Oliver Fiehn
University of California Davis Genome Center, 451 E Health Sci. Dr., Davis CA 95616, USA
email ofiehn@ucdavis.edu
URL http://fiehnlab.ucdavis.edu

Cellular metabolomics: the quest for pathway structure

Metabolomics can be used for two major and very different purposes: the screening for differences between global metabolic fingerprints of cohorts of populations, which is often referred to as metabonomics, or efforts to understand the regulatory structure of metabolic pathways, its connectivity, control of cellular concentrations and fluxes of metabolites, and partitioning of metabolic products between cellular compartments and excretion. Almost certainly, any biomarkers or major differences that have been identified in metabolic fingerprinting (resp. metabonomics) will lead to the next level of query, a quest for an in depth mechanistic understanding, e.g. why certain biomarkers were specific for a given biological condition or perturbation. *Vice versa*, once metabolic events are understood in a comprehensive way for a given cellular system, the logic next step is to ask how the observed regulatory structure responds to external stimuli and how its metabolic responsiveness relates to the integration of the cellular regulation into larger systems, be these on the level of tissues, organs or the organism. Even for unicellular systems it is known that the excretion of metabolites and ultimately, the growth rates of cell cultures depend on the competition and interaction between these cells, especially when these are present in a mixture with very many other cell types or unicellular organisms. Although there is a different emphasis on understanding metabolic responses in a larger context between the two major approaches, eventually a better understanding of cellular regulatory circuits is always sought. On the other hand, the emphasis of global metabolic fingerprinting[1] (metabonomics) is clearly the ability to distinguish different metabolic states and to make predictions on the eventual outcome and the fate of organisms in a given biological condition. A similar kind of predictability is also key for cellular metabolomics, which can be used for example in bioengineering efforts[2] with the aim of increasing specific metabolic levels[3], or for pharmacological approaches in which certain drugs are employed to exert control effects without redirecting unrelated metabolic fluxes in an unwanted way. Concluding, both major aims of metabolomics, ascertaining differences between metabolic systems and understanding metabolic control, are two complementary efforts which rely on each other. The major focus of this chapter is highlighting difficulties for cellular metabolomics in understanding metabolic networks. Why do we have so many unknown metabolites and how can we identify them in a rapid manner? Why are metabolic networks behaving in a highly flexible manner, yet are supposed to have a rigid structure of biochemical properties? Why is it so hard to predict metabolic systems despite decades of work, and the relative low number of enzymes involved in primary metabolism and energy regulation? These and related questions are needed to be answered in order to make full use of the prospects of metabolomics.

1. How large is the metabolome?

It is a long standing observation in many metabolomic research projects that the number of detected compounds is far higher than predicted by standard metabolic pathways. For example, the number of metabolites that were expected in the bacterium *B. subtilis* were estimated to account for a maximum of 800 compounds, but more likely even fewer, when metabolic pathways were reconstructed from the genome sequences and enzyme coding genes. However, a combination of three capillary electrophoresis (CE) methods coupled to mass spectrometry (MS) identified and quantified up to 1,500 metabolites from *B. subtilis*[4,5],

far more than expected from the simple calculations given above. One can argue that not all of these detected signals were accountable for genuine intracellular metabolites but also might be caused by partial degradation or artifact formation during the sample preparation procedure. On the other hand, many metabolites might still have been missed by the aforementioned analytical chemistry methods. The strength of these techniques is the detection of charged molecules, and arguably, most metabolic intermediates bear acidic or basic functional moieties that would enable good separation by CE/MS[6]. However, both capillary electrophoresis and mass spectrometry are notoriously poor in separating and ionizing nonpolar (hydrophobic) and uncharged metabolites which are very likely also to be present in plasma membranes and subcellular structures.

One may further look at mammalian cellular metabolomes. By a similar way of calculating the number of likely metabolites that originate from the sequenced genome and a consecutive prediction of biochemical pathways, some 2,400 compounds have been calculated to comprise the complement of internal human metabolites. Does this actually hold true? There is valid reasoning in arguing that mammals do not need to synthesize many compounds which can rather be digested through food sources, which may especially be true for those compounds that would be energetically costly to produce. However, there is also growing evidence that many enzymes are far less specific than previously assumed and might thus accept a variety of substrates that can be used for oxidative purposes, or that can be used as building blocks in biosynthetic routes. Given the converging role of catabolism and the great variety of foods that can be used by omnivores, it is imminent that a superficial view of mammalian enzymes does not hold to be as selective as is suggested by (reconstructed) genomic pathways[7]. It is simply not very realistic to exclusively reflect on (theoretical) substrate preferences and disregard the reality of mammals as food-eating organisms, for which the distinction between exogenous and endogenous compounds is far more artificial than, say, for photoautotrophic systems like plants. In addition, compounds that are produced and excreted by the gut microflora add to the complexity of mammalian metabolomic systems[8]. In humans, the metabolome therefore consists of a high complexity that relies on multiple parameters, not only on the human genome and reconstructed biochemical networks but also on food preferences and the intestinal microflora. Recently, first attempts have been reported on the prediction of metabolites that originate from ingestion of xenobiotics[9], but too little is yet known about the dynamics of adsorption, transport and metabolism of complex diets to directly infer the complete metabolome within the human body.

Therefore, a complete analysis of the metabolome of photoautotrophic organisms and unicellular systems should, in principle, be much simpler. Many of these organisms have a clear preference for the major source of carbon (carbon dioxide for photoautotrophs, simple small molecules like acetate or glucose for microbial systems). Why is the complete set of small molecules unknown even for these rather simple systems? One major reason is that still, many genes have only been vaguely annotated with properties such as ‘catalytic activity’. The elucidation of pathways in which the corresponding putative enzymes work is a tedious and not always straightforwardly interpretable. Consequently, functional genomics in such organisms necessarily calls for *de novo* structural elucidation of ‘unknown’ signals that were detected using advanced analytical chemical instrumentation.

2. Metabolite identification

The identification of metabolites is a less clear term than commonly believed[10]. First, a clear distinction is needed between annotating peaks as ‘known compounds’ from a series of chromatograms and *de novo* structural elucidation. Secondly, both tasks face different

challenges. For example, peaks in GC/MS may be annotated as ‘L-Aspartate’, however, in certain cases, also D-Aspartate might occur which would be indistinguishable from its L-isomer if not special precautions are taken. Conversely, even famous and Nobel-prize winning work for in depth *de novo* structural elucidations later turn out to be wrong[11], despite the many steps and various spectroscopic techniques that were involved. Therefore, any metabolite identification, especially in metabolomics, must precisely define how compound names are supported by experimental data and which analytical method and algorithm was used, including threshold levels.

It is now clear that there are different levels of confidence in naming metabolites. Many peaks will not safely match the criteria for clear-cut annotations, and accurate *de novo* identifications are too costly and too slow for most studies. However, if confidence levels are clearly ascribed, it might be valuable to annotate unknown analytical signals with a less precise but still biochemically meaningful name. Consequently, metabolite annotations may be structured into four major groups:

(i) compounds that are identified by data acquisition of at least two different physicochemical parameters and using authentic reference standards. Typical examples comprise the analyses of small molecules such as cholesterol by volatility (in gas chromatography) and mass spectra (ion fragmentation pattern). As stated above, confidence thresholds for the identification algorithm must be given (e.g. retention time window, mass spectral similarity).

