Integration of Proteomic and Metabolomic Profiling as well as Metabolic Modeling for the Functional Analysis of Metabolic Networks

Article *in* Methods in molecular biology (Clifton, N.J.) · November 2011

DOI: 10.1007/978-1-60761-977-2_21 · Source: PubMed

CITATIONS
11

READS
130

5 authors, including:

Patrick May
University of Luxembourg
147 PUBLICATIONS **2,471** CITATIONS
SEE PROFILE

Oliver Ebenhöh
Heinrich-Heine-Universität Düsseldorf
122 PUBLICATIONS **1,266** CITATIONS
SEE PROFILE

Wolfram Weckwerth
University of Vienna
349 PUBLICATIONS **8,810** CITATIONS
SEE PROFILE

Dirk Walther
Max Planck Institute of Molecular Plant Physiology
213 PUBLICATIONS **5,353** CITATIONS
SEE PROFILE

Chapter 21

Integration of Proteomic and Metabolomic Profiling as well as Metabolic Modeling for the Functional Analysis of Metabolic Networks

Patrick May, Nils Christian, Oliver Ebenhöh, Wolfram Weckwerth, and Dirk Walther

Abstract

The integrated analysis of different omics-level data sets is most naturally performed in the context of common process or pathway association. In this chapter, the two basic approaches for a metabolic pathway-centric integration of proteomics and metabolomics data are described: the knowledge-based approach relying on existing metabolic pathway information, and a data-driven approach that aims to deduce functional (pathway) associations directly from the data. Relevant algorithmic approaches for the generation of metabolic networks of model organisms, their functional analysis, database resources, visualization and analysis tools will be described. The use of proteomics data in the process of metabolic network reconstruction will be discussed.

Key words: Network reconstruction, Genome annotation, Metabolic modeling, Network expansion, Flux balance analysis, Expression analysis, Time-series data analysis, Granger causality, Systems biology

1. Introduction

Recent years have seen a rapid development of profiling technologies allowing to probe cellular systems across multiple levels of molecular organization, most importantly the metabolomic, transcriptomic, and proteomic systems levels. Although the degree of comprehensiveness still differs, available transcriptomics methods allow the near complete monitoring of the transcriptional activities of essentially all genes or genomic regions, whereas available proteomics, and even more so, metabolomics methods provide access to only a fraction of all proteins and metabolites, respectively, still,

Cathy H. Wu and Chuming Chen (eds.), *Bioinformatics for Comparative Proteomics*, Methods in Molecular Biology, vol. 694, DOI 10.1007/978-1-60761-977-2_21, © Springer Science+Business Media, LLC 2011

unseen opportunities for a holistic experimental approach creating an integral understanding of cellular systems upon applying these various profiling technologies have arisen.

Metabolic as well as signaling and regulatory pathways provide a natural framework for the integration of data from different molecular organizational levels. Pathways represent our accumulated scientific knowledge of molecular processes, structure the available data in a meaningful way, and allow the detection of coherent behaviors and, thus, a better separation of noise from real molecular signals. In particular, metabolic pathways can be expected to follow universal biochemical rules. Thus, metabolic pathways are expected to offer a suitable ordering framework even across different organisms. As a consequence, when studying system-wide responses of different organisms to external perturbations, the creation of this metabolic pathway reference framework, the metabolic network, frequently is among the first tasks when conducting systems biology experiments. Assuming that the underlying biochemical reactions are universal and catalyzed by similar enzymes, the task of assembling the metabolic network primarily means to detect all enzymes encoded in the organism's genomes – the so-called genome annotation. With this set of enzymes, all biochemically possible reactions can be derived, and thus the synthesizable set of metabolites can be determined. Comparison with actual experimental data then leads to the validation and refinement of the network, the detection of obvious gaps (missing enzymes), and a targeted search for filling these gaps (identification of enzymes in the genome).

At the same time, the available molecular profiling data sets also allow the reverse approach. Profile data, especially when followed over time, are frequently interpreted as results of as yet unknown pathways and other types of cause–effect relationships. To detect these pathways, various statistical data analyses techniques have been applied.

In this chapter, we describe the major steps involved in creating an integrated view of proteomics and metabolomics organizational domains. We will describe how the inventory of all enzymes encoded in a genome can be established, and how proteomics data can be used to obtain an improved view of the genomic complement. Flux balance analysis (FBA) as an approach to functionally characterize the resulting network is described in more detail. Furthermore, very basic statistical methods for the data-driven investigations to infer pathway associations between different molecules are introduced and relevant resources, software packages, and visualization means described.

2 Methods

2.1. Reconstruction of Genome-Scale Metabolic Networks Using Proteomics and Metabolomics Data

The basic steps involved in creating an integrated and network-based view of different molecule types in a given organism can be summarized as follows.

1. Functional gene annotation.
2. Automated genome-scale reconstruction.
3. Determination of discrepancies between the predicted network and measured data.
4. Expansion of the network to fill in the gaps and reconcile inconsistencies.

The reconstruction steps will be described in the subsequent paragraphs in more detail. Detailed description of the reconstruction process can also be found in (1, 2).

2.1.1. Functional Gene Annotation

As metabolites are processed by enzymes that in turn are encoded in the genome, the knowledge of the complete set of enzymes in a given organism is pivotal. The metabolic network reconstruction is normally done using all sequence and functional annotation data that is available in public databases combined with manual curation using literature and experimental data (see Note 1).

If available, all genomic, transcript (Unigenes or EST data), and protein sequences of the organism of interest should be downloaded from the webpage of the corresponding genome project [e.g., for *Chlamydomonas reinhardtii* from the JGI webpage (http://genome.jgi-psf.org/Chlre3/Chlre3.home.html) or the NCBI webpage (http://www.ncbi.nlm.nih.gov)). Functional annotations of genes and proteins can be retrieved from public databases or literature (see Table 1).

