Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data

Rainer Breitling,[a,b,*] Shawn Ritchie,[c] Dayan Goodenowe,[c] Mhairi L. Stewart,[b] and Michael P. Barrett[b]

[a]*Groningen Bioinformatics Centre, University of Groningen, 9751 NN, Haren, The Netherlands*
[b]*Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK*
[c]*Phenomenome Discoveries, Saskatoon, S7N 4L8, Canada*

Received 1 February 2006; accepted 19 May 2006

Fourier transform mass spectrometry has recently been introduced into the field of metabolomics as a technique that enables the mass separation of complex mixtures at very high resolution and with ultra high mass accuracy. Here we show that this enhanced mass accuracy can be exploited to predict large metabolic networks *ab initio*, based only on the observed metabolites without recourse to predictions based on the literature. The resulting networks are highly information-rich and clearly non-random. They can be used to infer the chemical identity of metabolites and to obtain a global picture of the structure of cellular metabolic networks. This represents the first reconstruction of metabolic networks based on unbiased metabolomic data and offers a breakthrough in the systems-wide analysis of cellular metabolism.

KEY WORDS: Fourier transform mass spectrometry; metabolic networks; network reconstruction; computational methods.

1. Introduction

The biological interpretation of post-genomic datasets depends on the ability to identify reliably the molecules that have been measured. For example, microarray analysis depends upon the hybridization behavior of complementary nucleotide sequences to enable detection of individual mRNA transcripts. For metabolomic analysis comparable means of simple identification have not yet been described; the chemical characterization of single molecular species being laborious and not currently amenable to automation. This has, to date, restricted the application of metabolomics largely to fingerprinting studies that detect diagnostic differences between samples but provide restricted biological insight (Allen *et al*., 2003; Goodacre *et al*., 2004; Kell, 2004; Nicholson *et al*., 2004).

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS or simply FTMS) (Brown *et al*., 2005; Zhang *et al*., 2005) has so far been used only in a handful of published studies into metabolomics (Aharoni *et al*., 2002; Hirai *et al*., 2004; Murch *et al*., 2004; Tohge *et al*., 2005). However, the technique has great potential as a technology to unravel metabolomes. The extreme mass accuracy of the technique, coupled to ultra high resolution of mass species means that thousands of metabolites can be identified simultaneously without the need for chromatographic separations (Brown *et al*., 2005; Zhang *et al*., 2005). The ultra high mass accuracy enables assignment of putative chemical formulae to metabolites since only a finite number of combinations of carbon, nitrogen, oxygen, hydrogen, sulfur and phosphorus can yield the same precise measured mass.

This capability to assign likely chemical formulae to a multitude of metabolites may allow the analysis of metabolomes at a level comparable to microarray analysis of the transcriptome. However, as we demonstrate in this manuscript, as masses of metabolites increase the ability to assign individual chemical formulae diminishes. However, judicious analysis of the data from a given metabolomics experiment, can go some way to resolving this problem. This is because, in addition to offering putative identification of formulae, ultra high mass accuracy has the potential to identify the connectivity between related metabolites, since chemically transformed species will be related by measurable, clearly defined mass differences. In the present study we are able to position observed molecules uniquely and accurately in comprehensive metabolic networks that are generated *ab initio* from the measured mass peaks. Links in these networks correspond explicitly to actual chemical reactions and thus go beyond the metabolite correlation networks used previously (Steuer *et al*., 2003). We show that the *ab initio* metabolic networks have a highly informative, non-random structure and can be used to assign putative molecular identities to metabolites. Moreover, they open up a novel perspective on the global structure of cellular metabolism by providing the first comprehensive experimental assessment of metabolite connectivities, unbiased by the historical contingencies of classical biochemistry. The topology of metabolic networks described to date has been inferred based upon reactions predicted to occur within a given cell type based

[*] To whom correspondence should be addressed.
E-mail: r.breitling@rug.nl

upon reactions that may be catalyzed by enzymes whose presence is predicted through genome analysis (Arita, 2004; Ma and Zeng, 2003a, b; Pfeiffer *et al.*, 2005). The networks generated to date fail to take into account the fact that enzymes need not be expressed constitutively, nor compartmentalized in a manner allowing them to contribute to a given sub-network. Moreover, roles of non-enzymatic metabolite interconversions within the cell are not accounted for, nor are roles for enzymes with promiscuous substrate specificity. In spite of these limitations, metabolic network building is an important discipline and one that will benefit greatly from the introduction of techniques that can directly measure the metabolites present within a cell and report upon their connectivity. The work that we present here demonstrates the potential of using ultra high resolution mass spectrometry to generate such networks *ab initio*.

Another important implication of the present work is the ability to construct metabolic networks for organisms that have so far been outside the focus of classical biochemistry, i.e. beyond yeast and *E. coli*. Indeed, genomic information of any kind is not required for these analyses. Construction of such *ab initio* networks will form a useful basis for future system-wide comparative studies of metabolism in a wide variety of species.

2. Materials and methods

2.1. Chemicals and standards

ATP, ADP, NAD, NADP and diminazene aceturate (berenil) were of the highest grade available from Sigma. Trypanothione (N1,N8 bis-glutathionylspermidine) was from G.H. Coombs (University of Glasgow). DB75 (2,5-bis(4-amidinophenyl)furan) was from D. Boykin (Georgia State University). Pentamidine isethionate was from Aventis (through the World Health Organisation). Cymelarsan (melarsen oxide in solution) was from M. Turner (University of Glasgow).