(ii) Secondly, compounds could be tentatively identified by deferring putative chemical structures from physicochemical properties. This process is called dereplication, and a potential workflow is detailed in figure 1. Further, comparison of spectral properties of unknown signals to spectral libraries may lead to tentatively identified annotations, if the spectral similarity score and matching physicochemical parameters are beyond reasonable doubt. Usually, this process would not distinguish between closely related isomers.

(iii) Next, the structural information that is gained along the analytical process may not be sufficient to derive exact chemical structures or a small list of tentatively identified compounds, but it might still enable a classification of unknown compounds to a certain chemical group such as ‘carbohydrate’. Such classifications may be aided by rules obtained supervised statistics carried out on larger mass spectrometric libraries or deduced by expert knowledge of mass fragmentation rules.

(iv) Last, in unbiased metabolomic surveys there will always be analytical signals that are just too low abundant, too uncommon or too poor in spectral information to be classified to a certain group of metabolites. These compounds can still be used for phenotyping approaches using relative quantification of signal intensities between genetic or treatment perturbations, but in order to derive biochemical or mechanistic insights, the isolation of microgram quantities of these metabolites would be necessary.

Obviously, the preferred route of consistent, reliable and routine metabolite identification is to work with authentic reference compounds. Given the theoretically unlimited size of metabolomes, especially in heterotrophic multicellular organisms, only a fraction of signals in metabolite profiles will generally be annotated by this way. It should become good metabolomic practice that at least the common set of conserved metabolites in primary metabolism are available as reference standards to verify certain signals. Publications should indicate which level of certainty is associated with metabolite naming, including referencing to authoritative chemical databases such as PubChem[12,13], the (commercial) chemical

abstracts service CAS or the InChI code[14]. In addition, if metabolite signals are reported that need yet to be structurally characterized, it is mandatory to label these signals in a way that enables tracking physicochemical characteristics (e.g. mass spectrum, quantification ion and chromatography retention times). Furthermore, it should be good reporting practice to label such unknown compounds in a way that is consistent for each specific laboratory in order to potentially learn about responses for this novel metabolite across biological studies, and hopefully even to exchange information about such compounds between laboratories. In 2005, an initiative has been formed under the umbrella of the metabolomics society that tries to foster such data exchange and re-use of metabolomic data by drafting and implementing 'reporting standards' on metabolomic studies[15]. The idea for this initiative is that metabolomic studies are so rich in data that novel conclusions may be derived if datasets are investigated from more than one point of view. Obviously, such efforts directly benefit from better strategies to structurally elucidate, annotate and report on metabolite identities.

As pointed out above, the identification of metabolic signals is best performed using authentic reference compounds. Where this is not possible, metabolite identity must be unraveled *de novo* as detailed as possible[16]. Classical methods have involved isolation of compounds followed by in-depth structural characterization using spectroscopic techniques. However, there are caveats. The parameter space for stereo configurations and positional isomers is tremendous, and even with elegant strategies and long time efforts, initial identifications have often proved wrong[11]. Furthermore, there is the risk that identifications are carried out on compounds that are later found to be already published in less well-known journals. Such problems may be circumvented by applying a strategy that aims at tentatively annotating potential structural candidates from databases before performing more laborious work on isolation and *de novo* identification. Annotations would need to combine all available physicochemical information from a separation and spectral characterization of a specific compound without necessarily isolating it, and a workflow is outlined in figure 1. The best way to approach such tentative identifications is by utilizing the combination of chromatography and mass spectrometry, calculating parameters from this information, and then matching these parameters with theoretical values that are calculated from molecular structures of database entries.

Two strategies may be distinguished: starting from elemental formulae and chemical databases or starting from mass fragment spectra and MS libraries. The first approach focuses on the elemental composition. Most recently, we have shown in a chemoinformatic approach that even 0.1 ppm mass accuracy would not enable unambiguous calculation of elemental formulae of low molecular weight compounds[17], whereas mass spectrometers with 3 ppm mass accuracy and 2% relative isotope ratio accuracy enables constraining compositional calculations in a most dramatic way to one or just a few candidate formulae. Such use of isotope ratio data can be applied to both LC- or GC-based methods[18,19]. Calculated elemental compositions will be used to query chemical and biochemical databases, resulting in potentially a couple of thousand candidate structures. Good databases for these purposes are the publicly available *PubChem* effort (5 million entries as of January 2006) or the commercial *Dictionary of Natural Compounds* DNP (200,000 structures). For these structures, physicochemical parameters can be calculated, e.g. boiling points, lipophilicity ($\log K_{ow}$) and pK_a values. In the next step, the same parameters can be calculated for the unidentified peaks using the experimental retention time information from liquid or gas chromatography. Matching the experimental to predicted parameters, along with the determination of elemental compositions, will constrain the candidate structure list to just a few structures, most of which will be positional isomers or closely related compounds. In

principle, the commercial *Chemical Abstracts Service* (CAS) could also be used for the identification of a list of potential structures. However, this service does not allow batch queries of structure searches and thus cannot be embedded into automatic algorithm routines because it needs to be handled in a manual way.

This constrained list of candidate structures can further be confined by matching tandem mass spectrometric information (MS/MS spectra) with theoretical fragmentations. Such theoretical MS/MS spectra can be predicted from commercially available software solutions based on proposed structures and known fragmentation rules (Mass Frontier)[20]. However, so far MassFrontier has only implemented positive ionization rules, and it obviously relies on fully characterized fragmentation mechanisms published in literature. Only if no structure from the initial list of tentative compounds remains, unidentified peaks must be regarded as novel metabolites which could be subjected to classical *de novo* identification, including isolation of the compound and two dimensional NMR analysis.

A second approach would query mass spectral libraries. Once enough mass spectra have been annotated with structures, such as the case for GC/MS spectra under electron impact ionization, further processes can be added to on-line structural annotations. For example, the large NIST 5.0 mass spectral database comprises some 5,000 compounds with trimethylsilylated moieties, the most common derivatization technique used in metabolomics. From these compounds, substructures can be generated to be used as training data sets for supervised statistical tools. Such algorithms aim at learning rules to automatically annotate unknown mass spectra to belong to certain chemical groups such as ‘carbohydrate’, ‘primary amine’, and other metabolite families. The advantage of this approach is that multiple supervised methods may be tested simultaneously (such as partial least square, linear discriminant analysis, tree-based models, feature selection, association rule models, and others) which can then be investigated with respect to false discovery rates and robustness. In conjunction with calculating boiling points from retention indices, it will be further possible to assign the size of the molecules, e.g. mono-, di- and trisaccharides, sugar alcohols and other structural information such as the presence of furanoside and pyranoside rings. For any peak for which very high similarity or even virtual spectral identity is found, physicochemical properties (e.g. boiling points) could be calculated from the structures and matched with the experimentally determined parameters.