Enzyme functions can be obtained by transferring functional annotations like EC numbers, GO terms (3) or MapMan (4) bins across organisms using comparative analysis (see Note 1). Typically, proteins are then annotated using BLAST (5) against annotated transcripts or proteins. Instead of BLAST, more sensitive methods like PSI-BLAST (6) or HHpred (7) can be used. The annotation is transferred if a certain hit identity and score threshold hold (standard values are 40% sequence identity and a blast score of at least 50 to ensure a sufficient alignment length). Another, more reliable, method to functionally annotate a set of genes is using orthology information of an annotated genome. The Inparanoid (8) software, the OrthoMCL-DB database (9), or the KEGG (10) Orthology (KO) can be used to obtain evolutionary relationships. An automated method to map sequences to KO groups and KEGG pathways and reactions is KAAS (11) (KEGG Automatic Annotation Server), which is based on

Table 1
Public resources for functional genome annotation

Database	Data	URL
Entrez Gene	Gene annotation	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
Entrez Genomes	Genomes	http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
Uniprot	Protein annotation	http://www.uniprot.org/
Interpro	Domain annotation	http://www.ebi.ac.uk/interpro/
TransportDB	Transporter annotation	http://www.membranetransport.org/
Brenda	EC numbers	http://www.brenda-enzymes.org/
KEGG	Pathways	http://www.genome.jp/kegg/
MetaCyc	Pathways	http://metacyc.org/
MapMan	Plant pathways	http://www.gabipd.org/projects/MapMan/
GabiPD	Plant annotation	http://www.gabipd.org/
PSORTdb	Subcellular localizations	http://db.psort.org/
Pubmed	Literature references	http://www.ncbi.nlm.nih.gov/pubmed

reciprocally best BLAST hits against all KO groups of functionally related genes assigned in the KEGG GENES database. To assign functional motifs and domains, InterproScan (12) can be used. The genome annotation can provide additional information such as subcellular localization, protein subunits, and protein complexes. If no experimental subcellular localization data is available, subcellular localization of proteins can be predicted using bioinformatics tools. A comprehensive list of methods is available at: http://en.wikipedia.org/wiki/Protein_subcellular_localization_prediction.

2.1.2. Automated Genome-Based Reconstruction

The genome annotation provides lists of metabolic enzymes that are present in the organism of interest catalyzing metabolic reactions (see Note 2). The next step in the reconstruction process is to determine which biochemical reactions are carried out by these enzymes. This can be determined manually or by using automated tools. Starting from the functional annotation of a genome given as EC number, KO group, GO term, or MapMan bin, there are a number of methods (see Table 2) that can be used to produce an initial draft metabolic network or to refine an existing metabolic network filling the missing reactions (see Subheading 2.1.3). Transport reactions have to be defined to connect the separated networks of the single compartments (see Note 3).

Table 2
Automated network reconstruction methods

Network reconstruction system	URL/reference
PathwayTools	http://www.biocyc.org/
GEM System	http://www.biomedcentral.com/1471-2105/7/168
metaShark	http://bioinformatics.leeds.ac.uk/shark/
SEED	(DeJongh 2007)
AUTOGRAPH	(Notebaart 2006)
KAAS	http://www.genome.jp/tools/kaas/

2.1.3. Determine Discrepancies Between the Predicted Network and Measured Data

A metabolic draft network that has been derived from sequence homologies to known enzymes may be incomplete. First not for all enzymes, the protein sequences are known and, second, homology matches may fail because of low sequence but high structural similarities (see Notes 5, 6). Metabolic profiles determined experimentally under well-characterized conditions can efficiently be exploited to identify metabolic capabilities missing in the derived draft network. Clearly, all observed metabolites must have been produced by the organism from the provided nutrients (see Note 10). To identify discrepancies between the predicted network and measured data, the draft network derived above is analyzed by structural modeling techniques to determine whether it is capable of carrying fluxes that allow for the synthesis of the observed metabolites from the applied nutrients. Evidently, the more growth conditions have been experimentally tested and the more metabolites could unambiguously be identified, the more discrepancies may be discovered. Furthermore, it is possible to exploit proteomics data to define even more synthesis routes that the network must be able to synthesize. If, for example, observed amino acid sequences strongly indicate a gene model for which a function is clearly assigned, then it is highly plausible that this reaction takes place and thus the participating substrates and compounds must be producible from the nutrients.

The underlying test for determining whether the draft network can carry the necessary flux can in principle be performed by the method of FBA (see Subheading 2.4.1). However, to avoid the tedious step in generating a manually curated stoichiometrically balanced model that is necessary for FBA, we propose the more robust method of network expansion (13). Although with this methodology, it is not possible to quantify flux ratios, the principle capability to produce metabolites can be tested very efficiently. In comparison to FBA, it is less mathematically stringent than relying on heuristics. However, for well-curated networks, it could be shown that almost identical results for the producibility of compounds can be expected (14), using only a fraction of the computing time.

For a given set of nutrients (the *seed*) and a given set of observed metabolites (the *target*), the identification of discrepancies involves the following steps:

1. Define a suitable set of cofactors that are assumed to be present (e.g., ADP/ATP, NAD(P)/NADH(P) and Co-A).
2. Denote by S the seed set of all nutrients.
3. Determine all reactions for which all substrates are either contained in S or belong to the cofactors defined in step 1.
4. Expand the set S by all products of the identified reactions.
5. Repeat the iteration with step 3 until no new products can be added.
6. Identify all those observed target compounds that are not contained in S. The draft network cannot produce these metabolites and therefore disagrees with the metabolite profile.

A web-based front end to the network expansion algorithm is available at http://scopes.biologie.hu-berlin.de (15).

2.1.4. Expand the Network to Fill in the Gaps and Reconcile Inconsistencies

There exist several attempts to fill gaps in metabolic networks (16–18) (see Note 2). Some methods are based on analyzing the local context of the reactions, e.g., by adding reactions that belong to predefined pathways if a certain number of reactions within this pathway have already been annotated. This bears the danger of missing possible solutions that are not contained in manually and rather arbitrarily defined pathways. Other approaches are based on FBA and apply mixed-integer linear programming techniques to identify minimal sets of reactions that are needed to allow for the network to carry a flux to synthesize a given set of products. This implies the disadvantage that a stoichiometrically balanced model has first to be built and embedded in a larger network derived from databases. In Christian et al. (19) these approaches are discussed and an alternative is presented that employs the method of network expansion, which was described above to identify discrepancies between the network draft and experimental observations. The presented method has the advantage that it can directly operate on networks derived from databases and thus the integration of the draft network into a larger reference network is greatly facilitated. In general, the identification of candidate reactions that should be added to the network relies on a *draft network* (see Subheading 2.1.2) and a *reference network* (derived from a database comprising known biochemical reactions from a large number of species, e.g., KEGG or MetaCyc (20)). The algorithm involves the following steps (see Notes 7–9):

1. All reactions from the reference network, which are not part of the draft network, are written to a list of possible candidate reactions (candidate list) (see Note 11).
2. The draft network is extended by this list.
3. The network expansion algorithm is used to test whether the extended draft network is in agreement with experimentally observed metabolite profiles. If this is not the case for some target metabolites, then our complete knowledge of biochemical reactions is not sufficient to explain their presence and for these metabolites, no extension can be predicted. In the following steps, we will therefore focus only on those target metabolites that may be produced from the fully extended network.
4. Remove the reaction from the top of the candidate list.
5. Test (with the method of network expansion) whether all targets can still be produced.
6. If this is the case, permanently remove the reaction. If not, add the reaction to the network and store it as a predicted extension.
7. Continue steps 4–6 until the complete candidate list is traversed.

