map of the sterol pathway that was reconstructed by integration of metabolic and transcriptomic data is presented in Figure 7.

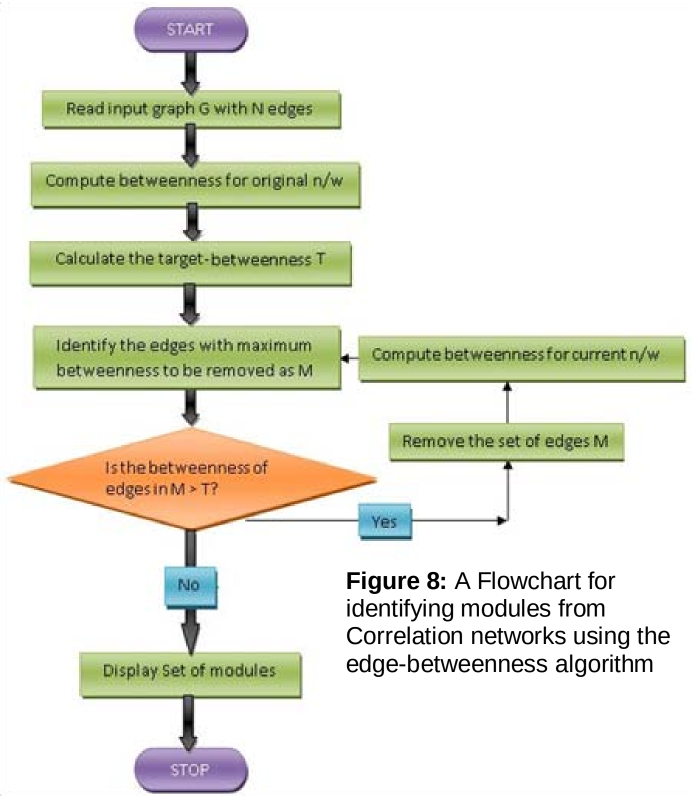
START
Read input graph G with N edges
Compute betweenness for original n/w
Calculate the target-betweenness T
Identify the edges with maximum betweenness to be removed as M
Compute betweenness for current n/w
Remove the set of edges M
Is the betweenness of edges in M > T?
Yes
No
Display Set of modules
STOP

Figure 8: A Flowchart for identifying modules from Correlation networks using the edge-betweenness algorithm

Identifying Pathway Modules from Correlation Data: We address the problem of clustering and modularity detection from correlation maps (graphs) by proposing a divisive algorithm, whose execution is directed by a deterministic termination criterion. The initial steps in our algorithm are motivated by the definition and use of edge-betweenness as a means of inferring community structure in networks by (Newman and Girvan, 2003). The edge-betweenness of an edge is computed as follows: The shortest paths between all pairs of vertices are found and the count of how many run along each edge, gives the measure of the betweenness value of that edge. Anthonisse, who is considered to have first introduced this, called it "rush", but Newman and Girvan termed it the edge-betweenness or the shortest-path betweenness. We will begin by considering the input correlation network of genes and metabolites as a graph consisting of nodes and edges, and calculating the edge-betweenness of all edges in the network. We will then calculate the target-betweenness T which will be used to guide the execution of the algorithm in subsequent steps. All the edges with maximum betweenness value are removed, and the betweenness is recalculated for edge after the removal. The key value-add in our proposal, is a novel and efficient criterion to stop the iterative removal of edges, in the context of biological networks. In particular, we propose that the recalculation and the removal of the edges be stopped at the stage when the edge to be removed is determined to be the edge with a betweenness value lower than the target-betweenness T.

Figure 8 illustrates the operation of our Target edge-betweenness Modularity Heuristic. The general form of their community structure finding algorithm is as follows:1. Calculate the betweenness for all edges in the network 2. Remove the edge with the highest betweenness 3. Recalculate the betweenness for all edges affected by the removal 4. Repeat from step 2 until no edge remains.

The algorithm's output is in the form of a dendrogram which represents an entire nested hierarchy of possible community divisions for the network. To identify where to cut the dendrogram to get a sensible division of the network, they define a measure of the quality of a particular division of a network, called *modularity (Q)****,*** which measures the fraction of the edges that connect vertices within the same cluster minus the expected value of the same quantity in the network. The division of the network with maximum value for Q is considered to be the best-split.

We have implemented this algorithm for protein-protein interaction networks, but in this application, we intend to utilize this to obtain modules from correlation maps. These modules will provide two types of insights. In the first, we will obtain co-regulated genes/metabolites that indicate diseased state or drug response mechanisms. In the second, we will obtain insights into downstream functional modules that emerge from the interaction of transcriptional and metabolic regulation.

Dynamical Model Development: We will have time series measurements of metabolites in cells post treatment with drugs. Using this data, a linear model will be developed using partial least square (PLS) method. Dynamic mapping ($y_t = f(y_{t-1})$) will be used to calculate the interaction coefficients. The model ($y_t = f(y_{t-1})$) represents the dependence of the level of measured PPs at time t on the level of all measured PPs at time t-1. If the mapping function (f) is linear, then it can be derived from the state-space representation of the system as follows: $\frac{d}{dt}X = \mathbf{A}X$, After discretization and rearrangement, we get $\left.\frac{d}{dt}X\right|_{t=t_k} \approx \frac{X_{t=t_k} - X\big|_{t=t_{k-1}}}{(t_k - t_{k-1})} = \mathbf{A}X_{t=t_{k-1}}$, or equivalently, $X_{t=t_k} = \mathbf{B}X_{t=t_{k-1}}$, where X (the vector of state variables) denotes the set of PPs. A and B are coefficient matrices with suitable relationships between the elements of A and B. From here on, we will work with Eq. 2. Even though PLS has the capability to handle multiple outputs, we will develop the models for individual metabolites (single output). This is motivated by a criteria for model selection, namely, the input matrix should capture significant variance in the output. Only those models will be selected for network reconstruction for which the input data will capture at least 50% variance of the output data.

