Predicting Network Activity from High Throughput Metabolomics

Shuzhao Li, Emory University
Youngja Park, Emory University
Sai Duraisingham, Emory University
Fred Strobel, Emory University
Nooruddin Khan, Emory University
Quinlyn A. Soltow, Emory University
Dean P Jones, Emory University
Bali Pulendran, Emory University

Journal Title: PLoS Computational Biology
Volume: Volume 9, Number 7
Publisher: Public Library of Science | 2013-07-01, Pages e1003123-e1003123
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1371/journal.pcbi.1003123
Permanent URL: https://pid.emory.edu/ark:/25593/s76xm

Final published version: http://dx.doi.org/10.1371/journal.pcbi.1003123

Copyright information:
© 2013 Li et al
This is an Open Access work distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Accessed January 31, 2019 9:18 PM EST
Predicting Network Activity from High Throughput Metabolomics

Shuzhao Li1,2*, Youngja Park3,4, Sai Duraisingham1,2, Frederick H. Strobel5, Nooruddin Khan1,2, Quinlyn A. Soltow3, Dean P. Jones3, Bali Pulendran1,2

1 Emory Vaccine Center, Emory University, Atlanta, Georgia, United States of America, 2 Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, United States of America, 3 Department of Medicine, Emory University, Atlanta, Georgia, United States of America, 4 College of Pharmacy, Korea University, Seoul, South Korea, 5 Mass Spectrometry Center, Emory University, Atlanta, Georgia, United States of America

Abstract

The functional interpretation of high throughput metabolomics by mass spectrometry is hindered by the identification of metabolites, a tedious and challenging task. We present a set of computational algorithms which, by leveraging the collective power of metabolic pathways and networks, predict functional activity directly from spectral feature tables without a priori identification of metabolites. The algorithms were experimentally validated on the activation of innate immune cells.

Introduction

Knowledge of many metabolic pathways has accumulated over the past century. For instance, glycolysis, citric acid cycle and oxidative phosphorylation fuel cellular processes through the generation of adenosine triphosphate; glycans and cholesterol are not only serve as structural blocks but also mediate intercellular communication. In fact, metabolites pervade every aspect of life [1,2]. Their roles are increasingly appreciated, as advancing tools allow deeper scientific investigations. The most notable progresses in recent years come from metabolomics and genome-scale metabolic models.

Metabolomics is the emerging field of comprehensive profiling of metabolites. As metabolites are the direct readout of functional activity, metabolomics fills a critical gap in the realm of systems biology, complementing genomics and proteomics [3–6]. The technical platforms of metabolomics are mainly based on mass spectrometry and nuclear magnetic resonance [4,7]. Among them, untargeted LC/MS (liquid chromatography coupled mass spectrometry), especially on high resolution spectrometers, produces unparalleled throughput, measuring thousands of metabolite features simultaneously [5,8–10].

On the other hand, genome-scale metabolic models have been largely driven by genomics, as the total list of metabolic enzymes of a species can be derived from its genome sequence [11,12]. The reconstruction of microbial metabolic network models is an established process [13,14]. Intense manual curation, however, was required in the building of two high-quality human models [15,16], which were followed by a number of derivatives [17–20]. The coverage of these metabolic models greatly exceeds the conventional pathways.

Even though they are a perfect match in theory, metabolomics and genome-scale metabolic models have had little overlap so far. The use of metabolomics in building metabolic models is rare [21], due to the scarcity of well annotated metabolomics data. The application of genome-scale metabolic models to metabolomics data is not common either [22]. The limited throughput of targeted metabolomics usually does not motivate large scale network analysis. Untargeted metabolomics cannot move onto pathway and network analysis without knowing the identity of metabolites.

A typical workflow of untargeted metabolomics is illustrated in Figure 1A. After ionized molecules are scanned in the spectrometer, the spectral peaks are extracted, quantified and aligned into a feature table. At this point, each feature is identified by a mass-to-charge ratio (m/z) and retention time in chromatography, but its chemical identity is not known. To assign a spectral feature to a bona fide metabolite, it usually involves tandem mass spectrometry to examine the fragmentation pattern of a specific feature, or coelution of isotopically labeled known references - both are inherently low throughput. Considerable effort is needed to build a spectral library, which is often of limited size and interoperability. Thus, metabolite identification forms the bottleneck of untargeted metabolomics [23].

