Probabilistic assignment of formulas to mass peaks in metabolomics experiments

Rogers, Simon; Scheltema, Richard A.; Girolami, Mark; Breitling, Rainer

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btn642

Publication date:
2009

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 4 2009, pages 512–518
doi:10.1093/bioinformatics/btn642

Systems biology

Probabilistic assignment of formulas to mass peaks in metabolomics experiments

Simon Rogers[1,*], Richard A. Scheltema[2], Mark Girolami[1] and Rainer Breitling[2]

[1]Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK and [2]Groningen Bioinformatics Centre, University of Groningen, 9751 NN Haren, The Netherlands

Received on August 22, 2008; revised on November 12, 2008; accepted on December 11, 2008

Advance Access publication December 18, 2008

Associate Editor: Trey Ideker

ABSTRACT

Motivation: High-accuracy mass spectrometry is a popular technology for high-throughput measurements of cellular metabolites (metabolomics). One of the major challenges is the correct identification of the observed mass peaks, including the assignment of their empirical formula, based on the measured mass.

Results: We propose a novel probabilistic method for the assignment of empirical formulas to mass peaks in high-throughput metabolomics mass spectrometry measurements. The method incorporates information about possible biochemical transformations between the empirical formulas to assign higher probability to formulas that could be created from other metabolites in the sample. In a series of experiments, we show that the method performs well and provides greater insight than assignments based on mass alone. In addition, we extend the model to incorporate isotope information to achieve even more reliable formula identification.

Availability: A supplementary document, Matlab code, data and further information are available from http://www.dcs.gla.ac.uk/inference/metsamp.

Contact: srogers@dcs.gla.ac.uk

1 INTRODUCTION

Recent advances in ultra high mass accuracy mass spectrometry are providing new tools for the comprehensive measurement of cellular metabolites (metabolomics) (Breitling *et al*., 2006b; Dunn *et al*., 2008; Lu *et al*., 2008). One of the major bottlenecks in the high-throughput application of mass spectrometry in metabolomics is the assignment of the observed peaks to empirical formulas. This is particularly challenging for broad, untargeted metabolome screens using liquid chromatography mass spectrometry, which can detect hundreds or thousands of uncharacterized metabolites in a biological sample. Lack of knowledge regarding what one expects to find in such samples makes it impossible to use traditional, target-based techniques. The ultra high resolution provided by the most recent generation of mass analyzers should allow the identification of empirical formulas in certain mass ranges based on mass alone, but even at 1 p.p.m. mass accuracy, this is not unambiguously possible for larger metabolites (Kind and Fiehn, 2006). Various pre-processing tools can be used to filter the list of potential empirical formulas (potential combinations of atoms and not necessarily their structural formula; see for example, Kind and Fiehn, 2007), but the problem cannot be completely eradicated and the development of more sophisticated methods is necessary. One possible solution is to pre-process the sample for easier identification. For example, Hegeman *et al*. (2007) propose creating four different isotopically labeled samples to be presented to the mass spectrometer. This greatly improves identification but adds a sizeable experimental and economic burden.

Often extra information will be available, in addition to the accurate mass of individual formulas. For example, mass spectrometry is usually coupled to gas or liquid chromatography, which provides retention time measurements that contain information about the general biophysical characteristics of a detected compound. Also, metabolites of interest are often observed in many related samples, e.g. from multiple patients or along time series, and the resulting information could be used for more reliable identification. Most importantly, however, metabolites are not detected in isolation, but together with hundreds of others that are produced by the same metabolic network. It is these network relationships between observed masses that we will concentrate on in this work.

An almost trivial example of mass peaks exhibiting informative relationships is provided by the isotope peaks detected for a single compound: if we observe an intense monoisotopic peak for a particular compound, we are very likely to also observe isotope peaks at exactly defined positions and with predictable intensity ratios. If no such isotope peaks are observed at the predicted position or intensity, taking into account the variability of the measurements, this is strong evidence against a particular identification. This principle has been used successfully for filtering potential empirical formulas (Kind and Fiehn, 2007).

But there are more possibilities for informative biochemical relationships between metabolites in a sample. Most importantly, a large proportion of the detected masses will be produced via chemical transformations within a densely connected metabolic network. With the possible exception of xenobiotics, every detected metabolite will therefore be connected to every other compound by a series of chemical reactions. Conveniently, there are only a limited number of chemical reaction types that make up the network (Breitling *et al*., 2006b)—a common example being hydrogenation (the addition or removal of H_2). Each transformation will correspond to a known exact mass addition/subtraction, so if a hydrogenated compound is predicted to be present in a sample, it might be sensible

*To whom correspondence should be addressed.

to look for the corresponding non-hydrogenated compound, which should be lighter by exactly $\Delta m = 2.01565$ Da. Just as in the case of isotope peaks, the presence or absence of the related metabolites can support or weaken a hypothetical identification. This will be the basis of the probabilistic approach introduced in this article.