Only if no candidate structure remains after exploiting the techniques outlined above, it is reasonable to assume that an unidentified compound is a truly novel metabolite which would need to be identified using classical de novo compound identification, combining various structural characterization techniques including NMR.

3. Pathway identification

Once all biologically relevant metabolic signals are annotated in the workflow schema given above, these structures need further be associated to a biochemical pathway and a biological function in order to aid cellular interpretation of metabolomic findings.
A first step utilizes common pathways that are compiled in consensus biochemical maps such as BioCyc[21], or specific maps like AraCyc and comparable overview charts. A general layout of reconstructed pathways is outlined in figure 2. These maps are certainly a good start for annotating compounds to pathways in which they are involved, but unfortunately, four major

problems can be outlined: (a) for any given organism or tissue, these maps are usually sparse and do not detail the complement of already existing biochemical knowledge, especially they do not contain information for many less common metabolites, (b) a number of metabolites that are believed to be involved in certain pathways cannot readily be determined by a given analytical-chemical technique (depicted as 'empty boxes' in figure 2), (c) many novel metabolites may be detected (depicted as Y1-Y9 in figure 2) for which no enzymes or biochemical reactions may be known and (d) the pathway topology and the directionality of enzyme reactions are often deferred from homology to other organisms and may reflect the *in vivo* function in the cellular system under study.

Therefore, pathways have to be elucidated using biochemical, genetic and molecular biologic approaches which may be summed as 'functional genomics'. Three potential outcomes are possible for functional genomics approaches to pathway elucidations.

(A) Gene products may comprise specific catalytic activity, converting
- already known substrates to known products in bypasses of classical reactions, or in cellular compartments that usually do not comprise these pathways.
- known substrates to novel compounds, e.g. in anabolic reactions to fulfill specific biological roles such as communication and defense.
- novel substrates to known compounds that may then merge into mainstream metabolism, e.g. in catabolic reactions to control turnover of compounds that were synthesized in the aforementioned process
- novel compounds to other novel compounds, which then derive a completely new pathway.

(B) In addition, unspecific enzymatic activity must be considered, for example processes that convert a variety of known substrates to a plethora of products which can then be utilized for cellular communication or defense. Examples would be P450 monooxygenases or enzymes involved in release of plant volatiles. The evolutionary and biological roles of such unspecific broad band anabolic processes are poorly understood. Emission of a variety of different compounds instead of a single specific metabolite might aid in inter- and intraspecies communication, where signals are perceived in patterns rather than in activation of a single (specific) receptor.

(C) Last, novel compounds may be produced by non-enzymatic processes such as oxidation, hydrolysation or even cleavage and condensation. Molecular biological analysis of non-annotated genes may lead to alteration of metabolite profiles, however, in order to conclude that enzymes encoded by these genes actually produce the final metabolic products, the exact reaction mechanisms must be worked out. Even more difficult to unravel are products that are entirely produced by physicochemical products but are only present at certain physiological conditions, say heat stress. Generally, metabolic products are variable by interactions of *genotype x environment x time x spatial location*. This matrix of parameters complicates any simple relationship of using metabolic phenotypes to understand gene functions, so any pathway hypothesis that is developed by metabolomics approaches needs to be verified by molecular and biochemical studies.

Given these constraints, novel pathways might be unraveled for compounds like Y3 and Y9 in the generalized pathway given in figure 3. Due to the reasons outlined above, such pathway elucidation is a rather tedious process that usually would not allow quick mapping of the other novel metabolites Y1,2,4-8 into the biochemical pathway structure.

Furthermore, the situation may even be more complicated than outlined so far. We have considered linear 1:1 relationships of enzymes, substrates and products that can be unraveled

using classical biochemical or modern molecular biology tools. However, there is still the question why there are seemingly more metabolites than enzymes, why there is so much diversity of metabolomes within a genus or between closely related species, and why crosses of these species will often reveal metabolites that are present in neither of the parents. This can hardly been explained by metabolite channeling or by kinetic parameters. This phenomenon might rather point to differences in substrate availability and transport between cellular compartments, organs, or even within a cellular compartment. At least in eukaryote cells, but likely also in prokaryotes, the intraplasmic space (e.g. in the cytosol) cannot be regarded as an aqueous solution that allows free diffusion of substrates. We might need to consider the interaction between different protein complexes that carry enzymatic activities, in addition to allelic complementation of missing pathway links that may distinguish crosses from parental lines. Protein complexes are focus of very active research in many areas of biology and biomedicine, but so far, enzymatic consequences have rarely been studied. It is known that protein folding, topology and protein complexes rely on posttranslational modification as well as allosteric modification, both of which may largely change in response to genetic differences (e.g. in crosses) or environmental perturbation (e.g. stress). Consequently, the formation of novel compounds may also largely depend on the actual formation and disassembly of such protein complexes in a given intracellular environment. Such events would easily be missed by typical molecular biological techniques which focus on the identity and activity of a single enzyme, e.g. by over expression of eukaryote enzymes into *E. coli* for purification purposes. It may be due to these theoretical difficulties that just a few examples have been reported so far where use of metabolite profiling actually led to the discovery of new pathways, such as for a direct pathway from glycine to glyoxylate in yeast[22].

4. Omics data integration

The paradigm of (molecular) biology implies a more or less linear hierarchy from genome to phenotype. This linearity would start at gene expression, splicing, translation to encoded proteins, posttranslational modifications, and eventually continue to metabolites as victims of the overall process, which are regarded as useful tools to monitor and predict the ultimate organismal phenotypes. In current research proposals, research panels and scientific boards, this view on the biological paradigm leads to demanding an integration of data across these levels of cellular organization. The integration of data from different levels of cellular regulation focuses on biochemical maps that are derived from genomic annotations and homology of genes and enzymes to well-studied organisms[23].

Theoretically, this is a compelling idea that might help understanding cellular biology by using the data to model the dynamics of the organism in a Systems Biology approach. However, the linearity of the paradigm represents an overly simplistic view that does not lend a good framework for actual data integration beyond biomarker detection for classification purposes. Despite a number of years that Systems Biology has been announced as a valuable goal, the number of research papers and the quality of these yet do not justify the verve with which the demand for ‘Omics data integration’ is put forward.

The problem is that the level of complexity increases with each level of cellular organization. The genome itself can readily be used for in depth studies and comparisons, but in many organisms, there is a *1:n* relationship on subsequent levels of organization. One gene may be spliced and transcribed into more than one gene product, one mRNA may be translated and modified to more than one protein, and one enzyme may work on one or more substrates and may be involved in many pathways. Consequently, there is no easy way to infer cellular

regulation from metabolite levels, at least not for higher organisms, and especially not for efforts integrating transcriptomics and metabolomics levels[24,25].