2.2. Preparation of trypanosome extracts

Bloodstream form trypanosomes (Lister 427 line) were collected by cardiac puncture at 5×10^8 parasites/ml from Wistar rats and separated from red blood cells as a buffy coat by centrifugation at 3000*g*. The same parasite line was grown in HMI-9 medium supplemented with 20% foetal calf serum to mid-log phase (8×10^5 parasites/ml) of culture medium. Parasites were then centrifuged at 3000*g* for 5 min with pellets and supernatant then flash frozen in liquid nitrogen prior to extraction.

2.3. Mass spectrometric analysis

Fourier transform mass spectrometry was performed as we have described previously (Tohge *et al.*, 2005). Briefly, cell pellets and media (300 μL) were extracted in solvents ranging from polar (aqueous) to non-polar with most proteins and nucleic acids removed during extraction. Extracts were stored at −80 °C. After appropriate dilution, samples were analysed on a Bruker Daltonics APEX III Fourier transform ion cyclotron resonance mass spectrometer equipped with a 7.0 T superconducting magnet (Bruker Daltonics, Billerica, MA). Samples were directly injected using electrospray ionization (ESI) and atmospheric pressure chemical ionization (APCI) at a flow rate of 600 μL per hour. Different sample extracts were analyzed separately, and the processed mass spectral data for each sample were combined. Sample peaks were calibrated using internal standards with peak mass error < 1 ppm relative to the theoretical mass. Measured masses were combined into a single table for exploration using *DISCOVAmetrics*™ software.

2.4. Computational analysis

Further analyses used a combination of Microsoft Excel, MATLAB and custom-written Perl scripts, which are available from the authors upon request. The degree distributions in figure 1A were calculated from the *ab initio* networks by counting how often a particular ‘‘commonly observed’’ mass difference or a certain ‘‘biochemical reaction’’ mass difference occurred in the network. The mass differences were then ranked by their number of occurrences and plotted in that order. The degrees in figure 1B were obtained by counting the number of mass pairs each observed mass was involved in, i.e. its number of edges in the undirected *ab initio* networks.

3. Results and discussion

3.1. Mass precision and resolution of FTMS

We started our study by analyzing a mixture of standard chemicals, including ATP, ADP, NAD, NADP, glutathione and a number of trypanocidal drugs important to our research. For the 13 standards that were detected in our sample, the average mass accuracy was 0.783 ppm (maximum 2.47 ppm; table 1). As we detect masses between 100 and 1500 atomic mass units, this resolution is sufficient to discriminate at least 50,000 molecular species, even when we assume that many of them will be represented by several peaks (isotope peaks and ion adducts). Indeed, for most of our standard molecules, single and double ^{13}C peaks were detected in the expected proportions, as were a large number of minor contaminants, confirming the high sensitivity of the method.

3.2. FTMS analysis of a parasitic protozoan, *Trypanosoma brucei*

We then proceeded with the analysis of metabolite samples from *Trypanosoma brucei*, a protozoan parasite

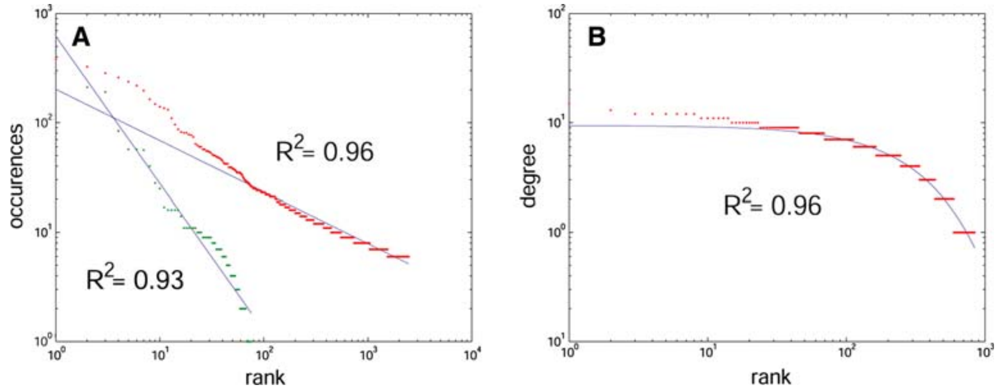

Figure 1. Zipf plots of the reaction (A) and metabolite (B) degrees in the *ab initio* metabolic network of *Trypanosoma brucei*. In (A), the red dots correspond to the reactions inferred automatically from the mass–mass differences, while the green dots are based on a pre-defined set of common biochemical reactions. The fitted lines are based on power-law (A) and exponential distributions (B), respectively.

Table 1
Molecular standards detected by FTMS

Compound	Predicted mass	Measured mass	ppm	Average S/N
Glutathione	307.083807	307.0835	1	438
Oxidized glutathione	612.152	612.1516	1	328
Trypanothione	723.3044	723.3036	1	16
Oxidized trypanothione	721.2887	721.2889	0	281
NADP	743.075458	743.0766	2	442
NAD	663.109125	663.1096	1	1229
ATP	506.99575	506.9945	2	289
ADP	427.029418	427.0293	0	118
AMP	347.063086	347.0633	1	14
Berenil	281.138894	281.139	0	9
Pentamidine	340.1899	340.1897	1	67
DB75	304.132411	304.1325	0	115
Melarsen oxide	292.00538	292.0053	0	113

Several of the standards were detected in multiple ionization modes. Only the measured mass with the strongest signal is listed in these cases. Average S/N is the average signal-to-noise ratio over all ionization modes that gave a detectable signal.

that causes the fatal disease sleeping sickness in Africa (Barrett *et al*., 2003). We collected metabolic profiles for parasites grown *in vivo* (in rats) and *in vitro* (in serum culture) and compared these to the profiles from their environment (rat serum and culture medium, respectively). As a parasite, *T. brucei* has a drastically streamlined set of metabolic enzymes, making it particularly suitable for pioneering studies in metabolomics. At the same time, metabolic enzymes represent key targets for drugs used in treating sleeping sickness, and new targets are urgently required (Butler, 2005).

