Date of acceptance Grade

Instructor

Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm

Markus Heinonen

Helsinki March 20, 2007

UNIVERSITY OF HELSINKI
Department of Computer Science

Contents

1 Introduction

The importance of reliable data is increasing as our models and theories are advanced to gain access into complex models. Systems biology is a new field studying the interactions of biological components and models containing several levels of biological insight [Fie01]. Ever more complex models place a progressive demand for high quality data. Reliability, robustness and repeatability are essential for experimental data acquisition.

One of the corner stones in the measurement technology in systems biology is mass spectrometry (MS), which is a methodology to measure accurately fine amounts of substances in a sample. The produced spectrum shows the compounds and relative abundancies of them.

Recently a paper was published by Kind et al. which studies the issue of annotating and identifying metabolites with MS [KF06]. This study aims to present the relevant computational methods of identifying compounds from a biological sample using spectroscopic methods.

In section 2 the reader is introduced to the necessary background in MS and theory of atoms with the focus on measurements of biological samples. In section 3 the methods of identification of metabolites from MS data is discussed in detail. Section 4 concludes this work.

2 Chemical background

The problems of analyzing biological samples and measuring their chemical compositions inherently requires some understanding of the relevant physical and chemical aspects. The mass spectrometer is an effective and accurate device capable of producing data in high-throughput manner from chemical samples [dH96]. MS enables measurements of whole cell's state and the mixture of compounds within. Ideally we want to measure the state of a single cell but in practise measurements over a culture of cells in different states are done.

MS measures the mass-to-charge ratio of ions. An inserted compound is ionized and its mass-to-charge measured. The amount of each compound in the sample is usually in the range of millions of molecules. The ionization process ionizes compounds with different amounts of charge. Superior amount of the compound is ionized to a

charge of one, which allows us to simplify the mass-to-charge to just mass without significant loss of evidence. In the case of a sample with mixture of compounds, all compounds can be measured at the same time. It is not uncommon for several molecules with the same elemental (atomic) composition to overlap in the spectrum and contribute to a single peak. Using a high-accuracy MS with low error (denoted usually in parts per million) helps distinguish molecules with close masses.

Molecular compounds, e.g. amino acids and metabolites consists of various amounts of elements. The elements are made of protons, neutrons and electrons and constitute atoms. E.g. the carbon atom (C) has 6 protons, 6 neutrons and 12 electrons, making it a charge-neutral and having a nominal (integral) mass of 12 Daltons (Da). One Da is defined as one twelfth of carbon's mass.

Carbon's exact integral mass in an exception: all other elements have a non-integral exact mass. E.g. oxygen has a nominal mass of 16 but exact mass of 15.9949146. Note that several molecules of different elemental composition can still have the same mass with respect to a certain accuracy. E.g. $CH_2N_2O_9$ has mass of 185.976030 and $C_4H_{11}PS_3$ has a mass of 185.976048, a difference of 1.92e-05. In this study we usually only consider the most common elements of metabolites, namely carbon, hydrogen, nitrogen, oxygen, phosphor and sulfur (CHNOPS). This is because metabolites are mostly of these elements. The range of metabolite's sizes is mostly well restricted. 96.5% of all molecules in the KEGG LIGAND database [KG00], which is a database of metabolites and organic compounds, are smaller than 1000 Da.

An elemental specie can hold differing amounts of neutrons in its nucleus, thus producing *isotopes* of the same element with otherwise the same chemical properties. We denote isotopes with a pair of nominal mass and natural abundance. Carbon has 2 naturally occurring isotopes, (C^{12}, 98.89%) and (C^{13}, 1.11%). Sulphur has four: (S^{32}, 95.02%), (S^{33}, 0.75%), (S^{34}, 4.21%) and (S^{36}, 0.02%). (S^{35}, 0%) doesn't occur in nature. The isotope with zero extra neutrons is called *monoisotopic* and sometimes denoted "+0" peak in the spectrum. It's mass is denoted by M_0. Respective to the focus of this study, higher isotopes quickly diminish in intensity. For a molecule C_{50} of 600 Da the +5, +6, ... sum up to less than 0.0002.

The isotopic pattern is clearly visible from the spectrum. Each molecule has its own unique isotope distribution coming directly from the elemental composition. This is highly valuable information in identification of molecule solely from it's M_0 peak and isotopic pattern.