Even for simple and well studied unicellular models, the dynamic of metabolism can only be modeled for a couple of seconds after a certain perturbation[26]. In eukaryotes, cellular compartmentation and metabolic specialization of organs further complicate any reasonable biological interpretation of findings beyond simple statements such as 'the rate of glycolysis is increased'. Recently, a study on yeast metabolism under sulfur deprivation using a combined approach of proteomics and metabolomics revealed that predictions of use of pathways could not be made on transcriptomics or proteomics alone[27].

For higher organisms, it is a truism these consist of many organs, each organ may include many tissue types and each tissue type may comprise various cell types. All published reports so far support the notion that different tissue types comprise varying metabolomes. Different biological roles of individual cell types support the further expectation of detecting striking differences on the low-level spatial resolution, e.g. between trichome and epidermis cells or between parenchyma and bundle sheath cells in plants[28]. Lastly, intracellular organization of metabolism is also highly structured into compartments, each of which serves specific functions which lead to large metabolic differences. For this reason, *in vivo* measurements are highly advantageous to study both the dynamics and subcellular localization of metabolites in real time, such as with genetically encoded fluorescent nanosensors[29]. Another report on subcellular studies of metabolites focused on metabolite profiling of isolated chloroplasts and subfractions including the envelope, the stroma and the thylakoids in a study on the activity of three 13-lipoxygenases under stress conditions[30]. So far, the integration of metabolomics data with proteomic or transcriptomic data has not gone beyond simple correlation analysis or statistical discrimination of phenotypes or treatment parameters. This use of data is inadequate to fulfill the vision of Systems Biology which aims at a comprehensive understanding and regulatory modeling of the complex interrelationships of cellular organisms[31], based on intensive computer simulations[32] followed by subsequent experimental testing of hypotheses derived from such models. In order to enable metabolomics (or proteomics!) to be a useful tool in such endeavor, metabolites need to be analyzed at high temporal and spatial resolution under carefully designed experiments in response to a range of genetic or environmental perturbations (not just plus/minus type of experiments such as 'sick vs. disease'). Today's analytical methods still seem to be inadequate with respect to acquiring the full complement of metabolites at ultimate sensitivity and for multiple biological snapshots. Instead of metabolomics approaches, hypothesis driven approaches seem currently to be more feasible for integrating gene, protein and metabolite levels.

5. Metabolic fluxes

The result of metabolomic analyses is a series of measurements of metabolite levels: snapshots of metabolism. Recently, attempts have been reported to use stable isotope incorporation for better quantification of cellular metabolism[33] whereas most researcher seem to utilize mature technology, such as GC/MS for high throughput profiling of steady state metabolite levels in yeast[34], algae[35] or plant cells[36] . However, such measurements just represent one side of the coin to study pathway structures. For example, with isotope labeled intermediates and isotopomer analysis, the different contributions of central carbon metabolic pathways can be unraveled for simple cell types[37]. Metabolic snapshot data, however, are usually not sufficient to directly derive enzyme activities and hierarchical structures of pathways, although metabolic changes caused by lack of enzyme activities are sometimes directly interpreted as alterations in metabolic fluxes[38]. Generally, changes of metabolite

levels may be due to drastically different causes: activities of membrane transport proteins may have been altered, rates of catabolism or anabolic reactions may have shifted, or branch point enzyme activities might have changed. Even if metabolite levels are found unchanged between different experimental situations, the underlying flux differences and enzymatic activities might still have changed. For example, if both anabolic and catabolic rates change in the same way and intensity, steady state levels of substrates and products involved in this reaction should not change although flux through the pathway would have clearly increased.

Therefore, metabolite snapshot data should be complemented by flux data, which has been proposed to be best accomplished by *in vivo* nuclear magnetic resonance measurements[39]. In principle, it should be also possible to derive enzyme activities and fluxes from snapshot measurements if we had the ability to measure true concentrations of all substrates and products in fast intervals, assuming we would know the total network structure. In practice, however, metabolomic methods miss important intermediates unless methods are tailored to meet these requirements, e.g. by using a range of different tools and technologies. Furthermore, even if snapshots were taken in time series, and even if all substrates and products of a pathway were covered, we are still unable to unravel the flux structure, i.e. the activity of reversible reactions, futile cycles or other back flows of products into the pathways via other routes through the metabolic network). Consequently, potential new side fluxes out of or into pathways by unforeseen additional enzymatic activities can only be detected by use of labeled compounds, either employing radioactive stable isotope tracers.

Unfortunately, these techniques are restricted in use by the need to feed in labeled substrates which (a) may not been taken up quickly enough into the cells and (b) are subsequently quickly diluted within the metabolic network. Therefore, only short distances or small parts of the total metabolic network can be imputed that have reasonably high metabolic turnover rates such as central C/N pathways in extremophiles[40] or that lead to and from strong carbon sinks such as starch in plant cells. A potential outcome of such a flux study is depicted in figure 4, with the compound 'A' used as labeled starting point from which relative fluxes A→ D and A→P→S can be followed. In this idealized map, the conclusion of such a flux study would be that the major flux of carbon is routed from A→ S via the H→O pathway but much less via the novel E→Y3 pathway. Significantly less carbon was partitioned here to the A→D pathway, and no net flux could be detected towards A→M in this artificial study result. Still, such a biochemical map would only include 35 compounds, which comprises a realistic and achievable aims in today's flux studies. A global view on all metabolic fluxes (a 'fluxome'[41]) is still out of reach by current techniques, even if fluxes are inferred from other metabolic sinks such as proteins. Using current methods, a combination of metabolomic snapshot data at a high number of biological replicates (to get the breadth of metabolic networks at high statistical significance levels) and flux measurements (on select and important pathways[42]) therefore seem to present the most practical solution to reach a more complete picture of metabolic control and regulation

6. Metabolic networks

The topology of metabolic networks, as being computed from genomic information, already comprises information about general systems properties[43]. Differences in metabolic levels can be used to interpret potential changes in pathway regulation using such known pathways. However, in general, it is not feasible to directly infer the global biochemical pathways structure or regulatory organization from omics data or even labeled intermediates and flux data. Still, metabolomics data can actually be used for gathering information that is complementary to differences in metabolite levels or patterns between genotypes or

treatments. Metabolomics workflows consist of one or many perturbations of a species with a number biological replicates per group according to the question and study design, and subsequently, data are compared by univariate and multivariate statistical means focusing on differences in average values between these statistical groups. However, one may ask if there actually is an average mouse, an average plant or even an average cell in a microculture? Since such a perfect average individual does not physically exist, how much biological interpretation and meaning can we generate out of the analysis of differences between averages of statistical groups?