A number of informatics tools have been developed for LC/MS metabolomics, ranging from feature extraction [24–26], pathway analysis and visualization [27–29] to workflow automation [30–32]. Yet, whereas pathway and network analysis is concerned, the existing tools require identified metabolites to start with. Computational prediction of metabolite identity, based on m/z alone, is deemed inadequate as a single m/z feature can match multiple metabolites even with high instrumental accuracy [33,34], and...
mass spectrometry based untargeted metabolomics can now profile several thousand of metabolites simultaneously. However, these metabolites have to be identified before any biological meaning can be drawn from the data. Metabolite identification is a challenging and low throughput process, therefore becomes the bottleneck of the filed. We report here a novel approach to predict biological activity directly from mass spectrometry data without a priori identification of metabolites. By unifying network analysis and metabolite prediction under the same computational framework, the organization of metabolic networks and pathways helps resolve the ambiguity in metabolite prediction to a large extent. We validated our algorithms on a set of activation experiment of innate immune cells. The predicted activities were confirmed by both gene expression and metabolite identification. This method shall greatly accelerate the application of high throughput metabolomics, as the tedious task of identifying hundreds of metabolites upfront can be shifted to a handful of validation experiments after our computational prediction.

In this paper, we report a novel approach of predicting network activity from untargeted metabolomics without upfront identification of metabolites, thus greatly accelerating the work flow. This is possible because the collective power in metabolic networks helps resolve the ambiguity in metabolite prediction. We will describe the computational algorithms, and demonstrate their application to the activation of innate immune cells.

Results

Full MS scan in an untargeted metabolomics experiment gives the most power of high throughput profiling, producing several thousand of features routinely. Our goal is to predict biological activity in a network of metabolites directly from the feature table, bypassing the bottleneck of metabolite identification (Figure 1A). As illustrated in Figure 1B, each spectral feature can match to several metabolite candidates, albeit the true identity is not known. The conventional work flow is to identify each metabolite experimentally before mapping them onto metabolic pathways/networks. Instead of treating metabolite identification and metabolic pathway/network analysis as two separate steps, we argue that there is collective power in the organization of metabolic networks, which can be leveraged when the two steps are unified under one theoretical framework. In other words, if a list of significant spectral features reflects a biological activity, the true metabolites they represent should show enrichment on a local structure in the metabolic network [22], while the falsely matched metabolites are distributed more randomly. This gives us a means to predict functional activities without upfront identification of metabolites (Figure 1C).

The software implementation of our approach is named mummichog (mummichog is an American Indian term for by groups, Figure 1). Mass spectrometry based untargeted metabolomics can now profile several thousand of metabolites simultaneously. However, these metabolites have to be identified before any biological meaning can be drawn from the data. Metabolite identification is a challenging and low throughput process, therefore becomes the bottleneck of the filed. We report here a novel approach to predict biological activity directly from mass spectrometry data without a priori identification of metabolites. By unifying network analysis and metabolite prediction under the same computational framework, the organization of metabolic networks and pathways helps resolve the ambiguity in metabolite prediction to a large extent. We validated our algorithms on a set of activation experiment of innate immune cells. The predicted activities were confirmed by both gene expression and metabolite identification. This method shall greatly accelerate the application of high throughput metabolomics, as the tedious task of identifying hundreds of metabolites upfront can be shifted to a handful of validation experiments after our computational prediction.

Figure 1. Mummichog redefines the work flow of untargeted metabolomics. A) In the work flow of untargeted metabolomics, the conventional approach requires the metabolites to be identified before pathway/network analysis, while mummichog (blue arrow) predicts functional activity bypassing metabolite identification. B) Each row of dots represent possible matches of metabolites from one m/z feature, red the true metabolite, gray the false matches. The conventional approach first requires the identification of metabolites before mapping them to the metabolic network. C) mummichog maps all possible metabolite matches to the network and looks for local enrichment, which reflects the true activity because the false matches will distribute randomly.

doi:10.1371/journal.pcbi.1003123.g001
also name of a small fish which live by groups). From user input, 
mummichog requires two lists of \textit{m/z} features, the significant list \textit{L}_{\text{sig}}
(e.g. selected by univariate statistics) and the reference list \textit{L}_{\text{ref}} (all
features detected in the MS experiment). From the \textit{m/z} features in
\textit{L}_{\text{sig}}, mummichog computes all possibly matched metabolites, and
searches the reference metabolic network for all the modules that
can be formed by these tentative metabolites. Random lists of \textit{m/z}
features are drawn from \textit{L}_{\text{ref}} many times to estimate the null
distribution of module activities. The statistical significance of
modules from \textit{L}_{\text{ref}} can then be calculated on this null distribution.
In return, the predicted metabolites in these significant modules
and both are included in 
mummichog.