In Breitling *et al*. (2006b), the idea of using mass difference between metabolites for data interpretation was first introduced and a systematic search of commonly occurring mass differences was used to create a metabolic map of the sample. Similar concepts have also been applied in petroleomics and natural organic matter analysis (Koch *et al*., 2007; Kujawinski and Behn, 2006). Alternatively, putative biochemical connections can be extracted from databases such as KEGG. In addition, once transformation information is used to filter the lists of potential metabolites, it is no longer necessary to restrict the search to known compounds, such as those found in PubChem or Metlin, and it becomes possible to predict potential empirical formulas using exhaustive enumeration of atomic combinations (as discussed in Kind and Fiehn, 2007).

In this work, we describe a probabilistic approach for assigning empirical formulas to a list of measured mass spectrometric peaks, given a list of potential formulas and possible biochemical transformations. As our statistical model defines the posterior probability distribution over the *complete* set of assignments from masses to empirical formulas, it is also able to capture the implicit dependencies between molecular identifications—if mass 1 is assigned to compound *a* then mass 2 is more likely to be assigned to compound *b*. Of course, enumeration of the full distribution over assignments is infeasible even for a small number of masses and formulas due to the rapidly increasing number of combinations. However, the conditional nature of the prior distribution over empirical formulas makes Markov Chain Monte Carlo (MCMC) sampling from the posterior very appealing and we present an efficient Gibbs sampling procedure to implement our approach.

2 THE MODEL

Assume that we are provided with a set of measured masses, x_m ($m = 1 \ldots M$), and a set of C potential empirical formulas, with masses y_c ($c = 1 \ldots C$). Our task is to assign each of the masses to one of the potential formulas. These potential empirical formulas can be collected from databases of chemicals and biomolecules, or they can be based on exhaustive enumeration of possible atomic combinations for each measured mass (Kind and Fiehn, 2007).

We first define a noise model for the mass part of the overall statistical model. The absolute accuracy of mass measurements is dependent on the mass being measured—it is normally given in parts per million—hence we need a noise model that reflects this. An appropriate choice is a Gaussian model on the ratio of the measured mass x_m to the mass of the potential formula y_c, although this could be replaced with any other noise model if desired. Defining the $C \times M$ binary matrix of assignments, $\mathbf{Z}$, where $z_{cm} = 1$ if mass m has been assigned to formula c, then the noise model (or likelihood) is defined as follows

$$p(x_m | z_{cm} = 1, y_c, \gamma) = \mathcal{N}\left(\frac{x_m}{y_c} \middle| 1, \gamma^{-1} \right),$$

where γ^{-1} is the mass accuracy (variance) that can be fixed a priori (if the approximate accuracy of the laboratory equipment is known) or inferred within the model based on the data. Note that $\sum_c z_{cm} = 1$, as each mass can only be assigned to one compound, while there is no restriction on $\sum_m z_{cm}$, as several masses could be assigned to the same compound. If we were to base our assignments on the mass alone, we would choose the closest mass (corresponding to the maximum of this likelihood) or we could normalize over the c possible formulas to give a discrete distribution. In this trivial case, if the masses of the potential empirical formulas are never closer than the mass tolerance, assigning each mass to the compound with the highest probability is sensible. However, this situation is unlikely in real metabolomics measurements, where hundreds of metabolites yield thousands of mass peaks, and there are tens of thousands of potential empirical formulas. In this case, we propose using a prior distribution over empirical formulas that takes into account the potential relationships between them. We will start by looking at biochemical relationships and discuss how one can easily incorporate isotope information and additional types of evidence later.

It is easiest to demonstrate the idea with a cartoon. Figure 1 shows a simple example. We have three measured masses, m_1, m_2 and m_3, denoted by mass peaks. Based on a database search, they could correspond to four potential empirical formulas—gluconic acid ($C_6H_{12}O_7$), galacturonate 1-phosphate ($C_6H_{11}O_{10}P$), caffeine ($C_8H_{10}N_4O_2$, 194.08 Da) and glucuronic acid ($C_6H_{10}O_7$, 194.04 Da). For two of the masses (m_2 and m_3), their identity is obvious based on the mass alone. However, for mass m_1, two potential identifications are available, as the potential empirical formulas have very similar masses. Both formulas initially have identical probabilities (the measured mass, 194.06, is half way between their masses of 194.04 and 194.08 Da).

First, we assign m_3 to galacturonate 1-phosphate (Fig. 1B). This compound is related to one of the candidates for mass 1 (glucuronic acid) via the addition/removal of a phosphate (denoted by an arrow). Hence, it is sensible to increase the probability of assigning m_1 to glucuronic acid—this can be seen in the change in probabilities in the figure. Second (Fig. 1C), we assign m_2 to gluconic acid. This compound is also related to glucuronic acid via hydrogenation/dehydrogenation. This further skews the assignment probabilities for m_1. In light of this evidence, if we had to make an assignment for m_1, it would be identified as glucuronic acid, rather than the alternative, caffeine—a decision that could not be made based on mass alone.