This greedy algorithm will result in one particular minimal extension that is sufficient to reconcile inconsistencies. To sample various possible minimal extensions, this algorithm is repeated a large number of times for different list orderings of the candidate reaction list. Comparison of the solutions can give hints about the plausibility of the occurrence of a reaction. Those reactions, for example, which are found in all solutions, are very strong candidates that indeed have to be included in the metabolic network.

The quality of the predicted extensions can be considerably improved by including genomic sequence information. By a systematic comparison of the amino acid sequences predicted by the gene models to protein sequences from other organisms, a likelihood score can be defined representing the probability that some gene in fact encodes a protein catalyzing a particular candidate reaction. This information can be used to randomize the candidate reaction list in such a way that there is a tendency for those reactions for which a strong signal is detected for a catalyzing enzyme which is encoded in the genome, placed at a later position and is thus more likely to be retained in the predicted list of reactions.

The sequence information is also useful to assign reactions in the predicted extensions to a particular gene. In this way, hypotheses are generated such as which particular genes code for which enzymes. These hypotheses are in principle testable either directly by isolation of the gene product and in vitro studies or indirectly by knockout experiments. Further hints whether predicted sequences are in fact translated are obtained by proteomics measurements as described in the following section.

During experimental validation of the predicted genes and proteins and their functions, new evidence is likely to arrive about the existence of so far unobserved proteins and metabolites. This information can then be used to reiterate the reconciliation process such that a repeated cycle of experiments and theoretical predictions result in an increasingly accurate description of the genome-scale metabolic network.

2.2. Using Proteomics Information to Improve Genome Annotation and the Metabolic Network

Often, not all genes encoding enzymes in an organism are known, or many different gene models from different gene prediction tools are available, or EST data are incomplete. The problem is even more evident in genomes of organisms that are not fully sequenced. Then, shotgun proteomics as well as transcriptomics methods (21) can be used to generate new or validate hypothetical gene models. Such a strategy also helps to eliminate metabolic reactions with experimentally unverified transcripts. Moreover, network gaps can be filled by alternative isoforms and better functional annotation of new or changed gene models can be generated.

Like EST sequencing, high-throughput, high-mass-accuracy proteomics profiling methods provide actual evidence for the presence of gene products and thus can serve as validation of gene models (22). In proteomics, peptides, and proteins are normally identified using annotated protein sequence databases. Besides applying de novo sequencing, alternative gene model predictions, exon splice graphs (23), or EST and genomic sequence translated in all six reading frames can be used to identify as of yet unannotated peptides and proteins. Exon-splice graphs compactly encode putative splicing events. In proteomics, it is standard to require at least two peptides per protein for identification. For new genes, only cases in which two or more previously unannotated peptides are mapping within a 1 kb of the genomic sequence are accepted. Identified peptides can then be used to predict new gene-models using software such as Augustus (24). The new gene models and their products can then be functionally annotated using the methods described in Subheading 2.1.1.

Figure 1 provides a schematic overview of the integrative approach using proteomics and metabolomics data and mathematical modeling to improve the quality of metabolic network.

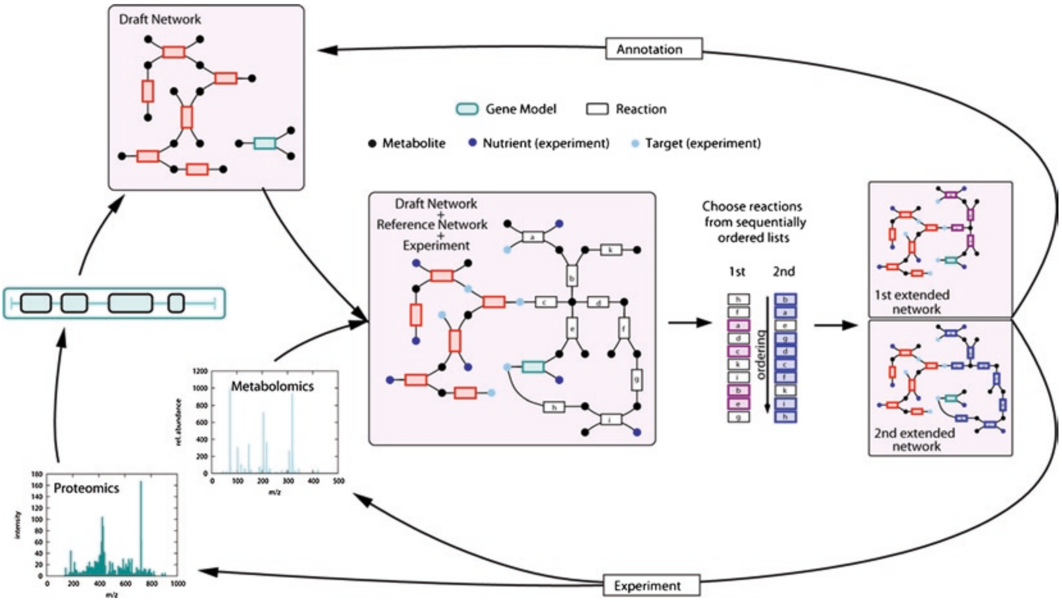


Fig. 1. Integrative approach using proteomics and metabolomics data and mathematical modeling to improve the completeness of the metabolic network. The initial network is derived from genomic data. The draft network may not be sufficient to explain the presence of all metabolites observed in metabolomics measurements or part of isolated reactions for new gene models predicted from proteomics experiments. The draft network is then embedded into a reference network consisting of reactions collected in databases such as MetaCyc or KEGG. A greedy algorithm calculates minimal sets of reactions (*extensions*) that have to be added to the draft network to make it compliant with all experimental data. A network is in agreement with observations if it is able to carry fluxes producing the measured metabolites from the applied nutrient medium. The calculation of a large number of extensions is achieved by initializing the algorithm with many differently ordered lists of reactions (*see* Subheadings 2.1.3 and 2.1.4). The solutions are compared and used to derive hypotheses about the existence of biochemical reactions and genes encoding the respective enzymes. These hypotheses can be tested experimentally or by bioinformatics methods. With this strategy, modeling, bioinformatics and experiment are combined in an iterative process to improve gene annotations and arrive at more complete genome-scale metabolic networks.