The reference metabolic network model

The genome-scale human metabolic network in mummichog is
based on KEGG [39], UCSD Recon1 [13] and Edinburgh human
metabolic network [16]. The integration process was described
previously [39]. The organization of metabolic networks has been
described as hierarchical and modular [40]. When we perform a
hierarchical clustering on the metabolic reactions in our network,
its modular structure is clear (Figure 2A). This modular
organization, as reported previously [41], often but not always
 correlates with conventional pathways (Figure 2B).

The module definition in this work is adopted from Newman and
Girvan [42,43], where a module is a subnetwork that shows
more internal connections than expected randomly in the whole
network. Modules are less biased than pathways, which are
defined by human knowledge and conventions, and outgrown by
genome-scale metabolic networks. Activity of modules may exist
within and in between pathways. Deo et al [22] convincingly
demonstrated the advantage of unbiased module analysis over
pathways. On the other hand, pathways have built-in human
knowledge, which may be more sensitive under certain scenarios.
Pathway analysis and module analysis are rather complementary,
and both are included in mummichog.

The reference metabolic network model contains both metab-
olites and enzymes. Since metabolomics only measures metabo-
lites, the model is converted to a metabolite centric network for
statistical analysis. Enzymes are only added later in the visualiza-
tion step to aid user interpretation.

Module analysis

Within the predefined reference metabolic network model \textit{M},
mummichog searches for all modules that can be built on user input
data, and compute their activity scores. This process is repeated
many times for the permutation data to estimate the background
null distribution. Finally, the statistical significance of modules
based on user data is calculated on the null distribution. The
specific steps are as follows:

1. A list of “input metabolites” is tentatively computed from the
   significant list of \textit{m/z} features, \textit{L}_{\text{sig}}. The \textit{m/z} - metabolite
   matching considers all common isotopic derivatives and
   adducts.

2. From \textit{M}, all subnetworks that can connect the input
   metabolites from Step 1 within one reaction are extracted.
   This is repeated for subnetworks within two, three and four
   reactions (by four steps \textit{M} should be exhausted, see Figure 2C).
   These subnetworks contain both “input metabolites” and other
   metabolites from \textit{M}.

3. Structural modules are identified within these subnetworks
   according to Newman’s spectral split algorithm [43]. This
   creates many modules of different sizes.

4. Both subnetworks from Step 2 and modules from Step 3 are
   compiled into a set of candidate modules (graphs), \textit{G}_1, \textit{G}_2, \ldots, \textit{G}_n.
   Each graph is cleaned of protruding edges if such edges do not connect input metabolites therefore not
   contributing to the activity score.

5. The activity score should reflect both the enrichment of “input
   metabolites” and the modularity. Activity score of a candidate
   module is computed as follows: For a subgraph \textit{G}, activity score
   \begin{equation}
   \hat{A} = Q \frac{N_{i,G}}{N_G},
   \end{equation}
   where \textit{N}_{G} is the number of metabolites in \textit{G}, \textit{N}_{i,G} the number of
   “input metabolites” in \textit{G}, \textit{Q} adjusted Newman-Girvan
   modularity:
   \begin{equation}
   Q = \sqrt{\frac{N_{i,G}}{N_G} \left( \frac{E_G}{m} - \sum_{ij} k_i k_j \frac{2m}{2m} \right)}, \quad i,j \in \textit{G}
   \end{equation}
   where \textit{k}_{i} is the network degree of metabolite \textit{i}, \textit{m} the total
   number of edges in the metabolic network \textit{M}, \textit{E}_{G} the total
   number of edges in \textit{G}, \textit{N}_{i} the number of “input metabolites”.
The original Newman-Girvan modularity has a bias towards
larger modules. The \sqrt{\frac{N_{i,G}}{N_G}} is added to reduce this bias.

6. A list of permutation features (equal length to \textit{L}_{\text{ref}}) is generated
   by randomly sampling \textit{L}_{\text{ref}}. The activity scores of modules
   from this permutation list are computed as in Steps 1–5.

7. Repeat Step 6 many times to populate a list of scores from
   random modules. Using maximum likelihood estimation, these
   scores are modeled as a Gamma distribution (this is the null
   distribution), and a cumulative distribution function (CDF) is
   calculated.

8. The p-value for each module from input metabolites is
   calculated on the CDF of null distribution.

Pathway analysis

The basic test for pathway enrichment here is Fisher’s exact test
(FET), which is widely used in transcriptomic analysis. The
concept of FET is, when we select \textit{k} features (\textit{L}_{\text{ref}}) from a total of
\textit{N} features (\textit{L}_{\text{ref}}), and find \textit{m} of the \textit{k} features present on a
pathway of size \textit{P}, the chance of getting \textit{m} in theory can be
estimated by enumerating the combinations of \textit{k}, \textit{N} and \textit{P}.