In this illustrative example, we have simplified the discussion by assuming that the assignment of masses m_2 and m_3 was independent of (and preceeded) any assignment to m_1, but in reality all assignments are interdependent. For the example, the interpretation is trivial, but for a realistic dataset many of the potential assignments are dependent on one another, as the density of the possible interaction network between formulas is increased. This makes the incorporation of biochemical relationships a computational challenge—we have to evaluate full sets of assignments and not each assignment on an individual basis. With large numbers of masses and potential formulas, the number of possible assignment sets is too large to enumerate. Fortunately, it is easy to define a prior distribution for one mass conditioned on the current assignments of all other masses and this can be used to build an efficient Gibbs sampling scheme. This enables us to draw samples from the posterior assignment of masses to formulas given this prior and the likelihood (noise model) already mentioned. A similar idea has recently been proposed by (Sanguinetti *et al*., 2008) for classifying genes as

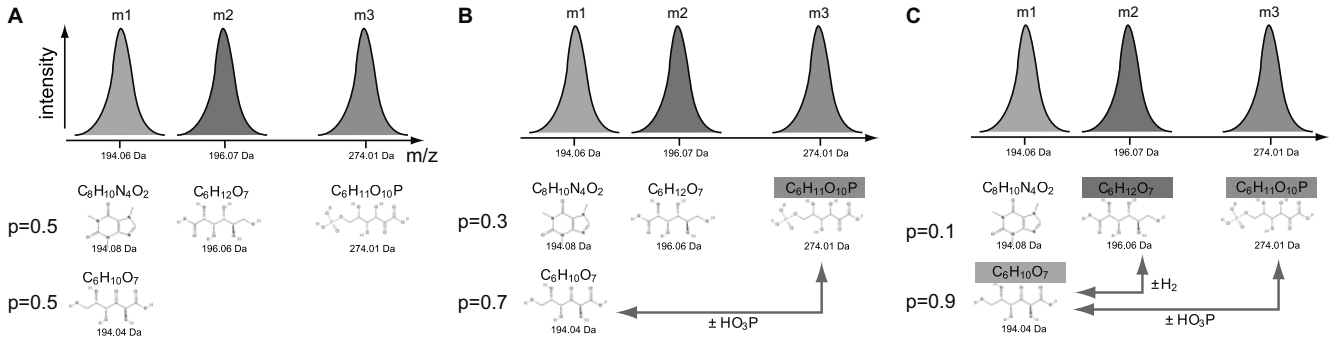

Fig. 1. Cartoon depicting the principle of our approach. Three peaks are observed in the mass spectrum. For mass m_1 two empirical formulas are initially equally likely ($P=0.5$), as they both differ by 0.02 Da from the observed mass. Assignment of mass m_3 as $C_6H_{11}O_{10}P$ provides support for the identification of m_1 as $C_6H_{10}O_7$ from which it differs by one phosphorylation reaction. Assignment of m_2 further increases the confidence in the assignment of m_1 to $C_6H_{10}O_7$, with the posterior probability increasing to $P=0.9$.

being over- and under-expressed based on gene expression data and connectivity information.

The conditional prior takes the following form:

$$p(z_{cm}=1|\mathbf{Z},\delta) = \frac{\beta_{cm}+\delta}{C\delta+\sum_{c'}\beta_{c'm}} \qquad (1)$$
$$\beta_{cm} = \mathbf{W}_{c\cdot}\mathbf{Z1}-\mathbf{W}_{c\cdot}\mathbf{Z}_{\cdot m},$$

where $\mathbf{W}$ is a $C\times C$ symmetric binary matrix encoding the chemical interactions, $\mathbf{1}$ is a $M\times 1$ vector of 1s and the notation $\mathbf{W}_{c\cdot}$ and $\mathbf{Z}_{\cdot m}$ denotes the c-th row and m-th column of $\mathbf{W}$ and $\mathbf{Z}$ respectively. Intuitively, β_{cm} can be thought of as the number of ways compound c can be produced from empirical formulas currently assigned to masses, without the assignments for the m-th mass. For later convenience, we will also define β_c, which is the number of ways c can be produced given all of the assignments (i.e. including the m-th assignment), and the two vectors $\boldsymbol{\beta}=[\beta_1,\ldots,\beta_c,\ldots,\beta_C]^T$ and $\boldsymbol{\beta_m}=[\beta_{1m},\ldots,\beta_{cm},\ldots,\beta_{Cm}]^T$. A more thorough description of the model is provided in the Supplementary Material.

Incorporating the likelihood based on the mass observation, we have the following posterior distribution

$$p(z_{cm}=1|\mathbf{Z},x_m,\mathbf{y},\delta,\gamma)\propto\mathcal{N}\left(\frac{x_m}{y_c}|1,\gamma^{-1}\right)\frac{\beta_{cm}+\delta}{N\delta+\sum_{c'}\beta_{c'm}},$$

where we must simply normalize over c to obtain a proper discrete distribution. If the hyperparameters γ^{-1} and δ are fixed, then our sampler proceeds as described in Algorithm 1. The first B samples should be discarded as a burn-in period, as discussed later.