2.3. Visualization Tools for the Integrated Metabolic Pathway Analysis of Profiling Data

Various functional genomics data from gene expression, protein expression, and metabolic profiling experiments can be visualized in the context of the reconstructed metabolic network using various visualization tools (see Table 3). These visualization tools enable the visualization of the user's own experimental data in the context of the reconstructed metabolic network.

2.3.1. PathwayTools Omics Viewer

The PathwayTools Omics Viewer (25) as part of the PathwayTools software is a user data visualization and analysis tool allowing lists of genes, enzymes, or metabolites with experimental values to be drawn on a diagram of the full pathway map for an organism for which Pathway Genome Databases (PGDBs) have been developed. Examples are EcoCyc (26), AraCyc (27), YeastCyc (28), or ChlamyCyc (29).

Table 3
Network visualization tools

Visualization tool	URL
PathwayTools Omics Viewer	http://www.biocyc.org/
Cytoscape	http://www.cytoscape.org/
MapMan	http://www.gabipd.org/projects/MapMan/
Vanted	http://vanted.ipk-gatersleben.de/
Pajek	http://pajek.imfm.si/doku.php
MetaViz	http://www.labri.fr/perso/bourqui/software.html
SimPheny™	http://www.genomatica.com/technology/technologySuite.html

2.3.2. Cytoscape

Cytoscape is an open source bioinformatics software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Cytoscape supports the standard network and annotation file formats used in systems biology: GML, BioPAX, SBML, and OBO.

2.3.3. MapMan

MapMan (30) is a visualization platform that has been developed for the display of metabolite, transcript, and proteomics data onto metabolic pathways of *Arabidopsis* and other plant genomes and thus features a special emphasis on plant-specific pathways (31).

2.3.4. VANTED

VANTED (Visualization and Analysis of Networks containing Experimental Data) (32) is a platform independent tool for analyzing biological networks. VANTED combines the following features: dynamic network editing and layout, mapping of medium- to large-scale experimental data sets from different time points or conditions on networks, statistical tests, generation of correlation networks, and clustering of similarly behaving substances.

2.3.5. Pajek

Pajek (Slovene word for Spider) (33) is a program, for the analysis and visualization of large networks providing efficient algorithms for network analysis, e.g., partitions, paths, components, flow, decompositions, reduction, etc..

2.4. Functional Metabolic Network Analysis

Once a metabolic network model has been created, several approaches have been developed to investigate their quantitative and qualitative behavior. For example, it is possible to predict the flux distribution; i.e., the metabolic throughput per unit time across all reactions in the network that optimizes growth of an organism. Quantitative network analysis includes methods such as kinetic modeling using differential equations (34, 35), Elementary Mode Analysis (36) to identify subpathways that can operate at steady state thus providing an objective criterion for the definition of pathways, and Flux Balance Analysis (FBA) (35). In this chapter, we will describe FBA as it directly relates to the reconstructed metabolic network (see Subheading 2.1.1) and uses additional proteomics data such as subcellular localization of enzymes.

2.4.1. Flux Balance Analysis

Ultimately, quantitative results are sought from an integrated network analysis, in particular in the context of metabolic engineering, where optimized reaction kinetics and fluxes through the metabolic network are determined that increase yields of certain desired product metabolites. FBA has become a popular quantitative network analysis approach (35, 37, 38). Unlike complete deterministic modeling using differential equations that require the determination of (prohibitively) many reaction parameters (rate constants etc.), FBA operates under the most basic assumption of conservation of mass as reflected in the stoichiometric matrix dictated by the chemical pathways. Thus, FBA explores the possible steady-state operating modes of a given network, modes that are consistent with the conservation of mass. All interconverting processes [including transport processes (see Notes 3 and 4)] are treated as fluxes (V) with reversible reactions split into two separate fluxes, a forward and reverse flux. The change of the level of particular compound, X_i, then is the integrative effect of all fluxes, V, acting on it:

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = V_{\text{synthesis}} - V_{\text{degradation}} - V_{\text{growth/use}} \pm V_{\text{transport}} \tag{1}$$

The time dependent change of all metabolites, X (vector notation), in the system can then be computed from the product of the stoichiometric matrix, S, and all fluxes, V:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = SV \tag{2}$$

At steady state, the net change of all metabolites is zero. Thus,

$$0 = SV \tag{3}$$

Assuming additional constraints – most importantly non-negative and bounded values for fluxes and concentrations, the solution of this equation can be determined that optimizes the yield of preselected target metabolites via linear programming techniques, such that

$$T = \sum_{i=1}^{N} c_i v_i \tag{4}$$

is maximized for T, where T represents the desired optimization parameter, c_i are the coefficients (weights) to be determined for all N fluxes.

As a result, a numeric solution for all fluxes in the system is found that are consistent with an optimal yield of a particular metabolite or target parameter that can be expressed as a result of fluxes.

In general, the FBA workflow includes the following steps:

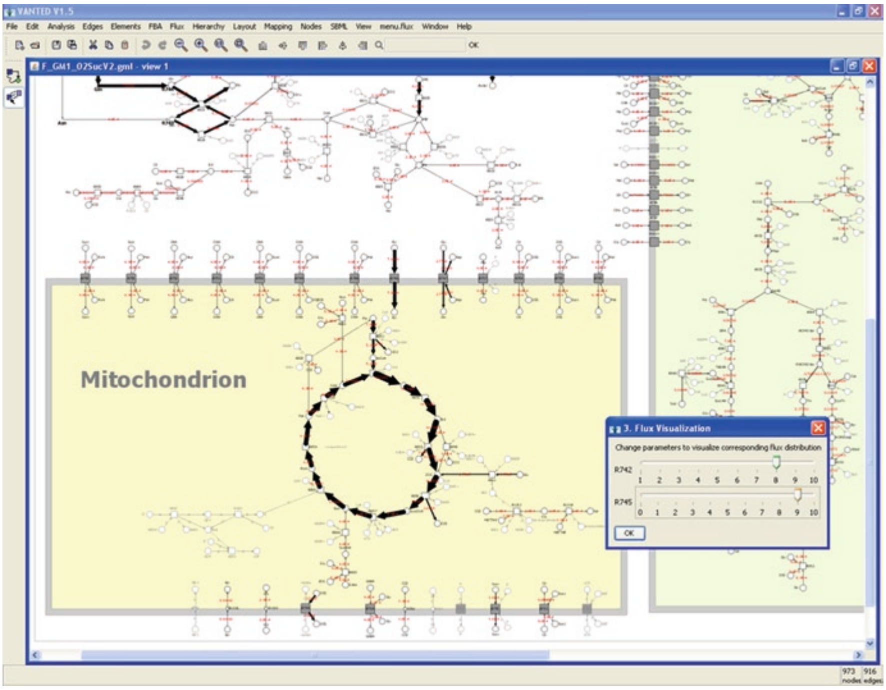
1. *Determine the metabolic network for the organism under study* (see Subheading 2.1). Of particular importance are the correct assignments of subcellular compartment in which the reactions are occurring. The reconstruction of the network also includes the generation of the stoichiometric matrix, which follows basic biochemical principles of conservation of mass (see Notes 7, 8, and 12).
2. *Define constraints.* The steady-state condition is already a limiting constraint. Other constraints on maximally possible flux values can be derived from consideration rate kinetics of particular enzymes determining the maximally possible conversion rate. Physical constraints such as thermodynamic considerations based on Gibbs free energy have recently been proposed (39) to avoid implausible solutions. The biomass composition that needs to be maintained adds additional constraints on elemental and compound composition.
3. *Specify the optimization criteria.* Define the objective function; i.e., the parameter that is to be maximized (T in Eq. 4). Examples are yield of a specific metabolite (ATP, for example), maximized growth rate and others.
4. *Solve the linear equation systems under constraints to maximize objective function.* Apply linear programming as a mathematical means to find the solutions with optimized results specified under step 3 and constraints defined under step 2. Several linear programming optimizers are available such as the ILOG CPLEX solver (ILOG, Inc. Mountain View, CA, http://www.ilog.com/products/cplex/) or the optimization routines available under the Matlab mathematical programming environment (for additional software tools, see http://en.wikipedia.org/wiki/Linear_programming).
5. *Analyze results.* FBA provides information on the possible operating modes of metabolic networks at steady-state and helps identify suitable sites for metabolic engineering efforts that aim at boosting the yield of a particular compound or rendering processes more efficient (reduced nutrient uptake).

The FBA framework also allows studying hypothetical flux distributions for knockout mutants by deleting the knocked-out gene (enzyme) from the metabolic network. Furthermore, questions of robustness (sensitivity analysis) can be addressed as well. Thus, FBA allows integrating the proteomics level (presence or absence of enzymes) with the functional consequences on metabolism. An approach to integrate gene expression levels into the FBA formalism has been described recently (40).

A visual inspection of resulting flux distributions from FBA mapped onto the metabolic network and additional analyses such as knockout studies and robustness as well as flux variability analysis can be conveniently performed using the FBA-SimVis plug-in (41) for the VANTED software. An illustration of the resulting flux distribution and their visualization in the FBA-SimVis tool is shown in Fig. 2.

Recently, the concept of FBA with focus on metabolic reactions has been expanded to also include time-dependent regulatory steps (42).

An example for the successful application of FBA to the study of primary metabolism of *C. reinhardtii* under three growth conditions (autotrophic, heterotrophic, and mixotrophic) based on a reconstructed network generated under consideration of subcellular compartmentalization was presented recently by Boyle and



Change parameters to visualize corresponding flux distribution

Fig. 2. Visualization of flux distribution in a model of barley seed metabolism with FBA-SimViz (41). Width of *reaction arrows* reflect flux values. Image courtesy Falk Schreiber.

Morgen (43). For the various conditions examined, the dominating metabolic routes were identified. Furthermore, FBA revealed the conditions associated with carbon efficiency. Thus, the study identified intervention sites for rational engineering of *Chlamydomonas* with the objective to modify its economically relevant or environmental properties (CO_2 fixation, for example).

2.4.2. FBA Software

Software programs for FBA computations include CellNetAnalyzer (44), the COBRA Toolbox (45), FBA (http://gcrg.ucsd.edu/Downloads/Flux_Balance_Analysis), and TinkerCell (http://www.tinkercell.com/).

2.5. Statistical Methods for the Integrated Analysis of Profile Data

To detect and understand relationships between molecules is a central goal of systems biology experiments that involve the parallel profiling of different molecule types (transcripts, proteins, metabolites). In the general sense, the interest is to determine, which molecules are involved in the same molecular processes. These associations can be inferred from profiling data derived from different samples taken at different steady states applying correlation followed by clustering techniques or from time series data monitoring the molecular response to external perturbations. Beyond functional associations, time series data also offer the potential to infer cause–effect relationships between molecules. The basic logic is that causes must precede effects. Thus, correlations of the time profile associated with one molecule with another molecule at later time points may be indicative, but not proof, of cause–effect relationships. In the following, we will focus our discussion of methods on the integrated analysis of protein with metabolite data. Evidently, the same concepts apply to other data types as well.

2.5.1. Correlation Analysis

As a hallmark of their association, molecules participating in the same process can be expected to follow a similar pattern of up- and down-regulation, in essence, to be correlated. Quantitatively, this is most frequently measured by their Pearson correlation. As profile data representing different molecule types (metabolites, transcripts, and proteins) can fall onto very different scales, other distance measures, such as Euclidean distance cannot be applied directly. Instead, all data sets need to be standardized beforehand, but transforming all values to a new range with zero mean and unit standard deviation. Correlation measures, on the other hand, are insensitive to absolute values, but identify similar patterns.

The linear correlation coefficient between two vectors (columns of data, e.g., level data for proteins, x, and metabolites, y, across n common samples) is defined as:

$$r_{xy} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}, \tag{5}$$

where $\overline{x}$ and $\overline{y}$ are the sample mean of x and y, s_x and s_y are the sample standard deviation of x and y and the sum is from $i = 1$ to n, the length of the data vector (see Notes 13, 14).

Based on the pairwise correlation coefficient as defined in Eq. 5, all variables (e.g., metabolites and proteins) can be clustered to identify subgroups of compounds and proteins that behave similarly. A number of different clustering techniques can be applied and depend to some degree on the question at hand (see Subheading 2.6.1, the Multiexperiment Viewer).