Excluding ^{13}C isotope peaks and common ion adducts, a total of 399 masses were identified from rat-derived trypanosomes, while for *in vitro* grown cells the total number was 262. Of these, about 30% could be matched to putative identities in the chemical database PubChem (http://www.pubchem.ncbi.nlm.nih.gov/). These matches offer reasonable certainty regarding the empirical formula (matches to two alternative empirical formulae within 2 ppm are rare) although mass alone does not allow discrimination of chemical connectivity of atoms within the molecules.

3.3. Generation of ab initio metabolic networks

A majority of masses in our sample did not match to any known compound. This is due to the prevailing lack of knowledge about the total complexity of the metabolome of *T. brucei* and most other biological species, rather than limitations of mass accuracy. We have systematically explored the accuracy requirements needed for database matching and found that about 1–2 ppm is sufficient for a unique hit in PubChem, which in the release used contained 72,634 unique empirical formulae (table 2). However, making use of the high mass accuracy of Fourier transform mass spectrometry, and particularly studying mass–mass differences, has allowed us here to make significant progress in surmounting this problem. Two approaches were used to generate *ab initio* reconstructions of metabolic networks from the available data:

(1) In a completely untargeted approach, all pairwise mass–mass differences were calculated. Considering all possible pairwise differences makes the approach unbiased, although not all molecules are related in a chemically feasible way. Thus, in a next step frequently occurring mass differences were identified (defined as clusters of more than five pairwise distances that differed by less than 0.0001 mass units). Such commonly observed distances are very unlikely to be observed by chance and can hence be expected to have a chemical basis. Compounds whose masses differed by one of these commonly observed masses were assumed to be related by a chemical transformation.
(2) In the first approach above we have focused on all measured relationships within the dataset. This totally unbiased approach will not distinguish between metabolic transformations within the cell and

Table 2
Average number of matching empirical formulae identified in PubChem at various mass accuracies, averaging over all masses in the present release of the database (mass range 2–9200, median mass 438)

ppm	0.01	0.02	0.05	0.08	0.1	0.2	0.5	0.8	1	2	5	10	20
Hits	1.00	1.00	1.00	1.00	1.00	1.00	1.03	1.06	1.09	1.32	2.03	3.40	6.08

chemical alterations induced during sample preparation and analysis by FTMS. As the full inventory of FTMS based artifacts becomes clear it will become possible to filter data for as many of these as possible. To work towards this approach we adopted a second, semi-targeted approach in which the observed pairwise mass–mass differences were compared to a list of 83 mass differences corresponding to common metabolic reactions compiled from biochemistry textbooks (table 3). Here, metabolites whose mass differed by the expected amount (within 2 ppm) were considered to be related by the corresponding metabolic transformation.

This approach is related to the technique used for the *de novo* sequencing of proteins using tandem mass spectrometry (in which masses are related by the masses of individual amino acids). The task in proteomics is facilitated by the linear nature of the examined