3 Identification of metabolites

The spectrum only provides a peak pattern which corresponds to a molecule. The isotopic pattern contains information about the molecule's elemental composition, as does the mass M_0. To identify a molecule based on it's deduced elemental composition is impossible task with purely computational means if there exists several plausible molecular structures for a certain peak pattern. This is most often the case.

A necessary step in identification of metabolites is using an metabolic database. Databases like KEGG [KG00], PubChem [NIH07], Dictionary of Natural Products (DNP) [CHE07] and Chemical Abstracts (CAS) [CAS07] contain known molecular structures. KEGG concentrates on naturally occurring compounds and metabolites, while e.g. CAS is the largest compound database, but also contains various artificially synthesized molecules.

Large amount of metabolites are not included in databases because of cell's dynamic metabolism and the almost limitless range of molecules composed of different structural motifs. Secondly for a certain mass various molecules of different molecular formulas can be found. Even if a single molecular formula can be found, there still might exists several structural *isomers*, i.e. molecules with same molecular formula but different structure. Analysis of isomers usually requires a chemist to manually identify chemically plausible and the most likely structures. Information about organism's pathways, biochemistry of cell and experimental setting might be utilized.

A sound strategy is to compute all chemically plausible *candidate molecules* and restrict their amount with context-dependent information and using isotopic patterns (See schematics of Figure 1). Computation of possible atomic compositions (molecular formulas) for a mass M_0 is discussed alongside with chemical constraints on the possible molecular formulas. Finally the issue of isotopic pattern matching is discussed in detail.

This usually results in a small set of highly scoring elemental compositions [KF06], which can then be queried from databases or studied manually to identify the structure of the peak. This approach can be extended by utilizing prior information about organism's biochemical pathways. Another promising extension is to match fragmentation patterns of the observed and candidate molecules with MS/MS databases [JS04, BCE+00] or to use fragmentation patterns directly in structural elucidation [HRM+06].

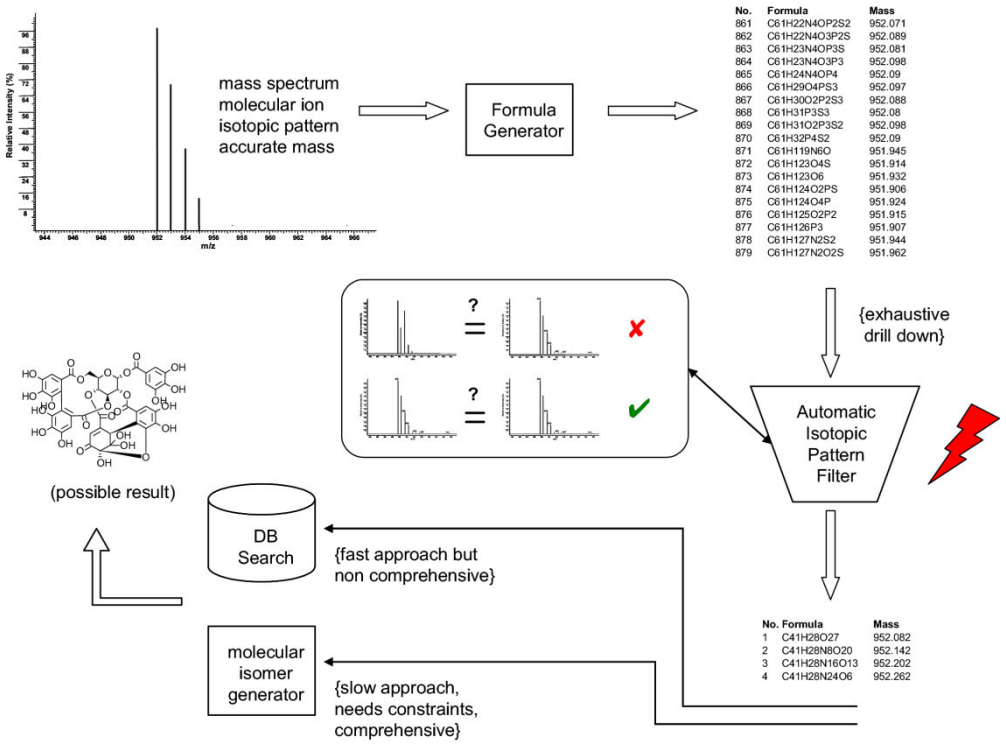

Figure 1: A schematics of identifying a molecule from spectral pattern. Figure reproduced from [KF06].