2.5.2. Time-Lagged Correlation Analysis of Time Series Data

Correlation analysis can be applied to identify groups of proteins, genes, metabolites that behave coherently and may thus be associated with similar processes. If time series data are available, the concept of correlation can also be used to identify potential cause–effect relationships with the ultimate goal to elucidate pathways from the data. For example, one could ask, whether a change of a particular metabolite is caused by a preceding change of enzyme levels. The basic assumption is that any cause resulting in an effect must precede the effect in time. Thus, time shifted (or time-lagged/time-delayed) correlation is performed to identify such shifted cause–effect patterns (see Fig. 3). Its use has been demonstrated for the detection of gene interaction networks (46). The conceptual expansion to correlate different molecule types is straightforward.

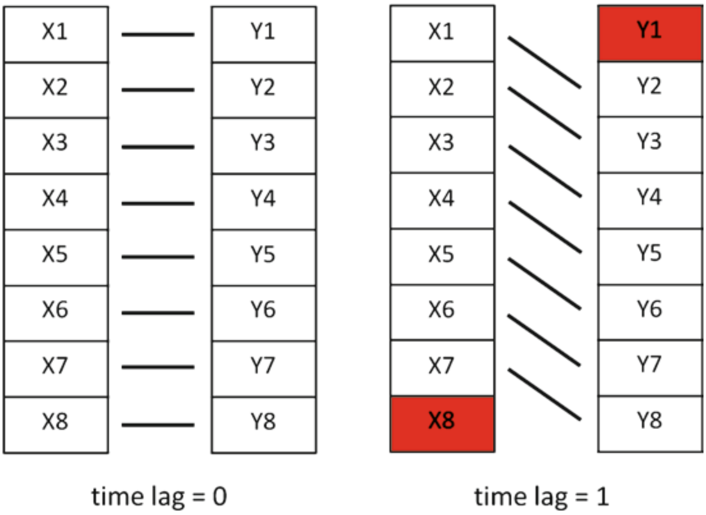


Fig. 3. Illustration of the concept of time-lagged correlation between two data vectors (e.g., protein and metabolite levels). The indexes denote the different time points in sequential order. The *lines* connecting the cells denote the pairwise association for which the correlation is computed. With every time lag increment, data points are lost (one for every variable in the example, shaded cells). Thus, the extent of possible time-delays is determined by the number of available time points. In the example, the scenario that X precedes Y is tested.

High time-lagged correlation levels are no proof of causal relationship as the observed correlations can also be caused by unconnected processes. However, absence of correlation is a strong indicator of independence.

2.5.3. Granger-Causality

As an alternative to time-lagged correlation analysis, Granger causality testing (47) can be applied to detect significant and directed (cause–effect) associations between metabolites and proteins (or any other combinations of molecule types). Granger causality tests whether past values of a time series associated with a variable (e.g., a particular metabolite) contain information that significantly improve the prediction of a future value of another variable (e.g., protein level) above and beyond the past values for this variable alone. Significance is established by applying a series of *F*-tests on the cross-term-coefficients for a linear regression model (see Eq. 6) for time dependent values of protein, $P(t)$, and metabolite data, $M(t)$, (or any other combination of variables) and computing associated *p*-values, with

$$\begin{aligned} P(t) &= \sum_{i=1}^{d} A_{P,i} P(t-i) + \sum_{i=1}^{d} A_{MP,i} M(t-i) + E_P(t) \\ M(t) &= \sum_{i=1}^{d} A_{PM,i} P(t-i) + \sum_{i=1}^{d} A_{M,i} M(t-i) + E_M(t) \end{aligned} \tag{6}$$

where $P(t)/M(t)$ denote protein/metabolite levels at time point t, the matrix A contains the linear regression coefficients, E the resulting residual error, and d is the maximal time lag (number of considered past values in the time series). In the model, if either one of the cross-term-coefficients (A_{MP} or A_{PM}) is significantly different from zero as tested by the *F*-test, past values of this variable improve the prediction of future values of the respective other variable. The variable is said to be Granger-caused by the respective other.

Granger causality was shown in the past to yield meaningful directed relationships between transcripts when applied to gene expression time series (48, 49).

Compared to correlation measures, Granger causality assigns very low mutual predictive values to variables showing monotonic behavior. Although in such cases, any time lag – forward or backward – will yield significant Pearson correlations, the Granger causality will be low, as the future values of a variable can be predicted from the variable itself. Thus, these trivial time-lagged correlations (that can, nonetheless, indicate true causal relationships) are eliminated under the concept of Granger causality.

Granger causality assumes covariance stationality, which in cases of perturbed systems is or may not be fulfilled. Nonetheless, Granger causality was shown to yield meaningful results even if this assumption is violated (49).

Granger causality computations can be performed using the MSBVAR-R package (http://cran.r-project.org/web/packages/MSBVAR/index.html; Method description: http://rss.acs.unt.edu/Rdoc/library/MSBVAR/html/granger.test.html). As discussed above for time-lagged correlation analysis, possible time lag values *d* will depend on the length of the available time series data. Obviously, the more time points available, the better.

2.6. Software for the Statistical Analysis of Multilevel Profiling Data

For the computation of pairwise correlations and clustering of data, many different software solutions and packages are available. At the most generic level, statistical computing environments, such as the freely available R or commercial solutions such as Matlab or Statistica, can be used to compute quantitative measures of interest. Because they essentially represent programming environments, they offer the greatest flexibility. By contrast, customized software packages that operate via a graphical application interface are more easily usable. Stand-alone applications [MultiExperiment Viewer (MeV)] are available as well as web-based software solutions (Metagenealyse) (see Table 4).