Table 3
Common metabolic transformations and corresponding formulae

Alanine	C_3H_5NO	Guanosine 5-diphosphate (–H_2O)	$C_{10}H_{13}N_5O_{10}P_2$
Arginine	$C_6H_{12}N_4O$	Guanosine 5-monophosphate (–H_2O)	$C_{10}H_{12}N_5O_7P$
Asparagine	$C_4H_6N_2O_2$	Guanine (–H)	$C_5H_4N_5O$
Aspartic acid	$C_4H_5NO_3$	Guanosine (–H_2O)	$C_{10}H_{11}N_5O_4$
Cysteine	C_3H_5NOS	Deoxythymidine 5′-diphosphate (–H_2O)	$C_{10}H_{14}N_2O_{10}P_2$
Cystine	$C_6H_{10}N_2O_3S_2$	Thymidine (–H_2O)	$C_{10}H_{12}N_2O_4$
Glutamic acid	$C_5H_7NO_3$	Thymine (–H)	$C_5H_5N_2O_2$
Glutamine	$C_5H_8N_2O_2$	Thymidine 5′-monophosphate (–H_2O)	$C_{10}H_{13}N_2O_7P$
Glycine	C_2H_3NO	Uridine 5′-diphosphate (–H_2O)	$C_9H_{12}N_2O_{11}P_2$
Histidine	$C_6H_7N_3O$	Uridine 5′-monophosphate (–H_2O)	$C_9H_{11}N_2O_8P$
Isoleucine	$C_6H_{11}NO$	Uracil (–H)	$C_4H_3N_2O_2$
Leucine	$C_6H_{11}NO$	Uridine (–H_2O)	$C_9H_10N_2O_5$
Lysine	$C_6H_{12}N_2O$	Acetylation (–H)	$C_2H_3O_2$
Methionine	C_5H_9NOS	Acetylation (–H_2O)	C_2H_2O
Phenylalanine	C_9H_9NO	C_2H_2	C_2H_2
Proline	C_5H_7NO	Carboxylation	CO_2
Serine	$C_3H_5NO_2$	CHO_2	CHO_2
Threonine	$C_4H_7NO_2$	Condensation/dehydration	H_2O
Tryptophan	$C_{11}H_{10}N_2O$	Diphosphate	$H_3O_6P_2$
Tyrosine	$C_9H_9NO_2$	Ethyl addition (–H_2O)	C_2H_4
Valine	C_5H_9NO	Formic Acid (–H_2O)	CO
Acetotacetate (–H_2O)	$C_4H_4O_2$	Glyoxylate (–H_2O)	C_2O_2
Acetone (–H)	C_3H_5O	Hydrogenation/dehydrogenation	H_2
Adenylate (–H_2O)	$C_{10}H_{12}N_5O_6P$	Hydroxylation (–H)	O
Biotinyl (–H)	$C_{10}H_{15}N_2O_3S$	Inorganic phosphate	P
Biotinyl (–H_2O)	$C_{10}H_{14}N_2O_2S$	Ketol group (–H_2O)	C_2H_2O
Carbamoyl P transfer (–H_2PO_4)	CH_2ON	Methanol (–H_2O)	CH_2
Co-enzyme A (–H)	$C_{21}H_{34}N_7O_{16}P_3S$	Phosphate	HPO_3
Co-enzyme A (–H_2O)	$C_{21}H_{33}N_7O_{15}P_3S$	Primary amine	NH_2
Glutathione (–H_2O)	$C_{10}H_{15}N_3O_5S$	Pyrophosphate	PP
Isoprene addition (–H)	C_5H_7	Secondary amine	NH
Malonyl group (–H_2O)	$C_3H_2O_3$	Sulfate (–H_2O)	SO_3
Palmitoylation (–H_2O)	$C_{16}H_{30}O$	Tertiary amine	N
Pyridoxal phosphate (–H_2O)	$C_8H_8NO_5P$	$C_6H_{10}O_5$	$C_6H_{10}O_5$
Urea addition (–H)	CH_3N_2O	$C_6H_{10}O_6$	$C_6H_{10}O_6$
Adenine (–H)	$C_5H_4N_5$	D-ribose (–H_2O) (ribosylation)	$C_5H_8O_4$
Adenosine (–H_2O)	$C_{10}H_{11}N_5O_3$	Disaccharide (–H_2O)	$C_{12}H_{20}O_{11}$
Adenosine 5′-diphosphate (–H_2O)	$C_{10}H_{13}N_5O_9P_2$	Glucose-N-phosphate (–H_2O)	$C_6H_{11}O_8P$
Adenosine 5′-monophosphate (–H_2O)	$C_{10}H_{12}N_5O_6P$	Glucuronic acid (–H_2O)	$C_6H_8O_6$
Cytidine 5′-diphosphate (–H_2O)	$C_9H_{13}N_3O_{10}P_2$	Monosaccharide (–H_2O)	$C_6H_{10}O_5$
Cytidine 5′-monophsophate (–H_2O)	$C_9H_{12}N_3O_7P$	Trisaccharide (–H_2O)	$C_{18}H_{30}O_{15}$
Cytosine (–H)	$C_4H_4N_3O$		

peptides and the well-defined set of possible building blocks, however the same principle, but using a more extensive set of transformation masses, should be informative in mass spectrometry as applied to metabolomics.

Table 4 summarizes the results of the first approach. About 25,370 mass–mass differences corresponded to one of 2472 "commonly occurring" mass differences. This is a dramatic excess over the number observed for random lists of masses (uniformly distributed between 100 and 1500 atomic mass units) shown in the same table. Thus, there are an astonishing 25,000 or more relationships between observed masses that can be explained by *ab initio* predicted chemical transformations. The most common of them are listed in table 4 and assigned to the most likely underlying chemical difference (including isotope variability). It is clear that not all of these enriched mass–mass differences will correspond to a catalyzed metabolic (or even chemical) reaction. Some of them may just be artificial fragmentation products, but just as in proteomics applications, where such artificial fragments are systematically exploited for peptide identification, they will also be informative in the case of metabolomics. More importantly, such artifacts provide an excellent "gold standard" for the evaluation of our approach: we know, for example, that isotope peaks should exist in our dataset, so re-discovering the corresponding patterns in an unsupervised manner demonstrates the general feasibility of the approach.

The semi-targeted approach confirms the results of the untargeted network reconstruction, and largely overcomes the issue of mass spectrometric artifacts. In this case, 1438 mass differences correspond to one of the major biochemical transformations, compared to 271 (±25) for a random list of masses of the same size. The most common mass differences correspond to hydrogenation/dehydrogenation (H_2; 284 occurrences), ethylene addition (C_2H_2; 211), ethyl addition (C_2H_4; 191), hydroxylation (O; 84) and palmitoylation ($C_{16}H_{30}O$; 57), all of them expected to be abundant in our membrane rich samples, based on general biochemical knowledge.

To determine the importance of mass accuracy for the ability of reconstructing metabolic networks, we added random noise of various size to the observed masses, i.e. a uniformly distributed random number from an interval indicated in the table was added or subtracted from each observed mass. We then performed the same untargeted analysis as before on these noisy data (table 5). The results show clearly that the reconstructed networks are robust against noise of this type – provided the accuracy of mass identification is ultra high. This analysis indicates that when mass accuracy falls to a number greater than 10 ppm in the order of 50% of the inferable transformations are lost. High accuracy spectra are essential for this approach to work.