To apply FET to an enrichment test of metabolic features on
pathways, we need to understand the additional layer of
complexity. Our metabolic features can be enumerated either in
the \textit{m/z} feature space or in the metabolite (true compound) space.
Since metabolic pathways are defined in the metabolite space,
either way needs to factor in the many-to-many mapping between
\textit{m/z} features and metabolites (Figure S1). This mapping is
effectively covered in our permutation procedure, which starts
from the \textit{m/z} features and reruns the mapping every time.
The overall significance of a pathway enrichment is estimated based on
a method by Berriz et al [44], which ranks the p-value from real
data among the p-values from permutation data to adjust for type
I error. Finally, a more conservative version of FET, EASE, is
adopted to increase the robustness [45]. The key idea of EASE is
to take out one hit from each pathway, thus preferentially penalize pathways with fewer hits. The specific steps are as follows:

1. A list of permutation features (equal length to $L_{\text{sig}}$) is generated by randomly sampling $L_{\text{ref}}$. This list is mapped to a list of tentative metabolites. These metabolites are looked up in each metabolic pathway. For each pathway, a FET right-tail p-value is calculated.

2. Repeat Step 1 many times to populate a list of p-values of all pathways under all permutations (null distribution). Slightly different from Berriz et al [44], all p-values not just the minimal p-values in each permutation are used. This is because the size and organization of metabolic pathways vary greatly, a different situation from the large gene categories in Berriz et al. Using maximum likelihood estimation, these p-values are modeled as a Gamma distribution (this is the null distribution), and a cumulative distribution function (CDF) is calculated.

3. Perform Step 1 on the significant feature list $L_{\text{sig}}$, and calculate FET p-value and EASE score for each pathway. An adjusted p-value per pathway is calculated based on the EASE score and CDF from Step 2. Both FET p-value and adjusted p-value are reported for each pathway.

**Data driven activity network**

Both the module analysis and pathway analysis above serve as a framework to estimate the significance of functional activities. In return, the predicted metabolites in significant
activities are more likely to be real. *Mummichog* collects these metabolites, and look up all their isotopic derivatives and adducts in *Lref*. A confidence rating system is applied to filter for qualified metabolites. For instance, if both the single-charged $^{12}$C form $M+\text{H}[1+]$ and the $^{13}$C form $\text{M(C13)}+\text{H}[1+]$ are present, this metabolite prediction carries a high confidence. All the qualified metabolites carry over their connections in the reference metabolic network, and form the "activity network" for this specific experiment (e.g. Figure 3).

The activity network gears towards a quality and clear view of user data, as modules and pathways can be redundant and fragmented. It also accentuates the activity in a specific experimental context, in contrast to the generic nature of the reference metabolic network.

We next illustrate the application of these algorithms to a novel set of immune cell activation data, and two published data sets on human urinary samples and yeast mutants.

**Application to the activation of innate immune cells**

The innate immunity plays a critical role in regulating the adaptive immunity, and the field was recognized by the 2011 Nobel Prize in Physiology or Medicine [46]. According to the nature of stimuli, innate immune cells direct different downstream molecular programs, which are still under intense scientific investigation [47,48]. In this study, we examine the metabolome of human monocyte-derived dendritic cells (moDC) under the stimulation of yellow fever virus (YF17D, a vaccine strain). We have shown previously that yellow fever virus activates multiple toll-like receptors, and induces cellular stress responses [49–51]. This data set is, to our knowledge, the first high throughput metabolomics on any immune cells (macrophages were previously studied by limited throughput).

The cell extracts from our activation experiment were analyzed by LC/MS metabolomics, and yielded 7,995 spectral features (denoted as *Lref*) after quality control. Among them, 601 features were significantly different between the infected samples and both the baseline and time-matched mock controls ($p<0.05$, student t-test; denoted as *Lsig*). Using *Lsig* and *Lref*, *mummichog* computes significant pathways and modules and the activity network.