It is easy to see that we can flexibly incorporate additional prior knowledge about the probability of observing certain metabolites in our sample. In the example discussed above, we may already expect to observe glucuronic acid (due to the source of our samples), while we do not expect to see caffeine, even before we made our measurements. This can be expressed in terms of empirical prior probabilities, which can be added to the model as an additional multiplicative term.

The two hyper-parameters γ and δ could either be marginalized out in the sampler or fixed by the user. We choose the latter option, setting γ to correspond to the MS accuracy and using $\delta=1$. Further discussion on setting δ is given in the Supplementary Material.

Input: M measured masses, $x_1,\ldots,x_m,\ldots,x_M$; C potential empirical formulas with masses $y_1,\ldots,y_c,\ldots,y_C$; and $C\times C$ connectivity matrix $\mathbf{W}$, hyper-parameters γ and δ
Output: A set of S samples of the $C\times M$ assignment matrix $\mathbf{Z}$, $\mathbf{Z}^1,\ldots,\mathbf{Z}^s,\ldots,\mathbf{Z}^S$
Initialize $\mathbf{Z}$ by randomly assigning masses to formulas.
Compute $\boldsymbol{\beta}=\mathbf{WZ1}$.
foreach *Sample s* **do**
  foreach *Mass m (selected in random order)* **do**
    Set $\boldsymbol{\beta}_m=\boldsymbol{\beta}-\mathbf{WZ}_{\cdot m}$
    Sample a compound c from the discrete distribution where the probability of the cth compound is proportional to
    $$\mathcal{N}\left(\frac{x_m}{y_c}|1,\gamma^{-1}\right)\frac{\beta_{cm}+\delta}{N\delta+\sum_{c'}\beta_{c'm}}$$
    Insert the new assignment into $\mathbf{Z}_{\cdot m}$
    Set $\boldsymbol{\beta}=\boldsymbol{\beta}_m+\mathbf{WZ}_{\cdot m}$
  end
  Set $\mathbf{Z}^s=\mathbf{Z}$
end

Algorithm 1: The Gibbs sampler

The algorithm was implemented in Matlab and all of our experiments were run on a standard iMac desktop computer. As an example of time required, generating 5000 samples for the Trypanosome data (Section 4) took ~4 min, comparing favorable with the several hours to generate the required list of potential formulas from KEGG (a necessary pre-processing step in any such metabolomic anaysis). The dominant memory requirement is the $C\times C$ connectivity matrix, $\mathbf{W}$ which could become unwieldy for particularly large sets of formulas. However, it will generally be very sparse and as such, even for very large numbers of formulas can be encoded very efficiently. More accurate timing details for the Trypanosome example are given in the Supplementary Material.

3 AN ILLUSTRATIVE EXAMPLE

We will illustrate the performance of the approach using a group of related metabolites involved in vitamin C metabolism. We simulated a metabolomics sample containing 12 compounds, including vitamin

Table 1. Assignment of 12 compounds from vitamin C metabolism based on mass alone

No	Compound	True mass	Measured mass	Assigned
1	D-Galacturonate $C_6H_{10}O_7$	194.0427	194.0395	Correct
2	L-Galactonate $C_6H_{12}O_7$	196.0583	196.0666	Correct
3	L-Galactose-1P $C_6H_{13}O_9P$	260.0297	260.0370	$C_4H_{13}N_4O_5SP$ (260.0344)
4	L-Galactose $C_6H_{12}O_6$	180.0634	180.0728	Correct
5	L-Galactono-1, 4-lactone $C_6H_{10}O_6$	178.0477	178.0429	$C_5H_{12}N_2OP_2$ (178.0425)
6	L-Ascorbate $C_6H_8O_6$	176.0321	176.0266	$C_2H_5N_6O_2P$ (176.0212)
7	L-Threonate $C_4H_8O_5$	136.0372	136.0409	Correct
8	3-Dehydro-L-threonate $C_4H_6O_5$	134.0215	134.0230	Correct
9	Monodehydro-ascorbate $C_6H_7O_6$	175.0243	175.0224	Correct
10	L-Dehydro-ascorbate $C_6H_6O_6$	174.0164	174.0114	$C_2H_10N_2O_3S_2$ (174.0133)
11	L-Ascorbate-6-phosphate $C_6H_9O_9P$	255.9984	255.9903	$C_5H_4N_7O_2P_2$ (255.9902)
12	3-Dehydro-L-gulonate-6-phosphate $C_6H_{11}O_{10}P$	274.0090	274.0042	$C_4H_{10}N_4O_6S_2$ (274.0042)

C (L-ascorbate) as well as an additional 100 masking compounds. The latter were drawn from a set of 3427 randomly generated metabolites with a relative mass difference of no more than 1% to one or more of the compounds of interest. The 3427 random formulas plus the 12 vitamin C-related compounds served as our database against which to match the measured masses.

A realistic experimental setting was simulated by adding noise to the masses of all 112 'measured' compounds. Gaussian noise, as described previously was used with $\gamma = 3 \times 10^8$ (1 SD corresponding to ~30 p.p.m.). In Table 1, we show the 12 vitamin C-related compounds, their true masses, their measured masses and how they would be assigned based purely on mass. If they are wrongly assigned, the formula to which they are assigned is given, along with its mass. Six of the 12 compounds are wrongly assigned based on mass alone because another compound in the database has a mass closer to the measured mass than the true formula of interest.