Further confirmation of the non-random nature of the observed mass–mass difference network is provided by

Table 4
Comparison of the most common mass differences in observed and random metabolite networks

Observed metabolite network				Random metabolite network	
Mass difference	Frequency	Formula	Exact mass	Mass difference	Frequency
2.01595	382	H_2	2.01565	92.70975	7
21.98312	326	Na–H	21.98194	205.30491	7
1.00320	284	^{13}C isotope	1.0033	52.82462	7
24.00000	260	C_2	24	193.60014	6
26.01629	237	C_2H_2	26.01565	243.29213	6
28.03188	218	C_2H_4	28.03130	254.75355	6
4.03201	197	H_4	4.03130	6.46724	6
1.01259	164	H_2–^{13}C isotope	1.01229	52.69339	6
3.01910	148	H_2+^{13}C isotope	3.01900	21.98649	6
22.99695	140	C_2–^{13}C isotope	22.99664	22.12482	6
Total	25,370 (in 2472 clusters of >5 members)			Total	115 ± 22 (in 19 ± 4 clusters of >5 members)

Table 5
Stability of network inference against noise

	Real data	1 ppm	2 ppm	5 ppm	10 ppm	20 ppm	100	1000	10,000	Random
Clusters	2472	2254	2040	1660	1365	1113	470	53	35	19
Explained distances	25,370	22,988	21,094	16,966	13,213	10,003	3261	328	213	115
% Excess over random	100	90.5	83.1	66.7	51.9	39.2	12.5	0.8	0.4	0

Uniformly distributed random noise of the indicated size was added to all observed masses and the network reconstructed as described in the text.

an analysis of the frequency of the various reactions. As shown in figure 1A, the number of times a specific mass difference is observed depends on its rank in the form of a power-law. This means that there are many rare reactions, but a few principal reactions/mass differences account for most chemical interconversions visible within the total dataset. Such a distribution would not be expected in a random network, but has been reported as an organizing principle for various metabolic networks (Jeong *et al*., 2000; Wagner and Fell, 2001; Ravasz *et al*., 2002; Almaas *et al*., 2004). These previous studies were generally based on a select series of enzymes and metabolites reported in the basic biochemical literature, or from genome-wide analysis of enzymatic reactions putatively present in an organism, superimposed on this historical view of metabolism (Edwards *et al*., 2001; Schilling *et al*., 2002; Forster *et al*., 2003; Covert *et al*., 2004). In striking contrast, the networks that we have identified, based on mass spectrometric data, reveal the potential of generating network connectivity "on the fly" from experimental results, without biasing outcomes based on well-established, but clearly incomplete, biochemical pathways. Interestingly, the degree distribution of the observed metabolites (i.e. the number of metabolic reactions in which each is predicted to participate) does *not* fit a power-law distribution in our data, but rather follows an exponential distribution with only slightly heavy tails (figure 1B). This is not consistent with those earlier reports (Jeong *et al*., 2000; Wagner and Fell, 2001) describing network properties extrapolated from enzymatic pathways predicted from whole-organism genome sequence information. It is, however, important to reiterate that we only reliably measure metabolites in the range 100–1500 atomic mass units. Thus, many central metabolites (e.g. water, CO_2, pyruvate, glutamate) fall outside the mass window that we explore. This results in an absence of numerous major network "hubs", and this influences the overall topology of the network and in particular removes the corresponding heavy tails in the degree distribution. In contrast, when examined from the point-of-view of metabolic transformations, which will take many important "hub metabolites" into account implicitly, the degree distribution is clearly following a power-law, although in the case presented here this distribution is also influenced by other chemical relationships that result from our mass spectrometric analysis. A future challenge will be to refine networks to include maximal information derived from the metabolome, while minimizing interference related to technical effects associated with sample preparation and analysis. In spite of this, the *ab initio* metabolic networks described here are in good agreement with the *in silico* networks derived through interpretation of genome content and biochemical literature. Technical refinements and variations in experimental design will certainly lead to further improvements in the amount and quality of information that can be used to build networks *ab initio* using Fourier transform mass spectrometry. Our results indicate that the effort required for these technical refinements is clearly warranted by the potential of the method to provide comprehensive and relatively unbiased overviews of the cellular metabolome.

Figure 2 shows an extract of the metabolic network, focusing on compounds that are greatly enriched in parasites compared to their environment. The same diagram also demonstrates the ease with which predicted transformations may be visualized within the network. Mass 809.5939 was predicted by database matching to be a choline phospholipid with four unsaturated bonds and 38 carbons in the lipid side chains. While mass alone cannot provide the identity of such a lipid, 1-stearoyl, 2-arachidonoyl-phosphatidylcholine (calculated mass = 809.5935) falls within the limits determined for these FTMS experiments. Moreover, a phosphocholine of this class has previously been identified as predominant in the *T. brucei* phospholipidome (Patnaik *et al*., 1993) making this a very like candidate. This identification was then used to predict the molecular identity of the connected metabolites, and all but one of the network's 44 members were successfully assigned putative formulae in this way. All of them are phospholipids with various side chain compositions and different headgroups, which again conforms to expectations, as the parasite samples are rich in membrane material. This identification is confirmed by the clear pattern that emerges when one looks for metabolites whose mass-to-mass difference can be explained by side-chain elongation and side-chain (de-)saturation. Figure 3 shows the resulting pattern. It demonstrates that the abundance of the predicted phospholipid masses follows a clear trend, with higher degrees of unsaturation at larger side chain length. This is a well-known phenomenon supports our mass identification. Even stronger support is obtained when we compare the abundance of masses in parasites and serum. Three ether phospholipids stand out as dramatically enriched in parasites compared to their environment. This overabundance is in perfect agreement with reports in the literature (Patnaik *et al*., 1993). Figure 3 also shows that many of the putative phospholipid masses correlate with mass 809.5939 in abundance in the parasite samples. This indicates that correlations in ion abundance can also be used as indicators of connectivity, although at a coarser level than provided by the mass–mass differences. In a large-scale system perturbation study, such correlations could thus be an important piece of complementary information.