Viral infection induced a massive shift of metabolic programs in moDCs (pathways in Table S1, modules in Figure S2). The predicted activity network is shown in Figure 3A, and we will focus our investigation on a small subnetwork (Figure 3B), which includes the metabolisms of nucleotides, glutathione/glutathione disulfide and arginine/citrulline. Nucleotides are required for viral replication, and the hijacking of host nucleotide metabolism by virus has been well described [52–54]. Glutathione is best known as intracellular antioxidant, where it is oxidized to glutathione disulfide (GSSG). However, our data show that both glutathione and GSSG are depleted in activated moDCs, departing from this conventional wisdom. The across-the-board depletion is consistent with the down-regulation of genes for glutathione synthesis (Figure 4B). Our data support the notion that glutathione is released by dendritic cells and conditions the extracellular microenvironment during their interaction with T cells [55–57]. Arginine is known to be an important regulator in innate immune response [58,59]. Arginine metabolism can lead to two pathways: to ornithine (catalyzed by arginase) or to citrulline (catalyzed by nitric oxide synthase). The decrease of arginine and increase of citrulline suggests the latter pathway, which is the main reaction of producing intracellular nitric oxide. We indeed detected the inhibition of eNOS and iNOS expression later (Figure 4C), a well documented feedback effect by nitric oxide [60]. We also performed tandem mass spectrometry on

![Figure 3. Metabolic activity network in dendritic cells stimulated by yellow fever virus. A) Prediction by *mummichog* directly from m/z feature tables (cell extracts after 6 hours of infection). Metabolites are colored according to log2 fold change. A high resolution copy is given in Figure S9. B) Further investigation was focused on the subnetwork on the right. Glutamate was not significantly altered, but included for network connectivity.](https://doi.org/10.1371/journal.pcbi.1003123.g003)
the metabolites in Figure 3B, using authentic chemicals as references. All the metabolites, except glutamylcysteine and thyroxine, were confirmed (Figure 5, Figure S3). The depletion of arginine and accumulation of citrulline in moDC was also triggered by dengue virus but not by lipopolysaccharide (LPS, Figure S4). It is known that iNOS is induced in dendritic cells by LPS but not by virus [47,61]. Our data suggest a different nitric oxide pathway in viral infection, driven by constitutive nitric oxide synthases. The intracellular nitric oxide has a fast turnover and we did not detect its accumulation by fluoremetric assays (data not shown). We previously demonstrated that the phosphorylation of EIF2A was induced by YF17D [50]. An upstream mechanism is now suggested by this metabolomic experiment, as both the production of nitric oxide and depletion of arginine induce the activity of EIF2A kinases [62].

Evaluation on human urine data and yeast data

The nature of metabolomics data often varies by platforms and sample types. We thus extend our mummichog approach to two published data sets on human urinary samples [63] and on yeast cell extracts [64]. Both data sets carry metabolite annotation by the original authors, which can be used to evaluate the prediction by mummichog.

The human urinary data contained both formal identification by matching to local library of chemical references and putative identification by combining multiple public resources [63]. We used mummichog to investigate the gender difference in this data set, and predicted an activity network of 45 metabolites. Among them, 13 were not found in the original annotation. For the remaining metabolites, 97% (31/32) were agreed between mummichog and the original annotation (Figure 6). There is an option in mummichog to enforce the presence of $M+\text{H}[+]$ form (for positive mode, $M−\text{H}[−]$ for negative mode) in metabolite prediction. With this option, 3 out of 44 metabolites were not in the original annotation, and the remaining 41 metabolites were in 100% agreement.

The mummichog algorithms are not tied to a specific metabolic model. We adopted the yeast metabolic model from BioCyc database [11] for the yeast data [64], to predict an activity network contrasting mutant and wild type strains. This data set was only annotated for 101 metabolites through the authors’ local library. As a result, the majority of metabolites in the predicted network by mummichog were not found in the original annotation. Out of the remaining 28 metabolites, 24 (86%) were agreed between mummichog and the original annotation (Figure 6). Enforcing the presence of primary ion $M−\text{H}[−]$ (data collected in negative ion mode) had little effect to the result, since the original annotation was already biased to metabolites that are ionized easily.

These results show that the prediction by mummichog is robust cross platforms and sample types.

Discussion

We have described the computational method of mummichog and its application to three case studies. While the description was intended to be concise, computational metabolomics is complicated and facing challenges from many fronts. Quantification and feature selection, the steps upstream of mummichog, can still be problematic. Data from in vivo studies often involve multivariate analysis, where validation is particularly challenging. In this
section, we discuss several confounding factors that are relevant to our approach.

Impact of metabolic models

Critical to the success of mummichog is the integration of genome-scale metabolic models. We have used in this study a recent human metabolic model. An alternative human model from BioCyc [11] produced comparable results (Figure S6). The coverage of the models in all three case studies is shown in Table 1. These genome-scale metabolic models are more extensive than conventional pathways, and were shown to capture activities in between pathways [22]. The pathway organizations differ between the two human models, as the BioCyc model tends to use smaller pathways. This creates some model dependency in the pathway analysis, but little effect to the “activity network”, as mummichog is more network centric. The two test cases in Figure 6 also indicate that these models tend to capture more information than conventional annotations.