We now create a connectivity matrix for all 3439 compounds in the database using the following five transformations: hydrogenation/dehydrogenation ($\pm H_2$), hydroxylation ($\pm O$), ketol group addition ($\pm C_2H_2O$), condensation/dehydration ($\pm H_2O$) and phosphorylation/dephosphorylation ($\pm HO_3P$).

For real datasets, a larger number of potential transformations can be used; e.g. based on the list of common biochemical reactions given in Breitling *et al*. (2006b). The resulting connectivity matrix, the list of potential compound masses, and the list of measured masses (12 masses from the pathway plus 100 masking masses) are passed to the sampler which produces assignment probabilities.

This leads to a dramatic improvement in identifications: all the vitamin C-related compounds in the dataset are now correctly assigned. We can also quantify the improvement. As already mentioned, probabilistic assignments based on mass alone can be produced by normalizing the likelihood over all formulas. These can be compared with posterior probabilities from the sampler produced by computing (for a specific mass) the proportion of times it is assigned to each formula in the Gibbs samples. In Figure 2, we can see the relative gain in predictive probability (the probability assigned to what we know to be the correct formula) of the sampler compared to probabilities based on mass alone. We can see that there is a significant increase in almost all compounds, with the probability of being assigned to the correct compound more than double in one case (mass 5: L-galactono-1, 4-lactone). Of the three cases without a significant increase, Compounds 4 and 7 show very little change whilst Compound 9 has a decrease, the reason for which we will describe later. Figure 3 shows the masses represented by the correctly assigned compounds and the inferred connections between them. Whilst connections between compounds are fixed a priori, they are not fixed between masses. However, we can compute a posterior probability of connection from our sampler. Specifically, for every pair of masses, we count the number of times in which the two masses are assigned empirical formulas connected by an edge and divide this by the total number of Gibbs samples—giving the posterior probability of a connection. In the figure, we have shown all connections with posterior probability greater than 0.01. All connections that should be present as defined by our set of five transformations are present.

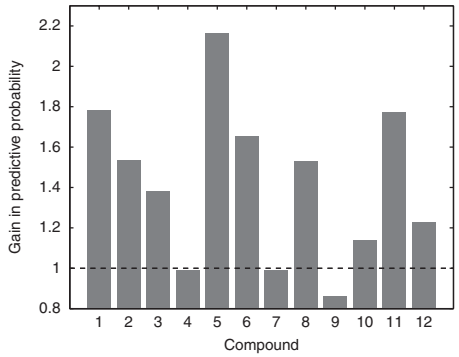

Fig. 2. Relative gain in predictive probability of the sampler compared to assignment based on mass only. For all but one masses the correct identification becomes much more confident when metabolic relationships are taken into account.

As already mentioned, the sampler is less likely to assign Compound 9 (monodehydro-ascorbate) correctly than the mass only method. If one restricts the analysis to the five common transformation types described above, Compound 9 is not connected to any other compounds in the vitamin C network. Hence, its probability of being assigned to the right formula is not increased by the connectivities, while it may well happen that one of the masking compounds has increased probability due to spurious connections. Importantly though, Compound 9 is still assigned correctly. It differs by a single hydrogen atom from Compound 10 and it is likely that if this very rare transformation type (mono-dehydrogenation)

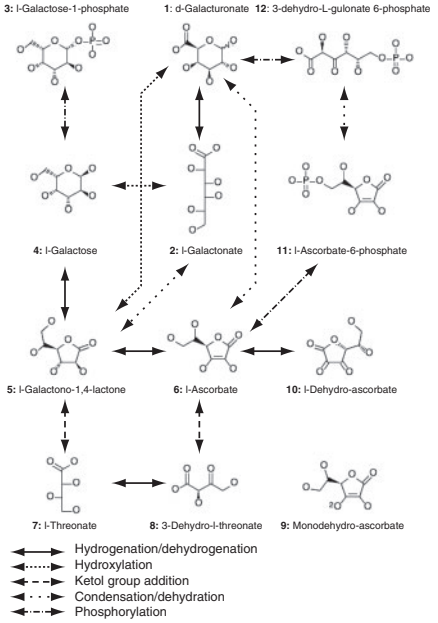

Fig. 3. Masses and their correctly assigned compounds. Connections with posterior $P > 0.01$ are shown with different arrow types corresponding to different transformation types. They do not necessarily correspond to an enzymatically catalyzed reaction, but can also just represent chemical relationships.

were to be included, it would have increased the confidence of the identification of Compound 9. The set of reactions to be considered is flexible and can be extended by the user to contain all reaction types that are considered relevant for the biological system being studied.