The resulting set of candidate molecules depends on the MS's accuracy of mass determination. The mass accuracy of MS is denoted as *mass accuracy error*, which is defined as $\frac{|mass_{obs} - mass_{true}|}{mass_{true}} \cdot 10^6$ parts per million. Modern MS devices are capable of 1-5 ppm, with state of the art devices achieving even 0.1 ppm accuracies. Unfortunately even this level of accuracy is not enough to determine a unique elemental composition on most metabolites. Using 0.1 ppm accuracy, a total of 5 elemental compositions are found for a molecule of mass 600 Da. The smallest molecule which has a non-unique elemental composition has a mass of only 185.9760 Da with 0.1 ppm accuracy. For this particular molecule there exists thousands of valid chemical structures. Thus the problem can't be solved by increasing the accuracy of MS in the foreseeable future [Bal04].

The amount of atomic compositions for any mass is exponential. For most metabolites exhaustive methods can still be applied with success, but e.g. protein analysis relies of heuristic methods.

3.1 Computation of atomic compositions

To determine the atomic composition of a molecule based on its mass can be formalized as computing the decompositions of mass m over the individual elements $a_1, a_2, \ldots, a_\sigma$, i.e. we find a non-negative integer linear combination for

$$a_1c_1 + a_2c_2 + \ldots + a_nc_n \in [M_0 - \varepsilon, M_0 + \varepsilon], \tag{1}$$

where M_0 is the monoisotopic mass, a_i are the real-valued masses of elements (one can assume $a_1 < a_2 < \ldots < a_\sigma$), ε denotes the measurement inaccuracy and c is the integer solution vector. This problem is a special case of well known integer knapsack problem [KPP04]. In knapsack problem a value of a set of chosen items is maximized while having the weights of the items below some limit. Atomic composition problem is a special case of knapsack problem called *subset sum problem*, where each item has a value equal to its weight and the set of items is unbounded and corresponds to the atoms. Thus we try to find a subset of unbounded set of atoms that add up to the desired mass. Both problems are known to be NP-hard.

A naive approach is to exhaustively go through all valid combinations of c. This results in exponential number of decompositions. For alphabet CHNOPS there are over $7 \cdot 10^8$ sum formulas below 1000 Da [KF06], which might be feasible depending on the algorithmic implementation and the scope of the analysis.

If we assume integral coefficients, i.e. integral masses, the solution is computed with dynamic programming. Böcker et al. [BL05] proposed an algorithm which requires runtime of $\mathcal{O}(a_1\sigma\tau(M_0))$, where a_1 is the mass of the smallest element, σ is the size of the alphabet and $\tau(m)$ the number of decompositions of M_0. The solution still leaves a lot be desired with a great number of false negatives resulting from integral precision.

Böcker et al [BLLP06] introduced a Dimension Reduction (DR) algorithm for the real-valued mass integer knapsack problem. The solution relies on a new formalism of joint decompositions, or multiple knapsack problem formulation. The results are consistently better than earlier algorithms.

3.2 Filtering the results on chemical knowledge

The numerous candidate molecular formulas can be reduced using chemical knowledge of plausible molecules. There exists several simple rules dictating plausible

molecules based on their molecular formulas, which greatly reduces the search space [KF06].

A well known LEWIS rule dictates that compounds have to account for an even number of electrons with atoms that all obey the octet rule. Octet rule states that atoms usually have eight electrons in their shell. If an atom lacks electrons, it tries to gather them by sharing electrons with its neighbouring atom, i.e. forming a bond.

Valence, a natural property of all atoms, dictates the bonding properties of atoms. A common abstraction of valence is the amount of single bonds it can form. A carbon atom has four bonds in its neutral state, oxygen two, etc. The SENIOR theorem [Sen51, MN03] places common restrictions on the valence properties of a compound. The *degree of unsaturation* rule states that $DU = -\frac{v_1}{2} + \frac{v_3}{3} + v_4 + 1$ [Pel83] is a non-negative integer if all elements are assumed to be on their lowest valency state. Variable v_1 denotes the number of monovalent atoms (H), v_3 trivalent atoms (N,P) and v_4 tetravalent atoms (C). Also unreasonably high or low amounts of hydrogen atoms can also, depending on context, be excluded.