However, as mentioned earlier, the new data from metabolomic studies are yet to be integrated into these genome-scale metabolic models. For example, a number of metabolites in metabolomics databases [36,65,66] are not in any of these metabolic models. In general, the features from a high resolution profiling experiment by far exceed the current annotations in metabolite databases. This leads to a de facto filtering when data are run on mummichog (similar situation in database searches). Meanwhile, the features

Figure 5. Identification of metabolites by tandem mass spectrometry. Arginine is shown as an example while the full data are given in Figure S3. From top to bottom: the fragmentation pattern (MS² followed by MS³ on peak 157) from biological sample, from biological sample spiked with authentic chemical and from authentic chemical reference.
doi:10.1371/journal.pcbi.1003123.g005
that can be mapped to the current metabolic model are more likely to be biologically relevant. This "filtering" is pertinent to the metabolic model, not to *mummichog* algorithms - *mummichog* still has to choose the true metabolites from multiple possible candidates (Figure S1B).

It will be an important future direction to advance metabolic modeling with the chemical data. We also expect the metabolic models to improve on lipid annotation, physiological context and tissue specificity.

### Robust design of the algorithms

As lessons learned from transcriptomics, pathway and network analysis not only provides functional context, but also the robustness to counteract noises at individual feature level, which are commonly seen in omics experiments. Similarly, the prediction on activity by *mummichog* is tolerant to errors at individual feature level. In the moDC data, we chose $L_{sig}$ by a cutoff value $p = 0.05$. When we vary this cutoff from 0.001 to 0.2, the program returned networks of a stable set of metabolites (Figure S7).

The module finding procedure in the program was designed to extensively sample subnetwork structures. Among the modules will be many variations, but the subsequent "activity network" will collapse on stable results. In deed, we tested an alternative algorithm of modularization [67], and it returned almost identical predicted networks, in spite of moderately different intermediate modules (Figure S8).

In theory, there are merits to incorporate a statistical matrix from the feature selection step prior to *mummichog*’s analysis and mass flow balance of metabolic reactions [22,68]. While these are appealing directions for future research, the current version of *mummichog* confers some practical robustness, such as tolerance to technological noise and biological sampling limitation. For example, mass balance is almost impossible within serum or urine samples, because the reactions producing these metabolites are likely to occur in other tissues.

Figure 6. Application of *mummichog* to additional data sets. Metabolite prediction by *mummichog* is in good agreement with annotation in the original studies, 97% for the human urine data [63] and 86% for the yeast data [64]. The metabolites not in the original annotation (yellow) can not be compared. The "-z" option in *mummichog* enforces the presence of primary ion (M+H+) for positive mode, M–H[−] for negative mode. This shifts the coverage in the huamn data set, but not much for the yeast data of limited annotation.

![Diagram](https://example.com/diagram.png)

Table 1. How metabolomics data match metabolic models.

<table>
<thead>
<tr>
<th></th>
<th>human model</th>
<th>human model (BioCyc)</th>
<th>yeast model (BioCyc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of reactions</td>
<td>4204</td>
<td>2091</td>
<td>1034</td>
</tr>
<tr>
<td>metabolite edges</td>
<td>18504</td>
<td>14100</td>
<td>6962</td>
</tr>
<tr>
<td>number of metabolites</td>
<td>3565</td>
<td>1536</td>
<td>811</td>
</tr>
<tr>
<td>number of metabolites with exact mass</td>
<td>2016</td>
<td>1308</td>
<td>742</td>
</tr>
<tr>
<td>case moDC: reference features</td>
<td>7713</td>
<td>7713</td>
<td></td>
</tr>
<tr>
<td>case moDC: tentatively matched metabolites</td>
<td>1384</td>
<td>874</td>
<td></td>
</tr>
<tr>
<td>case urine: reference features</td>
<td>10248</td>
<td>10248</td>
<td></td>
</tr>
<tr>
<td>case urine: tentatively matched metabolites</td>
<td>1453</td>
<td>930</td>
<td></td>
</tr>
<tr>
<td>case yeast: reference features</td>
<td>5366</td>
<td></td>
<td>533</td>
</tr>
<tr>
<td>case yeast: tentatively matched metabolites</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The top section gives statistics of each of the three metabolic models. The bottom section gives the specific number of measured features and tentative metabolite matches in each study case. These tentative matches contain excessive ambiguity, which *mummichog* aims to resolve.

![Diagram](https://example.com/diagram.png)
Acceleration of untargeted metabolomics

In spite of the fantastic progress in mass spectrometry, these are the early days of metabolomics. Effective computational integration of resources, the combination of cheminformatics and bioinformatics, will greatly benefit the field [69,70]. As data accumulate, further method refinement will become possible.