What these statistical methods usually ignore is the within-group variance, or the individuality of every given sample, for example by applying Bayesian likelihood calculations[44], Pearson's correlations[45] or partial correlation analysis[46] to construct undirected dependency graphs from metabolomics data. This is a complementary way of looking at the data, which allows constructing a snapshot of a metabolic network for a given biological situation reflecting its underlying regulatory structure. Two theoretical papers have outlined the origin and meaning of metabolite : metabolite covariance and correlation[47,48], and some further reports have based biological conclusions from comparisons of metabolomic correlation networks[49,50]. The general outline of metabolic correlation networks is depicted in figure 5, from which it becomes clear that such networks hardly resemble the underlying biochemical network (figure 1). However, when comparing correlation networks under different conditions (e.g. environmental perturbation I and II, left and right panel of fig. 5), such networks can still be utilized in several ways: (1) novel metabolites Y2-Y9 can be mapped to other compounds for which the biological role and cellular compartment is supposed to be known, thus hypotheses about these novel compounds can be generated in an easier way compared to flux studies. (2) Network topology differences emphasize differences in overall regulation and carbon partitioning. For example, compound *A* lost its dominant hub role under the hypothetical biological situation II, whereas metabolite *B* becomes much more connected to other compounds. Such findings can best be interpreted if further biological data are added, such as specific enzyme or transport activities, or activation of transcription factors that would elicit a range of biological pathways and therefore impact biochemically unrelated metabolites in a similar direction and magnitude. More detailed analysis might also focus on the strength (the statistical power) of a given linear relationship, or the slope of the linear regression (which is equivalent to the ratios of pairs of metabolites). Biochemically, changes in metabolite ratios (such as ADP/ATP or Glu/Gln) can readily interpreted as important physiological parameters. Correlation network analysis may add an overview on regulation of metabolite pairs, hence potentially bridging the analysis of metabolic snapshots ('steady state levels') to flux data by detailing the relative partitioning of metabolite pools under different biological conditions.

There are further theoretical considerations that support this notion of utilizing within-group variance as surrogate for the actual dynamics of the intracellular metabolic network. All biological systems share network properties which are called 'robustness and flexibility'. Cells are hit in short time intervals by stochastic factors such as influx deviations of external transport metabolites, intensity differences in environmental parameters or subtle physical interferences such as 'wind'. Metabolic systems would become very unstable if each of these short-term pulses would be taken up in immediate responses. There are a number of regulatory steps that inhibit metabolic overreactions but instead introduce response lag times by using threshold systems, active transport steps, or reversibility of reactions. In total, these delay steps render the system to become 'robust' which is an important property to maintain the system at a given steady state. Complementary to such robust regulation of the network structure is the necessity to quickly alter metabolite levels depending on certain stress conditions or developmental needs. The responsible general system property is called

'flexibility'. System flexibility is a prerequisite of the capability to 'control' or alter defined steady states without affecting other parts of the system, depending on external or internal stimuli. Any system needs capabilities to react in a fast and coordinated manner on immediate needs and threats, even if the triggering signals for such needs are of low abundant and transient nature. Examples might be heat shock, wounding responses, or herbivore attacks, among others. Very fruitful research has been reported on a combination of a calculation of the metabolic feasibility space of prokaryotes and confirmation of predictions of system responses using flux data[51]. One further step comprises use of metabolite concentrations to inform the feasible parameter space of enzyme kinetics in yeast[52]. It is this kind of model-based computation that needs further refinements and application for predicting metabolic responses in complex eukaryote systems and higher organisms.

It is interesting to note that there is a difference in terminology between theory of metabolism and molecular biology. David Fell has pointed out in his famous book "Control of metabolism" (1997)[53], that the terms 'control' and 'regulation' point to biochemical properties that are rather different in their respective meanings. Regulation is the ability of a complex system to maintain its basic properties (e.g. metabolite levels) independent of external factors that continuously try to push the system out of balance, whereas control was defined as a system property that enables changes between different states of a system. In this respect, terminology of metabolic theory reflects the understanding of network properties (flexibility and robustness), whereas in molecular biology, regulation and control are used as synonyms describing only the ability to change a system, but not how to maintain it. Examples of 'control' are found in classic physiology. In terms of plant physiology, cold acclimation (by increased values in carbohydrates) or leaf senescence (altered ratios of catabolism versus anabolism) are examples of 'control' or 'system flexibility', whereas the tendency to keep metabolic fluxes in a narrow range under a given set of environmental parameters (the steady state) is an example for metabolic 'regulation' or system 'robustness'.

Apart from kinetics and flux rates, further properties of metabolic networks are the stoichiometric structure which may be used to define metabolic feasibility spaces for cellular growth, and connectivity[54,55], which define the relative importance of metabolites as branching points to allow redirection and partitioning of the ratio of carbon, nitrogen, phosphor and sulfur between pathways, organs and compartments. Interestingly, metabolic correlation networks seem to reflect the different needs and partitioning between pathways and thus may enable bridging information that is derived from steady state levels and from flux information. However, so far information garnered from theoretical or experimental metabolic networks has not enabled probing biochemical pathway structure with the aim at detecting novel metabolic routes. At least, such work seems to be more feasible within the foreseeable future than meaningful integration with data from other omics approaches.

Acknowledgments

Continuing discussions with Wolfram Weckwerth (MPI-MP, Potsdam, Germany) and Tobias Kind (UC Davis) have been helpful to elaborate these considerations.

References

[1] Fiehn O (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genom*. 2, 155-168

[2] Forster J, Gombert AK, Nielsen, J (2002) A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol. Bioeng.* 79, 703-712