Heavier elements can be ignored from the calculations. The most precise testing can be done if each compound is simulated on structural level, however, this is unfeasible on all but smallest molecules (less than 500 Da).

3.3 Utilizing isotopic information

The *isotope pattern* contains direct information about the molecule's atomic composition. Isotopes can be used to eliminate candidate molecules with invalid isotopic distributions. The isotope pattern is easily retrieved from the spectrum if overlap is ignored. For sake of simplicity we assume that patterns do not overlap.

All isotopic combinations of a compound is represented by expansion of product of polynomials (see Table 1 and 2)

$$(a_1 + a_2 + \ldots)^m (b_1 + b_2 + \ldots)^n (c_1 + c_2 + \ldots)^o \ldots, \tag{2}$$

where a_1, a_2, ...; b_1, b_2, ...; c_1, c_2, ... represents individual isotopes of the elements a, b and c, respectively, and m, n, o represent the number of atoms of corresponding element. Thus e.g. sucrose $C_{12}H_{22}O_{11}$'s polynomial is $(C^{12}+C^{13})^{12}(H^1+H^2)^{22}(O^{16}+O^{17}+O^{18})^{11}$.

The iteration of all permutations of above mentioned polynomial is usually a daunt-

C^{12}	C^{13}	H^1	H^2	O^{16}	O^{17}	O^{18}	nom. mass	mass (Da)	abundance %
12	0	22	0	11	0	0	342	342.116215	84.9204
11	1	22	0	11	0	0	343	343.119570	11.4383
12	0	22	0	10	1	0	343	343.120431	0.3558
12	0	21	1	11	0	0	343	343.122492	0.2803
12	0	22	0	10	0	1	344	344.120460	1.8727
10	2	22	0	11	0	0	344	344.122925	0.7062
11	1	22	0	10	1	0	344	344.123786	0.0479
11	1	21	1	11	0	0	344	344.124647	0.0007
12	0	22	0	9	2	0	344	344.125847	0.0378
12	0	21	1	10	1	0	344	344.126708	0.0012
12	0	20	2	11	0	0	344	344.128769	0.0004

Table 1: Isotope species of sucrose ($C_{12}H_{22}O_{11}$) sorted by mass up till nominal mass 344. Table reproduced from Bocker et al. [BLLP06]

peak	nom. mass	mean mass (Da)	abundance %
+0	342	342.116215	84.9204
+1	343	343.120831	12.0744
+2	344	344.124734	2.6669
...			

Table 2: First three isotopes of sucrose ($C_{12}H_{22}O_{11}$). Isotopes +3, +4, ... contribute 0.3384 % to the isotopic pattern.

ing task. The number of isotopic expansions for the polynomial with n_C carbons, n_H hydrogens, n_N nitrogens, n_O oxygens, n_P phosphor and n_S sulfur is

$$(n_C+1)(n_H+1)(n_N+1)\binom{n_O+2}{2}(1)\binom{n_S+3}{3}. \tag{3}$$

Note that phosphor (P) contains only single naturally occurring isotope and thus doesn't contribute to the number of permutations.

Sucrose has $13 \cdot 23 \cdot \binom{13}{2} = 23322$ isotopic permutations, a number well in the range of exhaustive computation. A sound strategy is to enumerate all isotopic permutations, compute each's abundance and combine the permutations into +1,+2 etc. peaks. Let's first consider the isotopic distribution of a molecule E_l consisting of l identical atoms of element $E \in \{H, C, N, O, P, S\}$. The abundance of a single isotopic specie is

multinomial distribution $X \sim P(n^{+0}, n^{+1}, \ldots; l; r_1, r_2, \ldots)$, with probability function

$$P(n^{+0}, n^{+1}, \ldots; l; r_0, r_1, \ldots) = \frac{l!}{(n^{+0})!(n^{+1})! \ldots} r_0^{n^{+0}} r_1^{n^{+1}} \ldots, \tag{4}$$

where r_i denotes the natural abundance of element E's i'th isotope and $n^{(+j)}$ the count of j'th isotopic element.

The factorials in the equation can get large for medium sized molecules. An immediate optimization method is to calculate the abundancies iteratively [Yer83]. Lets denote with A_i the probability of the i'th permutation using some isotopic permutation (see Table 1).