4 CASE STUDY ON EXPERIMENTAL DATA

We next evaluated the performance of our method on an experimental metabolomics dataset from *Trypanosoma brucei* (Breitling *et al*., 2006a). The data were acquired using Fourier Transform Ion Cyclotron Resonance mass spectrometry on whole-cell extracts of parasites grown *in vivo* and *in vitro* and pre-processed as described in Breitling *et al*. (2006a). To create a set of potential compounds, putative formulas matching each peak were identified by searching the KEGG, Metlin and Pubchem databases using the MetabolomeExplorer (Scheltema, R.A. *et al*., unpublished data) with a mass tolerance of 10 p.p.m. This lead to 339 of the 446 peaks being assigned to single empirical formulas, 87 peaks with two potential formulas, 18 with 3 and 2 peaks with 4 formulas—leaving a total of 379 unique formulas (some were assigned to more than one peak). To create the connectivity matrix, each pair of formulas were checked against a list of 83 possible transformations (the full list is given in the Supplementary Material). The Gibbs sampler was run for 3000 burn-in samples (recall that each Gibbs sample re-assigns each peak) and then a further 2000 samples from which posterior probabilities were calculated. The precision, γ, was fixed at 4×10^{10}, equivalent to a Gaussian distribution where 2 SD from the mean corresponds to a mass difference of 10 p.p.m.

Assigning peaks to the formula with the highest posterior probability resulted in 14 of the 446 peaks not being assigned to the

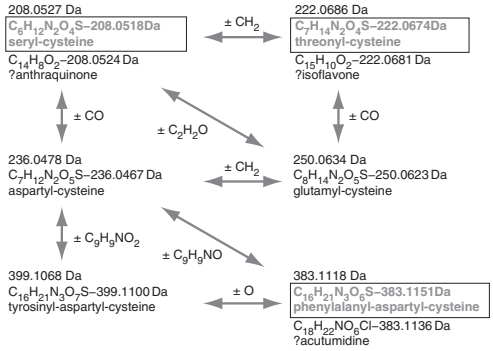

Fig. 4. Subnetwork from the trypanosome example. Black (un-boxed) assignments are made by mass alone, assignments in boxes (bold red) are made by the sampler. Where no boxed assignment is present, the two assignments agree.

compound with the closest mass (the default when assigning by mass alone). An example of 3 of the 14 reassigned masses are shown in their network context in Figure 4. Where the two methods disagree, the assignment from the sampler (when connectivity information is taken into account) are shown in boxes. All new assignments make more sense biologically than the initial identification: they indicate that all compounds in the network are di- and tri-peptides, which are expected to be abundant in the parasite samples, in contrast to the quite exotic formulas suggested by the mass-based assignment. Such an improvement might seem unexpected, given that the dataset was reported to have a mass accuracy of ∼1 p.p.m. (Breitling *et al*., 2006a), while the reassigned masses deviate by up to 4 p.p.m. from the measured mass. The reason is probably the limited dynamic range of mass accuracy (Makarov *et al*., 2006), with less accurate observations for less abundant compounds (all of the peptides except aspartyl-cysteine are observed at consistently low levels in the samples). This emphasizes the usefulness of our approach even for high-accuracy metabolomics experiments.

It is also interesting to note that the sampler provides more confidence in its predictions. Of the 63 masses for which the probability changes between the two methods, 41 (67%) show an increase in probability for the sampler compared to the mass-only approach (visualized in the Supplementary Material).

5 INCORPORATING ISOTOPE INFORMATION

We will now show how the framework described above can be extended to incorporate isotope information and other informative relationships between peaks in a mass spectrum.

Isotope information can be very useful in identifying metabolites of sufficient size and abundance. For that to be possible, we also require sufficient accuracy of the intensity measurements. Isotope peaks are commonly observed for reasonably large metabolites, as the probability of detecting molecules with at least one ^{13}C atom rather than only ^{12}C atoms becomes significant. Indeed, in Kind and Fiehn (2007) it is argued that incorporating isotope information is crucial for uniquely identifying particles above 186 Da (as opposed to identifying based on mass alone).

It is reasonably straightforward to add isotope information to our model. We can treat the isotopes as compounds with a connection to the monoisotopic version of the compound. We are not limited

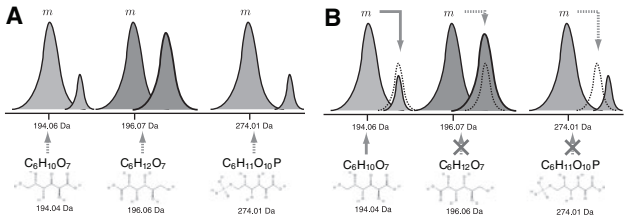

Fig. 5. Schematic representation of the incorporation of isotope intensity information. (**A**) Assignments are made based on the monoisotopic peaks. (**B**) For the first peak, the potential isotope peak has correct mass and intensity making the assignment of the monoisotopic peak more likely. In the second two examples, the isotope peaks have either the wrong intensity (second example) or wrong mass (third example). In these cases, they do not make the monoisotopic assignment more likely and the initial identification should be reconsidered.

to the number and type of isotope peaks that we can include in our potential compound list and we would anticipate that the number to include for each monoisotopic peak would be determined according to the theoretical isotope distribution and the intensity accuracy of the mass spectrometry device. The same argument holds for other types of non-biochemical relationships between peaks in a mass spectrum: for instance, several types of adducts, with characteristic mass difference, are very commonly observed. These could be treated in the same way as isotope peaks in our model.