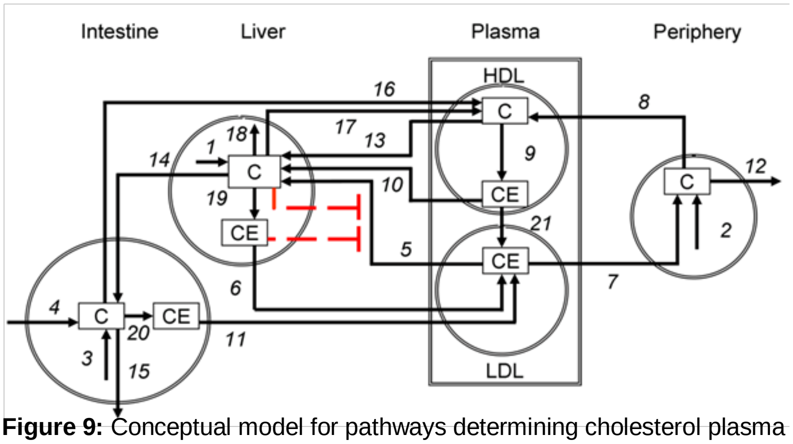

Figure 9: Conceptual model for pathways determining cholesterol plasma levels used as a basis to set up the in silico model. Process numbers stand for: 1, Hepatic cholesterol synthesis; 2, Peripheral cholesterol synthesis; 3, Intestinal cholesterol synthesis; 4, Dietary cholesterol intake; 5, Hepatic uptake of cholesterol from non HDL; 6, Hepatic Very Low Density lipoprotein cholesterol (VLDL-C) secretion; 7, Peripheral uptake of cholesterol from non HDL; 8, Peripheral cholesterol transport to HDL; 9, HDL associated cholesterol esterification; 10, Hepatic HDL-CE uptake; 11, Intestinal chylomicron cholesterol secretion; 12, Peripheral cholesterol loss; 13, Hepatic HDL-FC uptake; 14, Biliary cholesterol excretion; 15, Fecal cholesterol excretion; 16, Intestinal cholesterol transport to HDL; 17, Hepatic cholesterol transport to HDL; 18, Hepatic cholesterol catabolism; 19, Hepatic cholesterol esterification; 20, Intestinal cholesterol esterification and 21, CE transfer from HDL to LDL. C stands for cholesterol; CE for cholesterol ester. Red dashed lines indicate the regulation of the LDLR in response to the hepatic cholesterol level. (van de Pas et al, 2010, 2011)

Kinetic Model Development: We will use similar strategies to construct integrated dynamic and kinetic models of metabolomics networks. The time series data will provide us with concentrations and the rates and we can compute the rate constants. The kinetic models can be directly validated by isotopomeric experiments. Finally, metabolomics based kinetic models will be implemented on an “omics” integration, at physiological and clinically relevant level. This will be done using the metabolomics and genomics data of the above mentioned statin trials, where data on cholesterol biosynthesis, absorption, transport, esterification and excretion will be combined with metabolomics data on statin bioavailability, genomics data on genetic variations in statin transporter and other relevant genetic variations, together with (intermediate) pathological endpoint data. The kinetic model has been developed previously for mice and human applications and all relevant processes in cholesterol homeostasis are presented above (see Figure 9). The outcome of this model is clinically applicable in fine tuning statin based therapy and designing personalized combination therapies (“statins + second drug”).

6. Concluding Remark

We hope that the complimentary of approaches we bring will enable the metabolomics community derive biological insights from complex and rich datasets derived from ever evolving and complimentary analytical platforms. We also hope that this will enable and empower young scientists who are joining the metabolomics field to learn rapidly about issues related to mining of metabolomics data. We focus mainly on addressing complexities in mining and modeling metabolomics data but start to explore in specific aim 3 how to integrate metabolomics data with other layers of data. We plan to work closely with metabolomics researchers funded under common funded to enable sharing of all knowledge gained from this project.

MILESTONES AND TIMELINES

	Milestones	Y1	Y2	Y3	Y4	Y5
Aim 1a	Develop strategy to use QC, replicate and reference samples to monitor and correct for unwanted experimental variation	🟥				
	Develop strategy to use error models per metabolite per platform for subsequent data analysis of the integrated data sets		🟥	🟥		
	Develop strategy to assign iss to each metabolite for quantification purposes			🟥	🟥	
	How to use reference samples to build transfer models between different analytical platforms				🟥	🟥
Aim 1b	Develop the systematic metabolomics data analysis strategy	🟥				
	Implement the systematic strategy into software		🟥	🟥		
Aim 1c	Develop the statistical framework for the identification of metabolic pathways implicated in the mechanism of phenotype				🟥	🟥
Aim 2	Algorithmically reconstructed, literature curated, human atom transition mapping	🟦	🟦			
	Software interface for novel pathway discovery algorithm from metabolomic data			🟦	🟦	
Aim 3	Correlation analysis across “omics” data	🟩	🟩			
	Identify pathway modules from correlation data			🟩		
	Dynamical model development				🟩	🟩