Table 4
Selected software packages for integrated, multivariate data analysis

Software	Commercial/free	Source	Features
Multiexperiment Viewer	Free	http://www.tm4.org/mev.html	Menu-driven statistical analyses options including biclustering, principal component analysis, correlation network generation
Statistica	Commercial	Statsoft, http://www.statsoft.com	Implementation of most clustering and many multivariate data analysis techniques.
R	Free	http://www.r-project.org/	Statistical computing environment with implementations of essentially all known statistical procedures
Matlab	Commercial	http://www.mathworks.com	Mathematical and statistical programming environment
Metagenealyse	Free	http://metagenealyse.mpimp-golm.mpg.de	Web-based suite of statistical analyses including imputation of missing values, clustering, principal component analysis, independent component analysis

2.6.1. The MultiExperiment Viewer

Designed for the analysis of microarray gene expression datasets, the freely available MeV offers a broad spectrum of standard and advanced statistical data analysis methods that can also be applied to other data types. Most noteworthy is the very intuitive graphical user interface, in which the various applied methods remain neatly organized such that the results of the various approaches are easily comparable. For the purpose of integrated analysis, the various clustering methods [hierarchical, K-means performed optionally as biclustering; i.e., simultaneously clustering rows (e.g., representing protein levels) and columns (e.g., representing samples)] can be applied. The program allows choosing between different distance measures. Without data standardization, correlation measures are most appropriate as different variables can fall into different value ranges. From the computed correlations, network views can be generated (so called Relevance Networks), thereby quickly revealing any significant associations between the different molecules. The program also allows investigating whether particular functions are overrepresented in user selected clusters. For several organisms and gene expression platforms, built-in annotation files are available. For other custom data, custom annotation files need to be generated and uploaded. Note, as the program assumes to process gene expression data, some options and predefined labels may not apply. For data import,if treatment minus control datasets are to be analyzed, choose the

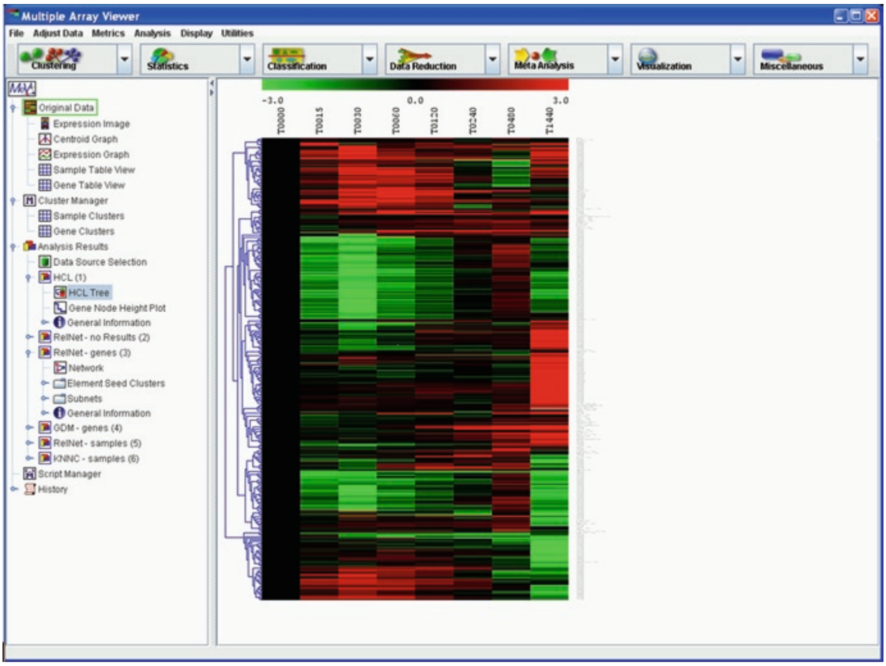


Fig. 4. Hierarchical clustering of molecular level data based on Pearson correlation distances using the Multiexperiment Viewer. Columns refer to different samples, rows correspond to the different gene transcript levels and metabolite levels. Evidently, any other data types can be treated similarly (protein levels, for example). Similar correlation patterns will result in a clustering of the corresponding molecules that may be indicative of functional association.

"Two-color Array" option as then also negative values will be colored appropriately, otherwise choose "Single-color Array" if only positive values are contained in the data matrix. Despite these idiosyncrasies, we find the program very valuable to easily perform sophisticated statistical analyses on the data at hand. Figure 4 shows an example of hierarchical clustering of molecular level data across different samples using the MeV.

A comprehensive review of integrated data analysis methods can be found in (50).

3. Notes

Here, we list some common problems encountered during automated genome-scale metabolic network reconstruction that can be used as a guide for the use of such methods (more details can be found in Feist et al. (2)) and add notes on the application of statistical concepts for the inference of pathways from data.

1. Functional annotations can change very quickly, but annotations are not continuously updated. Try to use regularly updated databases and automated methods for updating.
2. Manually check your reconstructed network for incorrect annotations or use methods that can test for inconsistencies (51).
3. Transporter reactions often have to be added manually, because annotation of transporters is still very insufficient.
4. The correct assignment of the enzymes to their respective subcellular compartment is of particular relevance for the analysis of metabolic pathways.
5. Protein–enzyme relationships are often not clearly defined. Problems can arise from the incorrect or missing annotation of isozymes, subunits, and protein complexes.
6. Reactions are often unspecifically defined. They can be associated with general classes of compounds, which can result in ambiguous connections in networks. Examples include electron carriers (NAD and NADP) or D-glucose (α-D-glucose and β-D-glucose).
7. Reactions are often unbalanced for H, C, P, N, O, or S in public databases (52).
8. Reactions are most often defined as reversible throughfully are not. Automated methods have been developed to address this problem (53, 54).
9. The protonation state of metabolites within reactions are often wrongly annotated.
10. Enzymes often need cofactors to be functional. The network must be able to produce them.

11. Often network and pathway annotation is derived by homology, but not all pathways are general across species, e.g., the photorespiration pathway between algae and higher plants.
12. In FBA, it is essential that the network is stoichiometrically fully balanced as otherwise the conservation of mass criterion is not fulfilled. Networks should also be balanced with regard to charge.
13. In correlation analysis, fewer data points will lead to increasing proportions of high correlation levels. In the extreme case of only two data points, the correlation will always be perfect, but trivial. Especially for the interpretation of time series data, where typically only few data points are available, this effect needs to be taken into account. Every additional time point significantly improves the statistical power. With six time points, there are 720 random orderings possible. By adding one more time point, this number goes up to 5,040. As a consequence, establishing statistical significance via randomized (shuffled) data set will yield much improved results in the latter case.
14. The Pearson correlation coefficient is sensitive to outliers and assumes Gaussian distributions. Thus, an apparent high degree of correlation can also result data points that are far removed from the majority of data points. To circumvent this problem, the rank-based Spearman correlation coefficient should be used. Instead of correlating the original values, the correlation is computed using the respective ranks associated with original level data in the different samples. Thus, the impact of outliers on the overall correlation is reduced significantly as the maximal rank difference can only be one unit. In practice, both measures should be used and/or the observed data points be examined beforehand for occurrences of outliers.