[3] Levin I, Lalazar A, Bar M, Schaffer AA (2004) Non GMO fruit factories strategies for modulating metabolic pathways in the tomato fruit. *Ind. Crops Prod.* 20, 29-36
[4] Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M, Nishioka T (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* 2, 488-494
[5] Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M, Nishioka T (2002) Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.* 74, 2233-2239
[6] Sato S, Soga T, Nishioka T, Tomita M (2004) Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant J.* 40, 151-163
[7] Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp, PD (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* 6: Art. No. R2 2005
[8] Nicholson JK, Holmes E, Lindon JC, Wilson ID (2004) The challenges of modeling mammalian biocomplexity. *Nat. Biotechnol.* 22, 1268-1274
[9] Mayeno AN, Yang RSH, Reisfeld B (2005) Biochemical reaction network modeling: Predicting metabolism of organic chemical mixtures. *Environmental Sci. Technol.* 39, 5363-5371
[10] Milman BL (2005) Identification of chemical compounds. *Trends Anal. Chem.* 24, 493-508
[11] Nicolaou KC, Snyder SA (2005) Chasing molecules that were never there: Misassigned natural products and the role of chemical synthesis in modern structure elucidation. *Angew. Chem. Int. Ed.* 44, 1012-1044 2005
[12] Pubchem project. URL cited April 30, 2006 [http://pubchem.ncbi.nlm.nih.gov/]
[13] Feldman HJ, Snyder KA, Ticoll A, Pintilie G, Hogue CWV (2006) A complete small molecule dataset from the protein data bank. *FEBS Lett.* 580, 1649-1653
[14] Murray-Rust P, Rzepa HS, Stewart JJP, Zhang Y (2005) A global resource for computational chemistry. *J. Mol. Model.* 11, 532-541
[15] Metabolomics Standards Initiative. URL cited April 30, 2006 [http://metabolomicssociety.org/mstandards.html]
[16] Steinbeck C (2004) Recent developments in automated structure elucidation of natural products. *Nat. Prod. Rep.* 21, 512-518
[17] Kind T, Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* 7:234
[18] Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* 72, 3573-3580
[19] Fiehn O, Major H (2005) Exact Molecular Mass Determination of Polar Plant Metabolites Using GCT with Chemical Ionization. *Waters application note*, URL cited Jan. 08, 2006 [http://www.waters.com/WatersDivision/SiteSearch/AppLibDetails.asp?LibNum=720001260EN]
[20] Tolstikov VV, Fiehn O (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal Biochem.* 301, 298-307
[21] BioCyc. URL cited April 30, 2006 [http://biocyc.org/]
[22] Villas-Boas SG, Akesson M, Nielsen J (2005) Biosynthesis of glyoxylate from glycine in Saccharomyces cerevisiae. *FEMS Yeast Res.* 5, 703-709
[23] Lange BM, Ghassemian M (2005) Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. *Phytochemistry* 66, 413-451
[24] Griffin JL, Bonney SA, Mann C, Hebbachi AM, Gibbons GF, Nicholson JK, Shoulders CC, Scott J (2004) An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver. *Physiol. Genomics* 17, 140-149
[25] Kant MR, Ament K, Sabelis MW, Haring MA, Schuurink RC (2004) Differential timing of spidmite-induced direct and indirect defenses in tomato plants. *Plant Physiol.* 135, 483-495
[26] Buchholz A, Hurlebaus J, Wandrey C, Takors R (2002) Metabolomics: quantification of intracellular metabolite dynamics. *Biomol. Eng.* 19, 5-15
[27] Lafaye A, Junot C, Pereira Y, Lagniel G, Tabet JC, Ezan E, Labarre J (2005) Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *J. Biol. Chem.* 280, 24723-24730
[28] Schad M, Mungur R, Fiehn O, Kehr J (2005) Metabolic profiling of laser microdissected vascular bundles of Arabidopsis thaliana. Plant Methods 1:2
[29] Deuschle K, Fehr M, Hilpert M, Lager I, Lalonde S, Looger LL, Okumoto S, Persson J, Schmidt A, Frommer WB (2005) Genetically encoded sensors for metabolites. *Cytometry A* 64A, 3-9
[30] Bachmann A, Hause B, Maucher H, Garbe E, Voros K, Weichert H, Wasternack C, Feussner I (2002) Jasmonate-induced lipid peroxidation in barley leaves initiated by distinct 13-LOX forms of chloroplasts. *Biological Chemistry* 383, 1645-1657

[31] Yang YT, Engin L, Wurtele ES, Cruz-Neira C, Dickerson JA (2005) Integration of metabolic networks and gene expression in virtual reality. *Bioinformatics* 21, 3645-3650
[32] Ishii N, Robert M, Nakayama Y, Kanai A, Tomita M (2004) Toward large-scale modeling of the microbial cell for computer simulation. *J. Biotechnology* 113, 281-294
[33] Kim JK, Harada K, Bamba T, Fukusaki E, Kobayashi A (2005) Stable isotope dilution-based accurate comparative quantification of nitrogen-containing metabolites in Arabidopsis thaliana T87 cells using in vivo N-15-isotope enrichment. *Biosci. Biotechnol. Biochem.* 69, 1331-1340
[34] Villas-Boas SG, Moxley JF, Akesson M, Stephanopoulos G, Nielsen J (2005) High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem. J.* 388, 669-677
[35] Boelling C, Fiehn O (2005) Metabolite profiling of Chlamydomonas reinhardtii under nutrient deprivation. *Plant Physiol.* 139, 1995-2005
[36] Broeckling CD, Huhman DV, Farag MA, Smith JT, May GD, Mendes P, Dixon RA, Sumner LW (2005) Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* 56: 323-336
[37] Marin S, Chiang K, Bassilian S, Lee WNP, Boros LG, Fernandez-Novell JM, Centelles JJ, Medrano A, Rodriguez-Gil JE, Cascante M (2003) Metabolic strategy of boar spermatozoa revealed by a metabolomic characterization. *FEBS Lett.* 554, 342-346
[38] Watkins SM, Zhu XN, Zeisel SH (2003) Phosphatidylethanolamine-N-methyltransferaactivity and dietary choline regulate liver-plasma lipid flux and essential fatty acid metabolism in mice. *J. Nutrition* 133, 3386-3391
[39] Mesnard F, Ratcliffe RG (2005) NMR analysis of plant nitrogen metabolism. *Photosynth. Res.* 83,163-180
[40] Maskow T, Kleinsteuber S (2004) Carbon and energy fluxes during haloadaptation of Halomonas sp EF11 growing on phenol. *Extremophiles* 8, 133-141
[41] Zamboni N, Sauer U (2004) Model-independent fluxome profiling from H-2 and C-13 experiments for metabolic variant discrimination. *Genome Biology* 5(12), Art. No. R99
[42] Fernie AR, Geigenberger P, Stitt M. (2005) Flux an important, but neglected, component of functional genomics. *Curr. Opin.Plant Biol.* 8, 174-182
[43] Giuliani A, Zbilut JP, Conti F, Manetti C, Miccheli A (2004) Invariant features of metabolic networks: a data analysis application on scaling properties of biochemical pathways. *Physica A – Stat. Mech. Appl.* 337, 157-170
[44] Likelynet – software for exploring linear relationships in multidimensional data sets. URL cited April 30, 2006 [www.likelynet.com]
[45] Kose F, Weckwerth W, Linke T, Fiehn O (2001) Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics* 17, 1198-1208
[46] de la Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 2004, 20, 3565-3574
[47] Steuer R, Kurth J, Fiehn O, Weckwerth W (2003) Observing and interpreting correlations in metabolic networks. *Bioinformatics* 19, 1019-1026
[48] Camacho D, de la Fuente A, Mendes P (2005). The origin of correlations in metabolomics data. *Metabolomics* 1, 53-63
[49] Fiehn O (2003) Metabolic networks of Cucurbita maxima phloem. *Phytochemistry* 62, 875-886
[50] Weckwerth W, Loureiro ME, Wenzel K, Fiehn O (2004) Metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA* 101, 7809-7814
[51] Wiback SJ, Mahadevan R, Palsson BO (2004) Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: The Escherichia coli spectrum. *Biotechnol. Bioeng*. 86, 317-331
[52] Famili I, Mahadevan R, Palsson BO (2005) k-cone analysis: Determining all candidate values for kinetic parameters on a network scale. *Biophys. J.* 88, 1616-1625
[53] Fell D (1997) Understanding the Control of Metabolism. Portland Press, London.
[54] Duarte NC¸ Palsson BO, Fu PC (2004) Integrated analysis of metabolic phenotypes in Saccharomyces cerevisiae. *BMC Genomics* 5: Art. No. 63
[55] Dandekar T, Moldenhauer F, Bulik S, Bertram H, Schuster S (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems* 70, 255-270

Figure legends

Fig. 1 Proposed workflow for a rapid annotation of unknown analytical signals in LC/MS or GC/MS metabolite profiles. Such schema might evolve as a general database query tool similar to MASCOT or SEQUEST in proteomics research, returning a list of tentative structures with an overall similarity score.

