**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

aGroningen Bioinformatics Centre, University of Groningen, 9751 NN, Haren, The Netherlands
bInstitute of Biomedical and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK
cPhenomenome 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 data-sets 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 \textit{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 \textit{E. coli}. Indeed, genomic information of any kind is not required for these analyses. Construction of such \textit{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 G.H. Coombs (University of Glasgow). DB75 (2,5-bis(4-amidinophenyl)furan) was from D. Boykin (Georgia State University). Pentamidine isethionate 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 G.H. Coombs (University of Glasgow). DB75 (2,5-bis(4-amidinophenyl)furan) was from D. Boykin (Georgia State University).

2.2. Preparation of trypanosome extracts

Bloodstream form trypanosomes (Lister 427 line) were collected by cardiac puncture at $5 \times 10^8$ parasites/ml from Wistar rats and separated from red blood cells as a buffy coat by centrifugation at 3000g. The same parasite line was grown in HMI-9 medium supplemented with 20% foetal calf serum to mid-log phase ($8 \times 10^7$ parasites/ml) of culture medium. Parasites were then centrifuged at 3000g 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 $\mu$L) were extracted in solvents ranging from polar (aqueous) to non-polar with most proteins and nucleic acids removed during extraction. Extracts were stored at $-80^\circ$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 $\mu$L 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 \textit{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 \textit{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, \textit{Trypanosoma brucei}

We then proceeded with the analysis of metabolite samples from \textit{Trypanosoma brucei}, a protozoan parasite
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 13C 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

| 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       | 721.2898      | 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              | 281.13899      | 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.
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

| Reaction | Formula | Reaction | Formula |
|----------|---------|----------|---------|
| Alanine  | C₃H₅NO | Guanosine 5′-diphosphate (–H₂O) | C₁₀H₁₃N₅O₁₀P₂ |
| Arginine | C₆H₁₂N₄O | Guanosine 5-monophosphate (–H₂O) | C₁₀H₁₂N₅O₇P |
| Asparagine | C₇H₁₈N₄O₂ | Guanine (–H) | C₇H₆N₂O |
| Aspartic acid | C₄H₇NO₄ | Guanosine (–H₂O) | C₁₀H₁₁N₄O₄ |
| Cysteine | C₃H₅NOS | Deoxythymidine 5′-diphosphate (–H₂O) | C₁₀H₁₃N₅O₉P₂ |
| Cystine | C₃H₄N₄O₂S₂ | Thymidine (–H₂O) | C₇H₄N₂O₄ |
| Glutamic acid | C₄H₇NO₄ | Thymine (–H) | C₇H₆N₂O |
| Glutamine | C₄H₇NO₂ | Thymidine 5′-monophosphate (–H₂O) | C₁₀H₁₃N₅O₈P |
| Glycine | C₂H₄N₂O | Uridine 5′-diphosphate (–H₂O) | C₁₀H₁₄N₄O₄P₂ |
| Histidine | C₇H₁₄N₃O | Uridine 5′-monophosphate (–H₂O) | C₁₀H₁₄N₄O₄P |
| Isoleucine | C₆H₁₂N₂O | Uracil (–H) | C₄H₃N₂O₂ |
| Leucine | C₆H₁₁N₂O | Uridine (–H₂O) | C₁₀H₁₂N₅O₆P |
| Lysine | C₆H₁₄N₂O | Acetylation (–H) | C₇H₂ |
| Methionine | C₅H₁₂N₂O | Acetylation (–H₂O) | C₁₀H₁₄N₄O₂ |
| Phenylalanine | C₅H₉NO | Carboxylation | CO₂ |
| Proline | C₅H₇NO | CHO₂ |
| Serine | C₄H₇NO₂ | CH₂ |
| Threonine | C₄H₇NO₂ | Condensation/dehydration | H₂O |
| Tryptophan | C₁₁H₁₄N₂O₂ | Diphosphate | H₂O₃P₂ |
| Tyrosine | C₄H₉NO₃ | Ethyl addition (–H₂O) | C₇H₇ |
| Valine | C₅H₈NO | Formic Acid (–H₂O) | CO₂ |
| Acetotetrate (–H₂O) | C₅H₁₂O₂ | Glyoxylate (–H₂O) | C₅H₂ |
| Acetone (–H) | C₃H₄O | Hydrogenation/dehydrogenation | H₂ |
| Adenylate (–H₂O) | C₁₀H₁₄N₃O₄P | Hydroxylation (–H) | O |
| Biotin (–H) | C₁₀H₁₄N₂O₂S | Inorganic phosphate | P |
| Biotin (–H₂O) | C₁₀H₁₄N₂O₄S | Ketal group (–H₂O) | C₅H₉O |
| Carboxamyl P transfer (–H₂PO₄) | CH₂ON | Methanol (–H₂O) | CH₃ |
| Co-enzyme A (–H) | C₂₁H₃₄N₇O₁₆P₃S | Phosphate | HPO₃ |
| Co-enzyme A (–H₂O) | C₂₁H₃₃N₇O₁₅P₃S | Primary amine | NH₃ |
| Glutathione (–H₂O) | C₁₀H₁₄N₃O₄S | Pyrophosphate | PP |
| Isoprene addition (–H) | C₁₀H₁₂N₃O₂S | Secondary amine | NH₂ |
| Malonyl group (–H₂O) | C₁₆H₃₀O₫ | Sulfate (–H₂O) | SO₃ |
| Palmitoylation (–H₂O) | C₁₆H₃₀O₫ | Tertiary amine | N |
| Pyridoxal phosphate (–H₂O) | C₁₀H₁₄N₃O₄S | | |
| Urea addition (–H) | CH₂N₂O | D-ribose (–H₂O) (ribosylation) | C₅H₁₀O₅ |
| Adenine (–H) | C₁₀H₁₄N₃O₄ | Disaccharide (–H₂O) | C₁₀H₁₄O₅ |
| Adenosine (–H₂O) | C₁₀H₁₄N₃O₄ | Glucose-N-phosphate (–H₂O) | C₁₀H₁₄O₅P |
| Adenosine 5′-diphosphate (–H₂O) | C₁₀H₁₄N₃O₄P₂ | Glucuronic acid (–H₂O) | C₇H₁₀O₅ |
| Adenosine 5′-monophosphate (–H₂O) | C₁₀H₁₄N₃O₄P | Gluconic acid (–H₂O) | C₇H₁₀O₅ |
| Cytidine 5′-diphosphate (–H₂O) | C₁₀H₁₄N₃O₄P₂ | Monosaccharide (–H₂O) | C₅H₁₀O₅ |
| Cytidine 5′-monophosphate (–H₂O) | C₁₀H₁₄N₃O₄P | Trisaccharide (–H₂O) | C₁₄H₁₄O₁₅ |
| Cytosine (–H) | C₅H₄N₂O | | |