3.4. Metabolic fragment analysis

Mass spectrometry fails to resolve structural isomers. Thus in spite of the high likelihood that our assigned chemical identification is robust (based on both exact mass calculation and metabolic connectivity) we have sought additional means of assigning an identity. The

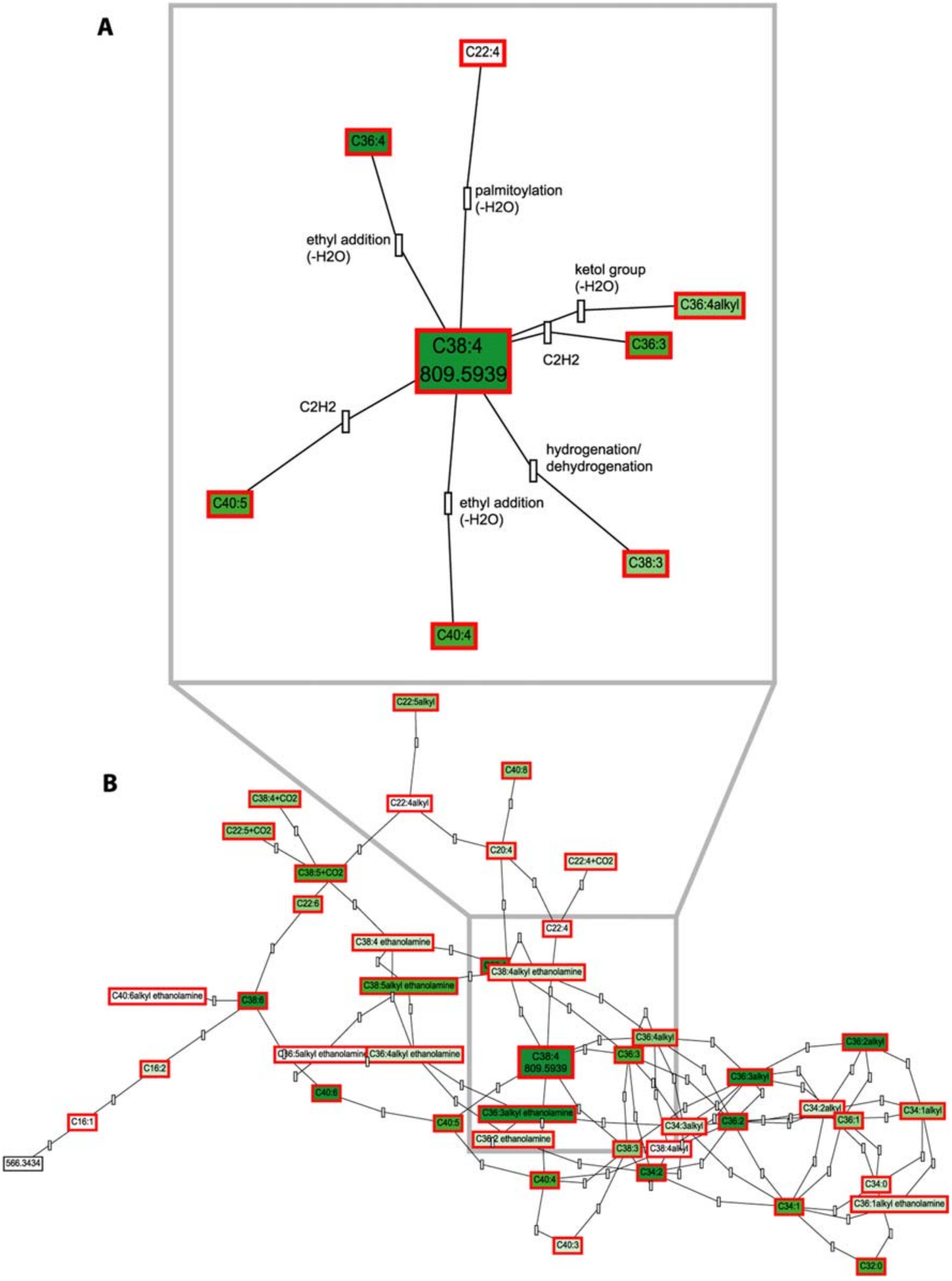

Figure 2. Extract of the *ab initio* metabolic network of *Trypanosoma brucei*. For clarity we show only metabolites that correlate in abundance with mass 809.5939, an unsaturated phosphatidyl choline phospholipid that is part of an enriched metabolite family in the parasite. The inset (A) highlights the first generation of transformations originating from mass 809.5939, the main figure (B) shows the entire subgraph, which connects more than 60% of the most strongly correlating masses (Pearson correlation $r > 0.85$). Assigned molecular identities for each metabolite are indicated in a shorthand notation, where C*n*:*m* stands for a phosphatidyl choline with *n* carbon atoms in the side chains and *m* unsaturated bonds. Alternative headgroups are explicitly mentioned in the labels. Shades of green indicate the abundance of the metabolites in the parasite. The graph layout was generated using aiSee (http://www.aisee.com).