Fig. 2 Left panel: Generalized metabolic pathway map (metabolites A-X) that may result from genomic reconstructions. Right panel: Result of a metabolite profiling study using a specific analytical technique. Six metabolites could not be detected (open squares) although these were supposed to be present by the reconstructed pathway map. In addition, nine novel metabolites were detected (Y1-Y9) which cannot immediately be mapped onto the pathway chart.

Fig. 3 Functional annotation of novel metabolites onto biochemical pathways depicted in fig. 2. Using a variety of genomic, analytical and biochemical experiments, pathways may be unraveled for some of the new compounds. Such pathway elucidations involve laborious wet laboratory work and thus leave many other uncommon metabolites without biochemical annotation.

Fig. 4 Potential outcome of a flux study using the labeled metabolite ‘A’ from the biochemical pathway depicted in fig. 2. Differences in use of alternative biochemical pathways result from flux studies, enzymatic activities can be calculated and the involvement of novel metabolites (Y3) is confirmed. Usually, flux data are not obtained for more distant pathways, or pathways with low overall metabolic turnover.

Fig. 5 Metabolic network graphs resulting from correlation or linearity analysis of metabolite pairs under two different conditions (left and right panel). Some correlations will reflect the underlying biochemical pathway structure depicted in figure 2, whereas other correlations refer to differences in overall metabolic regulation (e.g. by activation of transcription factors). Often, such network graphs enable generating improved hypotheses on the biological roles of pathways and known and novel metabolites (Y2-Y9).

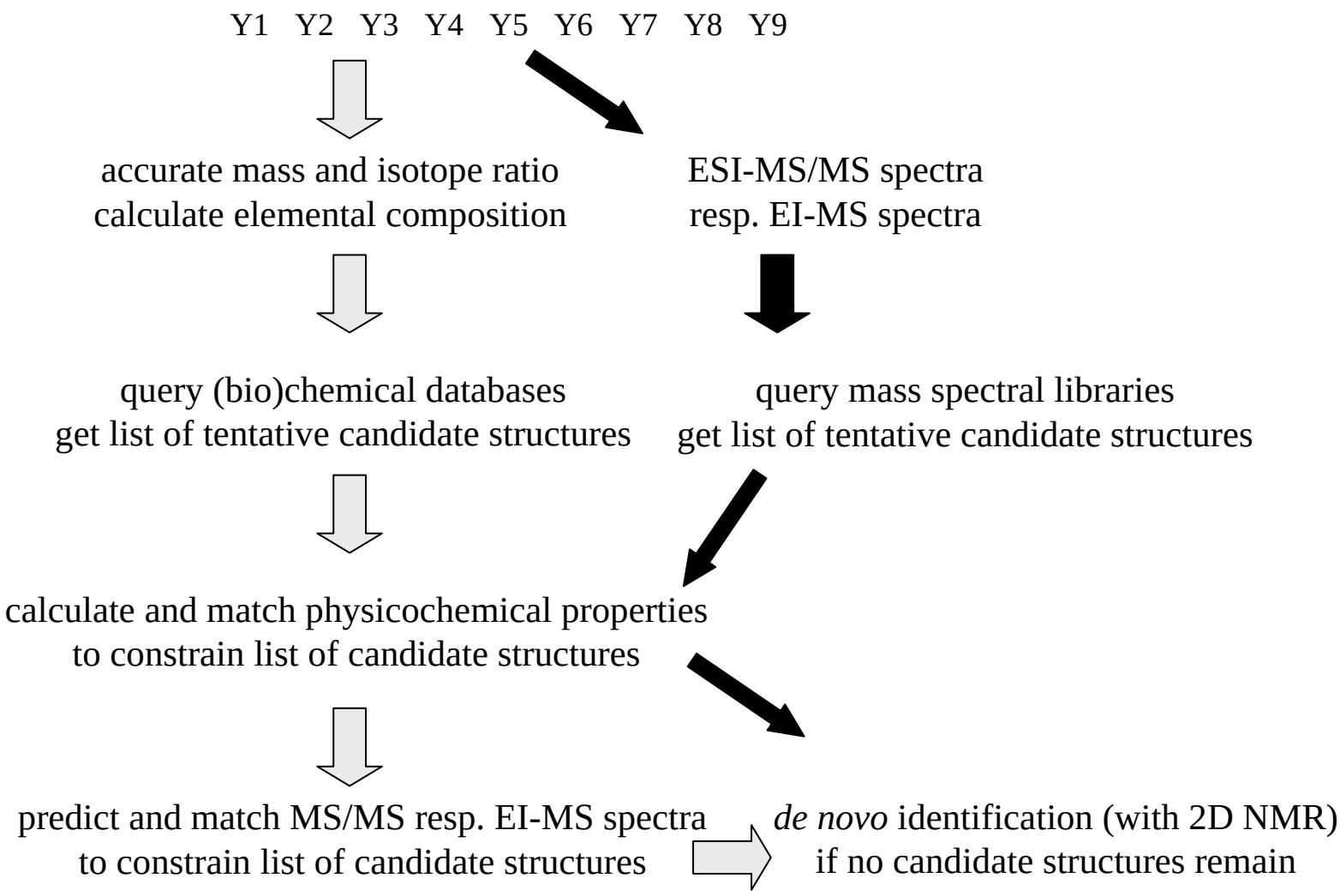
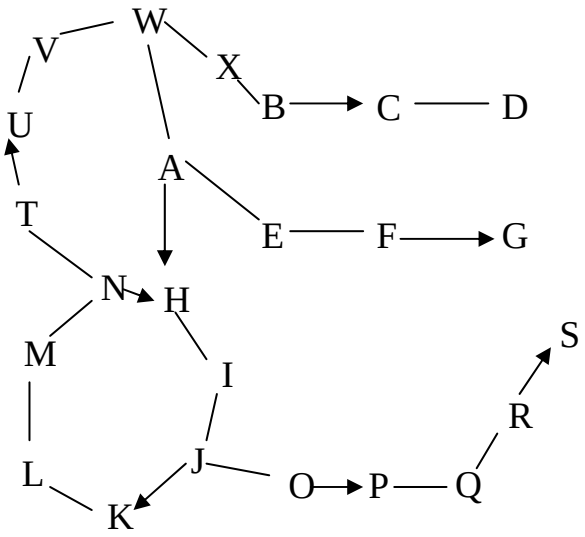
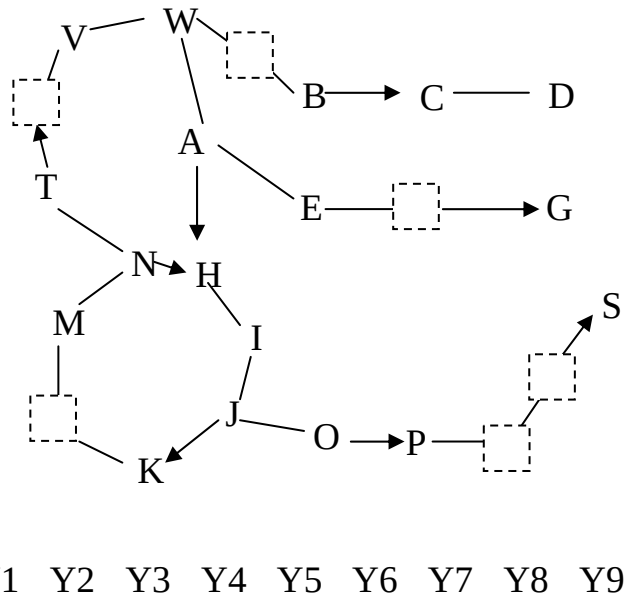

Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9
accurate mass and isotope ratio
calculate elemental composition
ESI-MS/MS spectra
resp. EI-MS spectra
query (bio)chemical databases
get list of tentative candidate structures
query mass spectral libraries
get list of tentative candidate structures
calculate and match physicochemical properties
to constrain list of candidate structures
predict and match MS/MS resp. EI-MS spectra
to constrain list of candidate structures
de novo identification (with 2D NMR)
if no candidate structures remain

Fig. 1





Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9

Fig. 2

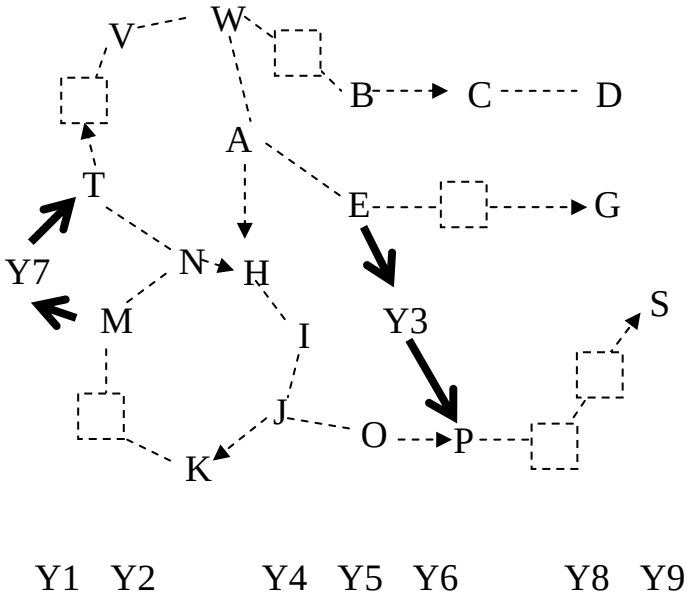


Fig. 3

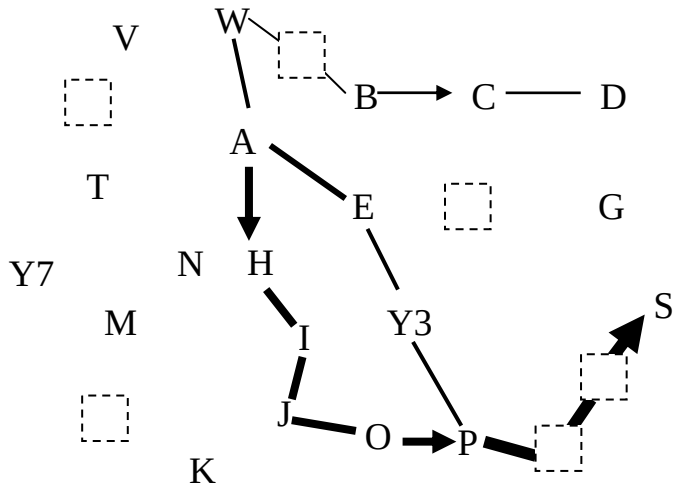
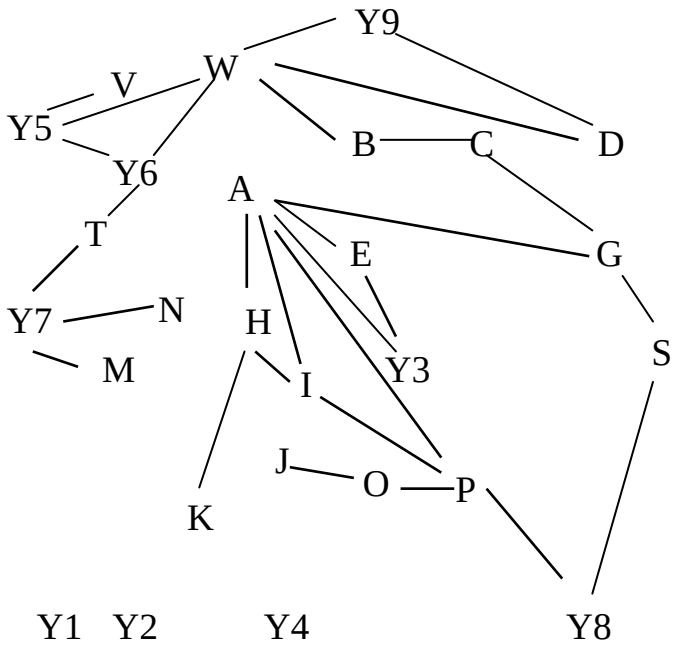


Fig. 4


Y9
W
V
Y5
Y6
B
C
D
A
T
E
G
Y7
N
H
S
M
I
Y3
J
O
P
K
Y1
Y2
Y4
Y8

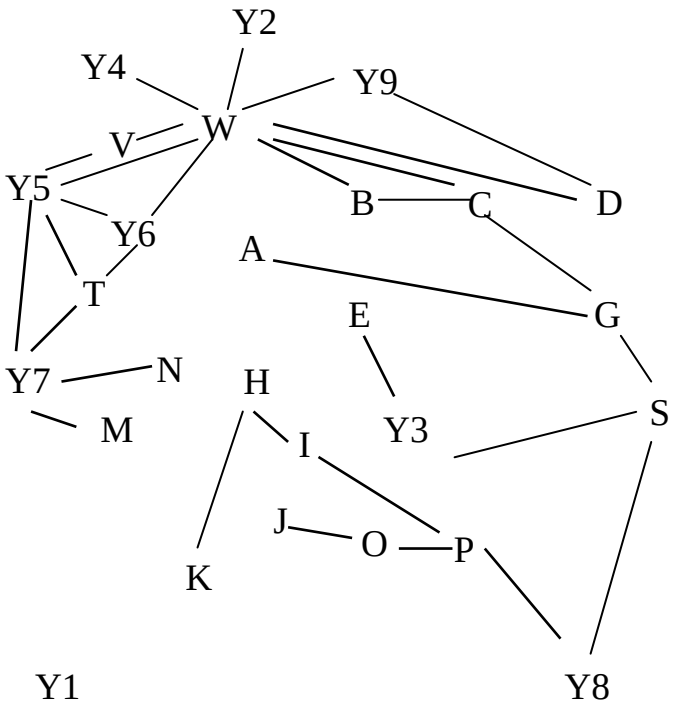

Y2
Y4
Y9
W
V
Y5
B
C
D
Y6
A
T
G
E
Y7
N
H
S
M
Y3
I
J
O
P
K
Y1
Y8

Fig. 5