Mummichog presents a practical solution of one-step functional analysis, bypassing the bottleneck of upfront metabolite identification. It trades off some sensitivity in the conventional approach for the much accelerated work flow of high throughput LC/MS metabolomics. Mummichog is not designed to replace tandem mass spectrometry in metabolite identification. It is the biological activity not metabolites per se that mummichog predicts. Even with some errors on individual metabolites, as long as the biology is pinpointed to a subnetwork structure, investigators can focus on a handful of validations, steering away from the lengthy conventional work flow.

In conclusion, we have demonstrated that mummichog can successfully predict functional activity directly from a spectral feature table. This benefits from the convergence of genome-scale metabolic models and metabolomics. Mummichog will continue to improve as the metabolic network models evolve. We expect this method to greatly accelerate the application of high throughput metabolomics. The mummichog software is available at http://atcg.googlecode.com.

Methods

Cell culture and infection

Human peripheral blood mononuclear cells (PBMCs) were isolated from Buffy coats by separation over a Lymphoprep gradient. CD14+ monocytes were isolated from the PBMCs with MACS beads (Miltenyi Biotec, Auburn, CA) and cultured for 7 days with 20 ng/ml GM-CSF and 40 ng/ml IL-4 (Peprotech, Rocky Hill, NJ). MoDCs were then harvested, washed twice and resuspended in serum-free medium.

MoDCs (2×10⁶) were stimulated in triplicate in 48-well plates in a 200 μL volume with Yellow Fever virus (M.O.I. of 1), or mock infected. After 2 hrs, 800 μL of 10% FBS-RPMI was added to all wells. MoDCs were harvested at 6 hr or 24 hr after infection and centrifuged. Supernatants were aspirated, and dry cell pellets were frozen at −80°C. Supernatants of moDC cultures were assayed for the concentration of IL-6 and TNF using ELISA kits (BD, San Diego, CA). Three biological replicates were used for LC/MS and the concentration of IL-6 and TNF using ELISA kits (BD, San Diego, CA). Three biological replicates were used for subsequent analysis. These 7,995 features constituted the reference list \( L_{\text{ref}} \). No normalization was used because total ion counts in all samples were very similar. Student’s t-test was used to compare infected samples (at 6 hr) versus mock infections (at 6 hr), and infected samples (at 6 hr) versus baseline controls (at 0 hr).

Features with \( p < 0.05 \) in both tests were included in the significant list \( L_{\text{sig}} \). The feature table, \( L_{\text{ref}} \), \( L_{\text{sig}} \) and predictions are given in Dataset S1.

Gene expression by QPCR

For gene expression quantification, mRNA was extracted by RNeasy Mini Kit (Qiagen) according to manufacturer’s protocol, where the cell lysate was homogenized by QIAshredder spin column. Reverse transcription was performed with SuperScript III reverse transcriptase and oligo-dT (Invitrogen) according to manufacturer’s recommendation. Real-time PCR was performed on a MyiQ Icyycler (BioRad), using SYBR Green SuperMix (Quanta Biosciences). The PCR protocol used 95°C 3 mins; 40 cycles of 95°C 30 seconds and 60°C for 60 seconds. The amplicons were verified by melting curves. Quantification was performed by the ΔΔCt method, normalized by β2 microglobulin levels. The primer sequences are given in Table S2.

Analysis of human urine and yeast data

Data on human urinary samples [63] were retrieved from MetaboLights server [71]. The positive ion feature table for study “439020” contained 14,720 features. A feature is only included if its ion intensity is above 100,000 in 5 or more samples. This leaves 11,096 features, which consist \( L_{\text{ref}} \) for this study. Data were normalized by total ion counts. We next compared the metabolite difference between females (8 samples of low testosterone glucuronide level) and males (11 samples of high testosterone glucuronide level). \( L_{\text{sig}} \) is consisted of 524 features (\( p < 0.01 \) by student t-test). The original authors annotated 3,689 metabolite features, and their annotation was used to compare with the prediction by mummichog.

The yeast data [64] were downloaded from MAVEN website [32] in mzXML format. Feature extraction was performed in MAVEN through two approaches: targeted approach and untargeted approach. The targeted approach used chemical library from the same lab and produced 177 annotated features, which corresponded to 101 metabolites. The untargeted approach extracted 6318 features without annotation. After the same processing procedure as in our moDC data, \( L_{\text{sig}} \) contained 5707 features. We thus used mummichog to predict on the untargeted data, and compared the result to MAVEN annotation. The \( L_{\text{sig}} \) consisted of 426 features that were significantly different between the prototrophic wild type and the auxotrophic mutant (\( p < 10^{-4} \) by student t-test). The yeast metabolic model was compiled from BioCyc data [11].