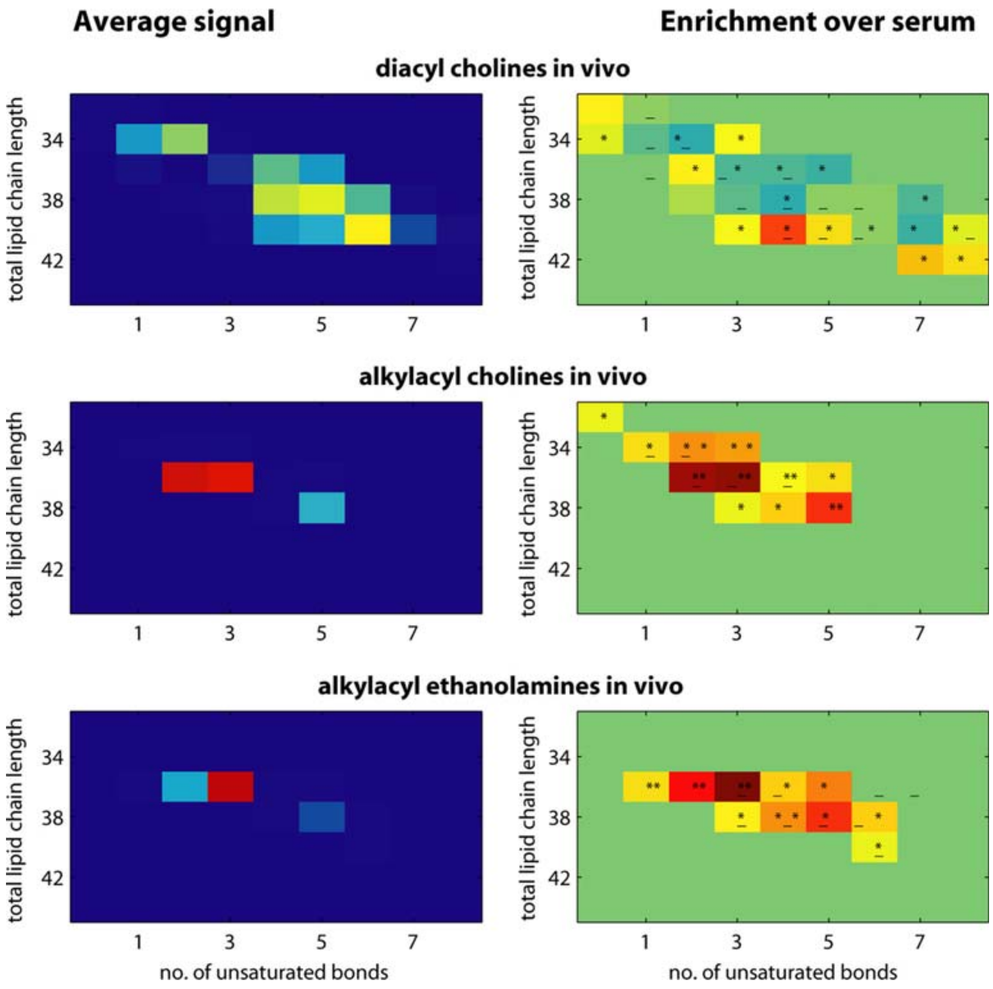

Figure 3. Abundance profile of various phospholipid classes. Diacyl cholines, alkylacyl cholines and alkylacyl ethanolamines are shown. The number of unsaturated bonds increases from left to right, the number of carbons in the side chains from top to bottom. The upper left corresponds to a saturated C16, C16 phospholipid, the bottom right to an 8-fold unsaturated C22, C22 molecule. The left column shows the absolute signal strength in trypanosomes *in vivo*. The right column shows the relative abdundance of the lipids compared to their concentration in the serum supernatant. Shades of blue indicate depletion in the parasites, yellow and red enrichment. One star denotes that the difference is significant at $p < 0.05$ (two-tailed t-test), two stars indicate that the same significant difference is also seen in vitro. A bar highlights masses that correlate in abundance with mass 809.5939. All three lipid classes show the same overall trends, with higher unsaturation at higher chain length. The highest abundance is found for three types of alkylacyl lipids, which can be putatively identified as C18:2,C18:0 alkylacyl phosphatidyl choline, C18:2, C18:0 alk-1-enylacyl phosphatidyl choline, and C18:2, C18:0 alk-1-enylacyl phosphatidyl ethanolamine.

gold standard in determining positive identification involves targeted fragmentation of selected masses followed by a second mass spectrometry step. This tandem mass spectrometry process is, however, itself challenging and requires additional sample preparation and technical development.

Careful analysis of the FTMS dataset offers an additional route to add supporting data towards assignments, based on what we call ‘‘metabolic fragment analysis’’. The technique is based solely on peaks derived from the dataset. Most biomolecules (including phospholipids) are formed by the condensation of building blocks and these may also be catabolized back to the building blocks by hydrolysis. For phospholipids, these building blocks will comprise the side chain fatty acids and the polar head groups. Hence, we searched for all triples of masses that could be explained by condensation/hydrolysis reactions (i.e. $\text{mass}_1 + \text{mass}_2 = \text{mass}_3 + \text{mass}H_2O$, at 2 ppm accuracy). About 581 masses (about half of all those detected) are putative condensation/hydrolysis products of other masses within the dataset, with a total of 1637 inferred reactions. Fifteen masses are putatively involved in at least 20 condensations each, and four masses in more than 30 each (table 6). With the exception of phosphocholine these all have masses in a narrow range between 280 and 370, and most of them are putative sidechain fatty acids. Other common ‘‘metabolic fragments’’ are choline phosphate (183.0661) with 26, glycerylphosphorylcholine (257.1029) with 15, and palmitoyl lysolecithin (495.3316) with 17 condensation reactions. This information can be used to infer the side chain composition of the phospholipids. For example, mass 727.5509, the most abundant phospholipid of the trypanosome pellet,

Table 6
Masses that occur in at least 30 putative condensation reactions among masses in our dataset. Their relative abundance in the various types of samples is indicated on an arbitrary scale. n.d., not detectable