Now the probability of the first permutation $(l, 0, 0, \ldots)$ is computed with

$$A_0 = P(l, 0, \ldots; l; r_0, r_1, \ldots) = r_0^l. \tag{5}$$

The probabilities of following isotopic species are obtained with

$$A_{i+1} = A_i \cdot \frac{(n_i^{+0})!(n_i^{+1})! \ldots}{(n_{i+1}^{+0})!(n_{i+1}^{+1})! \ldots} r_0^{n_{i+1}^{+0} - n_i^{+0}} r_1^{n_{i+1}^{+1} - n_i^{+1}} \ldots. \tag{6}$$

Note that for elements in $\{H, C, N\}$ the abundance is a simpler binomial distribution. The joint distribution of several elemental species is the product of monoelemental species.

There also exists other methods. Rockwood et al. [RVOS05] has concentrated on applying fast fourier transformation (FFT) on the problem. Later the method was optimized for high resulution [RVOS96] and for large sized molecules [RVO96].

Finally measured isotope pattern is compared to the simulated isotopic distributions. An adequate solution for this problem is to discriminate candidate molecules with root mean square (RMS) analysis. We minimize cost C

$$C = \sum_{i=1}^{k} (f_{obs}(M_i) - f_{sim}(M_i))^2, \tag{7}$$

where $f_{obs}(M_i)$ is the i'th peak's observed intensity and $f_{sim}(M_i)$ is the corresponding simulated intensity. This method only measures differencies in isotope intensities and ignores the questions about isotope's mass error. However, it performs well in most cases (see Figure 2).

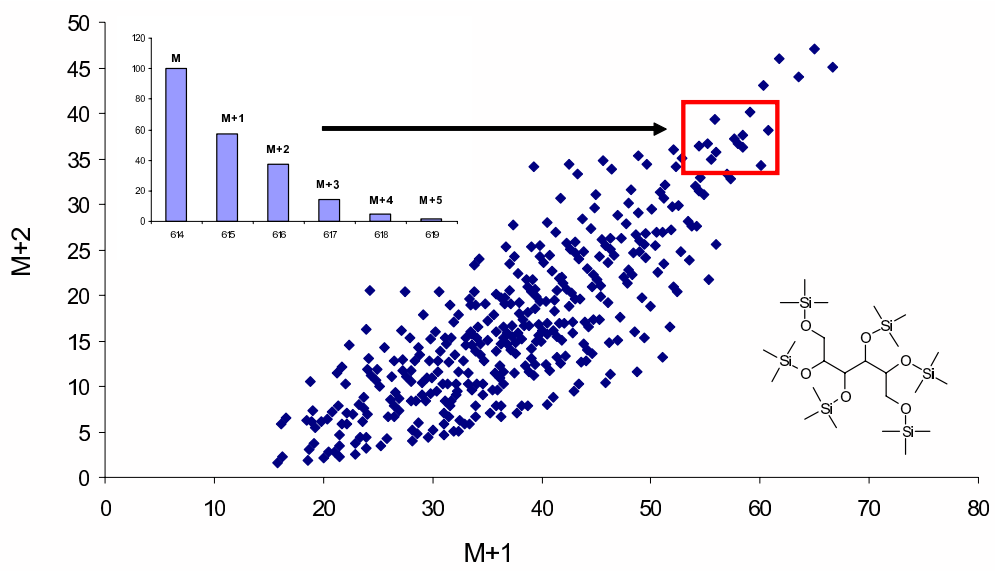

Figure 2: Candidate molecules for a target mass (Silylated sorbitol) are plotted on their isotopic distributions of M+1 and M+2 isotopes. The red square indicates a 5% RMS error area around the target. Figure reproduced from [KF06].

Another approach is to use bayesian statistics in distribution evaluation [ZAS02, ZC00]. A probability is assigned for simulated patterns with Bayes equation

$$P(M_j|D,B) = \frac{P(M_j|B)P(D|M_j,B)}{\sum_i P(M_i|B)P(D|M_i,B)}, \tag{8}$$

where D is the data (observed pattern), M_i are the models (simulated pattern) and B is the prior knowledge. B can be used to exclude impossible or invalid candidate molecules by setting prior $P(M_j|B)$ to zero for desired molecules. Probability of the observed intensity is calculated by $P(D|M,B) = \prod_j P(M_j|m_j) \prod_j P(f_j|p_j)$, where $P(M_j|m_j)$ is the probability to observe peak j at mass M_j with true mass of m_j and $P(f_j|p_j)$ probability to observe peak j with intensity f_j with real intensity of p_j. This formulation takes into account the error in peak position, which follows from mass spectrometry's inaccuracies and also from the fact that different isotopic species contain sligthly different masses even with the same nominal mass (See Table 1).