Supporting Information

Dataset S1 Metabolomics feature table and mummichog predictions. This file contains LC/MS Metabolomics feature table from the moDC experiment, \( L_{\text{sig}} \), \( L_{\text{ref}} \) and predicted metabolites in the activity networks of moDC activation. Metabolite predictions on the human urine data and yeast data are also included.

(ZIP)
Figure S1  Many-to-many relationship in m/z feature to metabolite matching. A) Among m/z features in Lref, about 1400 can be matched to various metabolites. B) For the 77 metabolites in the activity network (Figure 3A), on average, about five metabolites share the same m/z features. Mummichog chooses the most likely metabolites based on their network activities and spectral patterns.

(TIF)

Figure S2  Example of module analysis. Five significant modules were identified \((p < 0.05, \text{ default parameters})\) in moDC infected by yellow fever virus for 6 hours. This shows module 1, with 34 metabolites, \(p = 5.19E-7\). This is one of the figures automatically generated by mummichog, metabolites colored by log2 fold change, with connecting enzymes shown as EC numbers.

(TIF)

Figure S3  Tandem mass spectrometry data for the metabolites in Figure 3B. \(MS^2\) on the left and \(MS^3\) on the right. Each run contained sequentially: top) the candidate metabolite in cell extract, middle) cell extract spiked in with pure chemical reference, bottom) pure chemical reference alone. \(MS^3\) data were also obtained for major \(MS^2\) peaks except for Inosine. Elution peaks of these samples in liquid chromatography were also matched.

(PDF)

Figure S4  The depletion of arginine and accumulation of citrulline was also observed in moDCs stimulated by dengue virus, but not by LPS. Cells and supernatant were collected after 6 hours of infection and measured by LC/MS. Representative of two experiments.

(TIF)

Figure S5  Verification of metabolite quantification. Glutathione is used as an example (m/z 308.0896, \(M+H^+\)). Left panel shows intensities of three biological replicates. The arrows point to the corresponding signals in plots of LC retention time (x-axis) vs m/z (y-axis), where a large peak is seen in YF17D sample but missing in Mock sample. In the LC-m/z plots, each dot represents a raw data point in a single scan; warmer color codes for higher ion intensity.

(TIF)

Figure S6  Validation of mummichog on the human urine data set (Roux et al. 2012), using metabolic model from BioCyc database. Mummichog predicts a network of 60 metabolites, in which 19 are not found in the original annotation. Thirty nine of the remaining 41 are agreed to the original annotation. By enforcing the presence of \(M+H^+\) ion (z option), the predicted network contains 43 metabolite. The numbers of agreed to and not in the original annotation are 39 and 4, respectively.

(TIF)

Figure S7  The prediction of mummichog is robust from the cutoff of input significant features, \(L_{df}\). The x-axis shows the different number of input features by varying p-values. The 601 features by \(p < 0.05\) were presented in the main text, and are used as reference here. The y-axis shows the number of metabolites in predicted networks. All blue bars show the number of overlaps in metabolites in predicted networks, compared to those from \(p < 0.05\).

(TIF)

Figure S8  Details of module analysis. Using the moDC data with default parameters, mummichog produces 22 modules from user data and 2406 random modules by permutation. A) shows the distribution of all module sizes. B) Black horizontal bars show the distribution of activity scores from random modules, and the black line is the fitted Gamma distribution. The red vertical bars show the activity scores of modules from user data, where those on the right are more statistically significant. C) and D) follow the same format as A) and B), but use an alternative algorithm for module finding [67], which produces 84 modules from user data and 5097 random modules by permutation.

(TIF)

Figure S9  High resolution copy of Figure 3A.

(TIF)

Table S1  Pathway analysis by mummichog revealed a number of pathways activated in the infection of moDCs by yellow fever virus. Activities in many of these pathways are also identified in the significant network modules.

(PDF)

Table S2  Primer sequences and references for quantitative PCR.

(PDF)

Acknowledgments

We thank Bill Liang, ViLinh Tran, Raghu Tadagavadi, Rajesh Nair, and Young-Mi Go for technical assistance.

Author Contributions

Conceived and designed the experiments: SL DPJ BP. Performed the experiments: SL VP SD FHS NK QAS. Analyzed the data: SL FHS. Contributed reagents/materials/analysis tools: DPJ BP. Wrote the paper: SL. Developed mummichog algorithms and software: SL. Participated in software design: FHS QAS DPJ.