Mass	# Of condensations	Putative identity	*In vivo*	*In vitro*	Serum	Medium
280.2395	42	Linoleic acid	+ +	+ +	+ + +	+ + +
302.2245	39	Icosapentaenoic acid	+	n.d.	n.d.	n.d.
320.2342	39	Hydroxy icosatetraenoic acid?	+	n.d.	n.d.	n.d.
309.2753	35	?	n.d.	+	+	+
312.2666	34	Hydroxy nonadecenoic acid	+	n.d.	n.d.	n.d.
328.2407	30	Docosahexaenoic acid	+	n.d.	n.d.	n.d.

is a putative condensation product of masses 465.3207 and 280.2395. The latter corresponds to linoleic acid, leading to the prediction that 727.5509 is an 18:0 alk-1-enyl,18:2 acyl phosphatidylethanolamine. The single previous study (Patnaik *et al*., 1993) aimed specifically at characterizing the molecular species of phospholipids in trypanosomes also revealed that 18:0, 18:2 species were by far the most abundant in trypanosomes.

4. Concluding remarks

Our results indicate that the unprecedented mass accuracy of Fourier transform mass spectrometry can lead to qualitative, rather than merely quantitative, advances in the study of cellular metabolism. Issues of sample preparation (e.g. loss of labile metabolites) and metabolite detection (e.g. ion suppression), which currently restrict the numbers of metabolite visible in Fourier transform mass spectrometry, remain a challenge (as discussed in Aharoni *et al*., 2002; Tohge *et al*., 2005). However, our study shows that the technology, coupled to advances in bioinformatic data interpretation, has great potential to allow unbiased and comprehensive studies of complex metabolic systems. Increasing numbers of metabolites should become visible as sample preparation parameters are optimized, and further advances in ultra-high resolution mass spectrometry promise to lead to substantial increases in the quantity of high resolution mass spectrometry data available for analysis (see for example Olsen *et al*., 2005). This will have a dramatic impact on the way such systems will be perceived and analyzed by biologists.

Acknowledgments

RB was supported by a personal fellowship from the Caledonian Research Foundation. We thank Lesley Nicholson and Alix Hayden for critical discussions.

References

Aharoni, A., Ric de Vos, C.H., Verhoeven, H.A., Maliepaard, C.A., Kruppa, G., Bino, R. and Goodenowe, D.B. (2002). Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *Omics* **6**, 217–234.

Allen, J., Davey, H.M., Broadhurst, D., Heald, J.K., Rowland, J.J., Oliver, S.G. and Kell, D.B. (2003). High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692–696.

Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. and Barabasi, A.L. (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427**, 839–843.

Arita, M. (2004). The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. USA* **101**, 1543–1547.

Barrett, M.P., Burchmore, R.J., Stich, A., Lazzari, J.O., Frasch, A.C., Cazzulo, J.J. and Krishna, S. (2003). The trypanosomiases. *Lancet* **362**, 1469–1480.

Brown, S.C., Kruppa, G. and Dasseux, J.L. (2005). Metabolomics applications of FT-ICR mass spectrometry. *Mass Spectrom. Rev.* **24**, 223–231.

Butler, D. (2005). Parasitology: triple genome triumph. *Nature* **436**, 337.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. and Palsson, B.O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96.

Edwards, J.S., Ibarra, R.U. and Palsson, B.O. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125–130.

Forster, J., Famili, I., Fu, P., Palsson, B.O. and Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* **13**, 244–253.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**, 245–252.

Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T. and Saito, K. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **101**, 10205–10210.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651–654.

Kell, D.B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **7**, 296–307.

Ma, H. and Zeng, A.P. (2003a). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277.

Ma, H.W. and Zeng, A.P. (2003b). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19**, 1423–1430.

Murch, S.J., Rupasinghe, H.P., Goodenowe, D. and Saxena, P.K. (2004). A metabolomic analysis of medicinal diversity in Huangqin (Scutellaria baicalensis Georgi) genotypes: discovery of novel compounds. *Plant Cell Rep.* **23**, 419–425.

Nicholson, J.K., Holmes, E., Lindon, J.C. and Wilson, I.D. (2004). The challenges of modeling mammalian biocomplexity. *Nat. Biotechnol.* **22**, 1268–1274.

Olsen, J.V., de Godoy, L.M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S. and Mann, M. (2005). Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell Proteomics* **4**, 2010–2021.

Patnaik, P.K., Field, M.C., Menon, A.K., Cross, G.A., Yee, M.C. and Butikofer, P. (1993). Molecular species analysis of phospholipids from Trypanosoma brucei bloodstream and procyclic forms. *Mol. Biochem. Parasitol.* **58**, 97–105.

Pfeiffer, T., Soyer, O.S. and Bonhoeffer, S. (2005). The evolution of connectivity in metabolic networks. *PLoS Biol.* **3**, e228.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555.

Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S. and Palsson, B.O. (2002). Genome-scale metabolic model of Helicobacter pylori 26695. *J. Bacteriol.* **184**, 4582–4593.

Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019–1026.

Tohge, T., Nishiyama, Y., Hirai, M.Y., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D.B., Kitayama, M., Noji, M., Yamazaki, M. and Saito, K. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *Plant J.* **42**, 218–235.

Wagner, A. and Fell, D.A. (2001). The small world inside large metabolic networks. *Proc. Biol. Sci.* **268**, 1803–1810.

Zhang, J., McCombie, G., Guenat, C. and Knochenmuss, R. (2005). FT-ICR mass spectrometry in the drug discovery process. *Drug Discov. Today* **10**, 635–642.