### 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 | 1.03 | 1.06 | 1.09 | 1.32 | 2.03 | 3.40 | 6.08 |
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 \textit{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 ($\text{H}_2$; 284 occurrences), ethylene addition ($\text{C}_2\text{H}_2$; 211), ethyl addition ($\text{C}_2\text{H}_4$; 191), hydroxylation (O; 84) and palmitoylation ($\text{C}_{16}\text{H}_{32}\text{O}$; 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

| Mass difference | Frequency | Formula | Exact mass | Mass difference | Frequency |
|-----------------|-----------|---------|------------|----------------|-----------|
| 2.01595         | 382       | H$_3$   | 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$_3$H$_2$ | 26.01565 | 243.29213      | 6         |
| 28.03188        | 218       | C$_3$H$_4$ | 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         |
| 2.019110        | 148       | H$_2$–$^{1}$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

| Clusters | Real data | 1 ppm | 2 ppm | 5 ppm | 10 ppm | 20 ppm | 100 | 1000 | 10,000 | Random |
|----------|-----------|-------|-------|-------|--------|--------|-----|------|--------|--------|
| 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 within the total dataset. Such a distribution would not be accounted for most chemical interconversions, but a few principal reactions/mass differences 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 metabolite 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,stearyl, 2-arachidonoyl-phosphatidylethanolamine (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 likely 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).
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. $\text{mass}_1 + \text{mass}_2 = \text{mass}_3 + \text{massH}_2\text{O}$, 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-ene,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., Ottav, 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.J., Frasch, A.C., Cazzullo, J.J. and Krishna, S. (2003). The trypanosomiasis. 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., Awazuahara, M., Arita, M., Fujiwara, T. and Saito, K. (2004). Integration of transcriptomics and metabolomics for understanding 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 Huang-qin (Scutellaria baicalensis Georgii) 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.