4 Conclusions

Mass spectrometric data is abundant and the emphasis on the analysis of compounds in data. The annotation of compounds based on low error in peak measurements and isotopic information produces highly specific elucidations for the compounds. The completely automical annotation of compounds is not yet resolved, though. A meticulous care has to be taken for overlapping compounds and isotopes, technical and chemical aspects of experiments and error-ranges of the MS devices.

The various computational methods in this area are mature. The chemical databases also have increased in size and accuracy to keep in pace with systems biology advancements. Contextual information about pathways and feasible potential metabolites adds while for hard cases methods like fragmentation patterns can be utilized.

References

Bal04 Balogh, M., Debating resolution and mass accuracy. *LC GC North America*, 22, page 118.

BCE+00 Baumann, C., Cintora, A., Eichler, M., Lifante, E., Cooke, M., Przyborowska, A. and Halket, J., A library of atmospheric pressure ionization daughter ion mass spectra based on wideband excitation in an ion trap mass spectrometer. *Rapid Communications in Mass Spectrometry*, 14, pages 349–356.

BL05 Bocker, S. and Liptak, Z., Efficient mass decomposition. *Proceedings of ACM Symposium on Applied Computing*, 2005, pages 151–157.

BLLP06 Bocker, S., Letzel, M., Liptak, Z. and Pervukhin, A., Decomposing metabolomic isotope patterns. *Algorithms in Bioinformatics, WABI*, volume 4175 of *Lecture Notes in Computer Science*. Springer, 2006, pages 12–23.

CAS07 CAS, Chemical abstracts database, `http://www.cas.com`, 2007.

CHE07 CHEMnetBASE, Dictionary of natural products, `http://dnp.chemnetbase.com/`, 2007.

dH96 de Hoffmann, E., Tandem mass spectrometry: a primer. *Journal of Mass Spectrometry*, 31, pages 129–137.

Fie01 Fiehn, O., Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2, pages 155–168.

HRM+06 Heinonen, M., Rantanen, A., Mielikäinen, T., Pitkänen, E., Kokkonen, J. and Rousu, J., *Ab Initio* prediction of molecular fragments from tandem mass spectrometry data. *German Conference on Bioinformatics*, volume P-83 of *Lecture Notes in Informatics (LNI)*. GI, 2006, pages 40–53.

JS04 Josephs, J. and Sanders, M., Creation and comparison of ms/ms spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers. *Rapid Communications in Mass Spectrometry*, 18, pages 743–759.

KF06 Kind, T. and Fiehn, O., Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7, pages 234–244.

KG00 Kanehisa, M. and Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28, pages 27–30.

KPP04 Kellerer, H., Pferschy, U. and Pisinger, D., *Knapsack problems*. Springer, 2004.

MN03 Morikawa, T. and Newbol, B., Analogous odd-even parities in mathematics and chemistry. *Chemistry: Bulgarian Journal of Chemical Education*, 12, pages 445–450.

NIH07 NIH, Pubchem, `http://pubchem.ncbi.nlm.nih.gov/`, 2007.

Pel83 Pellegrin, V., Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. *Journal of Chemical Education*, 60, pages 626–633.

RVO96 Rockwood, A. and Van Orden, S., Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry*, 68, pages 2027–2030.

RVOS96 Rockwood, A., Van Orden, S. and Smith, R., Ultrahigh resolution isotope distribution calculation. *Rapid communications in Mass Spectrometry*, 10, pages 54–59.

RVOS05 Rockwood, A., Van Orden, S. and Smith, R., Rapid calculation of isotope distributions. *Analytical Chemistry*, 67, pages 2699–2704.

Sen51 Senior, J., Partitions and their representative graphs. *American Journal of Mathematics*, 73, pages 663–689.

Yer83 Yergey, J., A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, 52, pages 337–349.

ZAS02 Zhang, N., Aebersold, R. and Schwikowski, B., Probid: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, pages 1406–1412.

ZC00 Zhang, N. and Chait, B., Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Analytical Chemistry*, 72, pages 2482–2489.