It is important to note, however, that isotope peaks differ from metabolically related compounds or adduct peaks in two important respects: if one detects an isotope peak for a metabolite, the corresponding monoisotopic peak *must* also be present, and it must have the correct *intensity*. Isotope peaks, therefore, should have more influence in the sampling distribution as identification of such a compound provides very strong evidence for discovery of the monoisotopic version. For this reason, we incorporate the isotope information as a separate term in our model. Our conditional prior then becomes

$$\frac{\beta_{cm}+\delta}{N\delta+\sum_{c'}\beta_{c'm}} \times p(z_{cm}=1|\text{isotopes}).$$

To incorporate intensity information, we can weight isotope connections according to some function of their theoretical intensity and their measured intensity. This weighting must be positive and could for example be a likelihood value from some suitable intensity noise distribution. In this work, however, we will use the most simple option—weights will be 0 or 1 depending on whether or not the measured intensity value is within some fixed tolerance of the expected intensity (we arbitrarily use 10% here—in a real setting, this value could be tuned to reflect the equipment accuracy). To clarify, we will use the following form

$$p(z_{cm}=1|\text{isotopes})=\frac{\omega_{cm}+\delta}{N\delta+\sum_{c'}\omega_{cm}},$$

where ω_{cm} is the count of the number of currently assigned peaks that could have this compound as an isotope peak (or isotope peaks that could have this peak as their monoisotopic form) that satisfy the necessary intensity requirement. We illustrate the method of adding isotopes in Figure 5.

We will demonstrate the approach using 100 random biomolecules, for which we generate synthetic observed peak

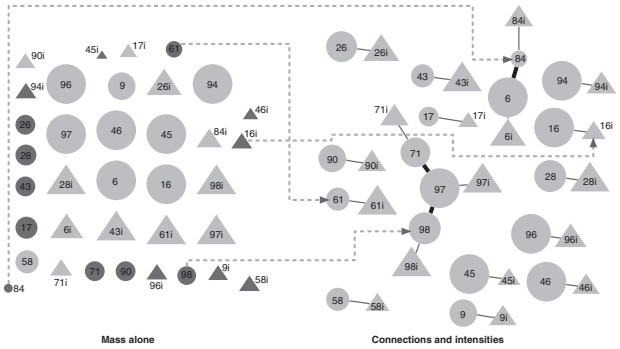

Fig. 6. Example masses from the isotope example. Circles correspond to monoisotopic compounds, triangles to ^{13}C isotope peaks. Color indicates whether or not the peak is correctly assigned, and size corresponds to the confidence of the assignment (posterior probability). Hence, a large light (green) object is correctly assigned with high confidence, and large dark (red) objects are confidently, but incorrectly assigned. The left panel shows assignments based only on mass, the right hand side uses connections (thick lines) and isotope peak intensities (thin lines). The dotted lines connect equivalent masses in the two panels.

masses by adding noise to their calculated accurate mass. We also generate potential ^{13}C isotope peaks for these 100 compounds by computing monoisotopic intensities from a Gamma(1, 1) distribution and then computing the intensity of the single ^{13}C isotope. If this intensity is above some pre-determined threshold (this would typically be defined by the accuracy of the equipment), this peak is included and its 'observed' mass is sampled by adding noise to the theoretical mass. Note that in this example, for illustrative purposes, we only generate one isotope peak for each compound. Additional isotopes could be added in the same manner. Seventy-nine of the 100 potential isotopes survive this threshold test.

We compare three approaches on this new dataset: mass alone, the sampler with connections (including isotope connections) and the sampler with connections and intensities. As we know what the correct assignment for each peak should be, we can compare the number of peaks that each method assigns correctly/incorrectly. Computing assignments based on mass alone results in 26 incorrect assignments (10 of which were isotope peaks). When connections are incorporated, this figure drops to 15 (six isotopes) and with intensity information included, this drops further to 11 (four isotopes), an improvement of 58%.

Of the 11 mistakes that the full isotope method still makes, the majority (10) are also made by both alternative methods, whilst one is also made by the method with connections. Hence, the more complex model is improving upon the simpler methods without introducing new mistakes of its own.

In Figure 6, we can see a subset of compounds from this example assigned by mass alone (left) and by the full model (right). Major improvements are immediately evident; even high-confidence misidentifications are corrected by the consideration of isotope information.

6 NETWORK INFORMATION AND FOLLOW-UP EXPERIMENTATION

Until this point, we have been primarily interested in assigning measured masses to empirical formulas and have seen that the

performance in this task can be greatly improved through the use of putative chemical transformations and isotope information. However, with this method, we obtain much more than just better assignment performance as we are also able to recreate useful biochemical networks. For example, the network shown in Figure 6 provides more than just assignments of mass peaks to empirical formulas—in addition it allows us to unravel the complex relationships between molecules present in the biological sample, which can then be examined for their biological or chemical relevance. Computing the posterior distribution over network topologies is straightforward from the samples drawn by our Gibbs sampling procedure.