Acknowledgements

This work was supported by BMBF-funded GoFORSYS systems biology project. We wish to thank Falk Schreiber for generously providing us with Fig. 2.

References

1. Reed, J. L., Famili, I., Thiele, I., and Palsson, B. O. (2006) Towards multidimensional genome annotation. *Nat Rev Genet* 7, 130–141.
2. Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L., and Palsson, B. O. (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* 7, 129–143.
3. Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D. J., and Moulton, V. (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 24, 2252–2253.

4. Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, S. Y., and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* 37, 914–939.
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* 215, 403–410.
6. Altschul, S. F., and Koonin, E. V. (1998) Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem Sci* 23, 444–447.
7. Soding, J., Biegert, A., and Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33, W244–W248.
8. Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314, 1041–1052.
9. Chen, F., Mackey, A. J., Stoeckert, C. J., Jr., and Roos, D. S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34, D363–D368.
10. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36, D480–D484.
11. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35, W182–W185.
12. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* 33, W116–W120.
13. Handorf, T., Ebenhoh, O., and Heinrich, R. (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J Mol Evol* 61, 498–512.
14. Kruse, K., and Ebenhoh, O. (2008) Comparing flux balance analysis to network expansion: producibility, sustainability and the scope of compounds. *Genome Inform* 20, 91–101.
15. Handorf, T., and Ebenhoh, O. (2007) MetaPath Online: a web server implementation of the network expansion algorithm. *Nucleic Acids Res* 35, W613–W618.
16. Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., and Palsson, B. O. (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 103, 17480–17484.
17. Satish Kumar, V., Dasika, M. S., and Maranas, C. D. (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 8, 212.
18. Green, M. L., and Karp, P. D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5, 76.
19. Christian, N., May, P., Kempa, S., Handorf, T., and Ebenhoh, O. (2009) An integrative approach towards completing genome-scale metabolic networks. *Mol Biosyst* 5, 1889–1903. DOI: 10.1039/b915913b.
20. Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., Walk, T. C., Zhang, P., and Karp, P. D. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 36, D623–D631.
21. Manichaikul, A., Ghamsari, L., Hom, E. F., Lin, C., Murray, R. R., Chang, R. L., Balaji, S., Hao, T., Shen, Y., Chavali, A. K., Thiele, I., Yang, X., Fan, C., Mello, E., Hill, D. E., Vidal, M., Salehi-Ashtiani, K., and Papin, J. A. (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat Methods* 6, 589–592.
22. May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhoh, O., Weckwerth, W., and Walther, D. (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics* 179, 157–166.
23. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci U S A* 105, 21034–21038.
24. Stanke, M., and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33, W465–W467.
25. Zhang, P., Foerster, H., Tissier, C. P., Mueller, L., Paley, S., Karp, P. D., and Rhee, S. Y. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol* 138, 27–37.

26. Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M., and Karp, P. D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* 33, D334–D337.
27. Mueller, L. A., Zhang, P., and Rhee, S. Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol* 132, 453–460.
28. Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., Hong, E. L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C. L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., and Cherry, J. M. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* 32, D311–D314.
29. May, P., Christian, J. O., Kempa, S., and Walther, D. (2009) ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii. *BMC Genomics* 10, 209.
30. Usadel, B., Nagel, A., Thimm, O., Redestig, H., Blaesing, O. E., Palacios-Rojas, N., Selbig, J., Hannemann, J., Piques, M. C., Steinhauser, D., Scheible, W. R., Gibon, Y., Morcuende, R., Weicht, D., Meyer, S., and Stitt, M. (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol* 138, 1195–1204.
31. Goffard, N., and Weiller, G. (2006) Extending MapMan: application to legume genome arrays. *Bioinformatics* 22, 2958–2959.
32. Junker, B. H., Klukas, C., and Schreiber, F. (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* 7, 109.
33. Batagelj, V., Mrvar, A. (1998) Program for large scale network analysis. *Connections* 21, 47–57.
34. Fell, D. A. (1992) Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J* 286 (Pt 2), 313–330.
35. Varma, A., and Palsson, B. O. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl Environ Microbiol* 60, 3724–3731.
36. Schuster, S., Fell, D. A., and Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* 18, 326–332.
37. Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003) Advances in flux balance analysis. *Curr Opin Biotechnol* 14, 491–496.
38. Lee, J. M., Gianchandani, E. P., and Papin, J. A. (2006) Flux balance analysis in the era of metabolomics. *Brief Bioinform* 7, 140–150.
39. Hoppe, A., Hoffmann, S., and Holzhutter, H. G. (2007) Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol* 1, 23.
40. Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T. Y., Moody, D. B., Murray, M., and Galagan, J. E. (2009) Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol* 5, e1000489.
41. Grafahrend-Belau, E., Klukas, C., Junker, B. H., and Schreiber, F. (2009) FBA-SimVis: interactive visualisation of constraint-based metabolic models. *Bioinformatics* 25, 2755–2757.
42. Lee, J. M., Gianchandani, E. P., Eddy, J. A., and Papin, J. A. (2008) Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol* 4, e1000086.
43. Boyle, N. R., and Morgan, J. A. (2009) Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC Syst Biol* 3, 4.
44. Klamt, S., Saez-Rodriguez, J., and Gilles, E. D. (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol* 1, 2.
45. Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. O., and Herrgard, M. J. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* 2, 727–738.
46. Schmitt, W. A., Jr., Raab, R. M., and Stephanopoulos, G. (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res* 14, 1654–1663.
47. Granger, C. W. J. (1980) Testing for causality: a personal viewpoint. *J Econ Dyn and Contr* 2, 329–352.
48. Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009) Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics* 25, i110–i118.

49. Mukhopadhyay, N. D., and Chatterjee, S. (2007) Causality and pathway search in microarray time series experiment. *Bioinformatics* 23, 442–449.

50. Steinfath, M., Repsilber, D., Scholz, M., Walther, D., and Selbig, J. (2007) Integrated data analysis for genome-wide research, *EXS* 97, 309–329.

51. Sauro, H. M. and Lugalls, B. (2004) Conservation analysis in biochemical networks: Computational issues for software writes. *Biophys Chem* 109, 1–15.

52. Gevorgyan, A., Poolman, M. G., and Fell, D. A. (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics* 24, 2245–2251.

53. Henry, C. S., Jankowski, M. D., Broadbelt, L. J., and Hatzimanikatis, V. (2006) Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys J* 90, 1453–1461.

54. Kummel, A., Panke, S., and Heinemann, M. (2006) Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* 7, 512.