Regarding isotopes, not only do they allow us to better identify masses, we automatically identify connections between each monoisotopic compound and its isotopes. Depending on the intended downstream analysis, it would then be trivial to use these connections to remove the isotope peaks before further study.

Having created a putative assignment of masses to empirical formulas, the experimenter may wish to validate some of the assignments and identify the corresponding structure using tandem mass spectrometry techniques (MS–MS or MS^n). Our method makes it easy to identify the most informative candidates for such follow-up experimentation—one could simply pick the compounds for which the assignment probabilities change the most when sampler results are compared to the assignments based only on mass. Alternatively, one could choose the most ambiguous or select the candidate that, if it were assigned exactly, would cause the largest increase in certainty over the remainder of the dataset. Selections based on any of these criteria can easily be made if one has access to the full collection of Gibbs samples. Such aid in directing further experimentation would not be available from analysis of mass alone.

7 CONCLUSIONS

In this report, we have introduced a probabilistic approach for the identification of metabolites from measured masses and a list of possible empirical formulas. The approach uses a prior distribution defined over a complete assignment of masses to formulas that uses information regarding putative biochemical connectivity between compounds. We have evaluated the performance of the approach on realistically simulated metabolomics mass spectra and on a real dataset from *T.brucei*. In addition to incorporating common biochemical transformations, we have extended the model to include isotope-specific information.

We have shown that for realistic metabolic networks, using connectivity information greatly improves our ability to reliably assign formulas to mass peaks. In addition to this, the posterior distribution reflects dependencies between the measured masses, which can be used to infer the hypothetical metabolic network structure and to join monoisotopic peaks to their isotope peaks. The lack of gold standard metabolomic datasets makes a realistic in-depth performance assessment impossible. However, as the field of metabolomics grows it is anticipated that such data will become available and more rigorous evaluation will be possible.

The inferred network can provide useful biological insights. Correctly assigned empirical formulas are essential for efficient database searches for further annotation of the results. The inferred network organizes the data in an accessible way for further manual exploration. It also contains numerous implicit hypotheses about potential enzymatic reactions and the metabolic context of particular measured masses. If a system was measured in two different conditions, comparing the different network topologies would identify major reorganizations of the metabolome. Furthermore, using the full posterior of our model can also help directing further experimentation as there are principled ways of choosing which masses to focus on for follow-up validation and identification, for example by applying tandem mass spectrometry, as described in the previous section.

There are several possible extensions to this work. The simple way in which the model is built up, and the conditional nature of the prior, means that incorporating additional forms of information is straightforward. For example, there may be evidence from previous experiments or the literature to indicate that specific compounds are more likely to be detected in a particular experiment. This prior information could easily be added as a multiplicative term to the prior that is currently used . We could also have more comprehensive intensity information (maybe over a time course) or information from multiple-related samples, which can be used to support or refute the identification of masses (e.g. correlated compounds may be more likely to be connected). As long as we can construct a suitable discrete distribution, this can simply be added as an additional term in the prior.

Our probabilistic approach to assigning empirical formulas is an important first step toward a more detailed and informative analysis of metabolomics experiments. The flexible structure of the underlying model will allow rapid incorporation of additional levels of information to provide a very powerful and general metabolomics annotation tool.

Funding: Engineering and Physical Sciences Research council (EP/E052029/1 to S.R., M.G.); Vidi grant from the Netherlands Organisation for Scientific Research (to R.B.).

Conflict of Interest: none declared.

REFERENCES

Breitling,R. *et al*. (2006a) Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics*, **2**, 155–164.

Breitling,R. *et al*. (2006b) Precision mapping of the metabolome. *Trends Biotechnol.*, **24**, 543–548.

Dunn,W. *et al*. (2008) Metabolic profiling of serum using ultra performance liquid chromatography and the LTQ-orbitrap mass spectrometry system. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **871**, 288–298.

Hegeman,A.D. *et al*. (2007) Stable isotope assisted assignment of elemental compositions for metabolomics. *Anal. Chem.*, **79**, 6912–6921.

Kind,T. and 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.

Kind,T. and Fiehn,O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

Koch,B.P. *et al*. (2007) Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Anal. Chem.*, **79**, 1758–1763.

Kujawinski,E.B. and Behn,M.D. (2006) Automated analysis of electrospray ionization fourier transform ion cyclotron resonance mass spectra of natural organic matter. *Anal. Chem.*, **78**, 4363–4373.

Lu,W. *et al*. (2008) Analytical strategies for LC-MS-based targeted metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **871**, 236–242.

Makarov,A. *et al*. (2006) Dynamic range of mass accuracy in LTQ-Orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.*, **17**, 977–982.

Sanguinetti,G. *et al*. (2008) MMG: a probabilistic tool to identify submodules of metabolic pathways. *Bioinformatics*, **24**, 1078–1084.