5 Decoy Databases and False Discovery Rates

> "[Back in 1915] Charlie Chaplin look-alike contest became a popular form of entertainment. At these events, contestants would compete to see who could best imitate the 'tramp' persona championed by Chaplin. [...] According to entertainment folklore, Chaplin himself once entered and lost one of these contests. [...] Charlie Chaplin did not come in second or third, he did not even make the finals." (Mario Cruz)

THE content of the following chapter is a little different from the rest of this book, as it deals with statistics and stochastic of mass spectrometry analysis but not combinatorics. This overview will be short and vastly incomplete: In fact, a complete textbook can be written about the statistical analysis of peptide and protein MS, which has many similarities but also some unique features compared to transcriptomics and microarray analysis, see e.g. Aaant [1]. The reason to include this chapter are twofold: First, **[TODO: THIS IS THE BASIC STUFF, AND EVERYBODY SHOULD UNDERSTAND AT LEAST THIS]**. Second, it must be understood that computational MS only make sense in light of statistics: Computational MS is about "real data" and, as such, full of inaccuracies, errors, misclassifications, and spurious signals. Usually, the best way to deal with these problems is statistics. In the remainder of this book, we will often indicate how to modify, say, a combinatorial algorithm so that results have "statistical meaning". Third, some ideas introduced in this chapter (decoy databases, p-values) can be reused in many other areas of computational mass spectrometry.

5.1 Introduction and data

In the previous chapter, we have described how to match a measured spectrum to a reference spectrum. Again, we focus on the task of identifying a peptide using MS/MS data, and just note that the methods presented here can be applied to similar problems as well. We search our measured spectrum against a database of reference peptide sequences, and we accept the reference spectrum and, hence, the reference peptide with the highest score as being the correct answer. This is called the *best hit* in the database, and the pair "measured spectrum" plus "best hit peptide sequence" is usually referred to as *peptide-spectrum match* (PSM). But the truth is that we often measure spectra that do not stem from peptides and proteins in the analyzed sample: These might be spectra where metabolites, glycans, or lipids are recorded instead of peptides; spectra that do not contain any real biomolecules but only "chemical noise"; or, spectra where we have recorded impurities in sample preparation such as the infamous Keratin.[1] For these spectra, our method will also find a best hit, and this will be called a *spurious* hit in the following.

[1] Keratin is the key structural material making up the outer layer of human skin.

How can we differentiate true hits from spurious hits? Is a score of 120 a good score and, hence, a true hit? We can compare it to other scores but maybe, all of our hits are spurious, and all scores are bad scores. The most reasonable way to deal with this dilemma, is to estimate the *significance* of a hit: Roughly speaking, this is the chance that a hit is spurious. We will introduce to basic concepts on how to compute such significances, namely p-values and q-values.

In a proteomics experiment, we usually do not search for a single spectrum inside the peptide database. Instead, proteomics experiments tend to produce thousands of spectra that all have to be searched in the database, see Sec. 11. So, it is reasonable to process all of these spectra in a *batch*, and then to assign how sure we are about the individual search results. This is made use of in the decoy database strategy. As an example, assume that we have 10 000 spectra that we want search in the peptide database. A possible outcome of our search might be that for 2 000 spectra, we do not assign any peptide; for 8 000 spectra we assign a peptide each, and estimate about 400 of these assignments might be wrong; and for each spectrum, we also give an individual assessment on the quality of the PSM, based on the complete batch of PSMs. In the following, we describe how this can be calculated.

5.2 Decoy databases

We find it very hard to decide if a particular hit is true or spurious. Can we produce a PSM that is necessarily a true hit? Only by changing the MS/MS spectrum, but but this would not make much sense, as the MS/MS spectrum is the measured data we want to interpret. On the other hand, can we produce a PSM that is necessarily spurious? This is much easier, as scoring the measured spectrum against any random peptide sequence, can only result in a hit that is necessarily spurious. In fact, there is a very small chance that the random peptide sequence is the true sequence; we will come back to this later.

We will refer to the peptide database that we use for searching, as the *target database*. The punchline of decoy database searching is: Create a second database, called *decoy database*, which looks similar to the target database, but only contains peptides which cannot be part of the sample. Combine both databases, and search in the combined database. Any PSM with a peptide from the decoy database *must* be spurious.

Now, some MS/MS spectra will result in hits to the decoy database, and can be excluded. But still, there will be many hits to the target database which are spurious, too. We can increase the size of the decoy database, to make it more likely that spurious hits in the decoy database result. But in fact, this is not an option: If the decoy database gets too large, such as 100 times the size of target database, it may happen that a MS/MS spectrum that truly belongs to some peptide in the target database, just by chance looks more similar to another peptide in the decoy database. But even for such a large decoy database, there will still be some spurious hits to the target database. Also, this will significantly increase the size of the peptide database to search in and, hence, result in highly increased running times for searching.

In the following, we want to use PSMs in the decoy database, to estimate spurious hits in the target database. For this, we will use that fact that spectra in a proteomics experiment are usually searched in batches, as described above. This allows us to estimate the significance of one PSM, taking into account all other PSMs in the batch.

As stated above, the decoy database should look “reasonably similar” to the target database while at the same time, all hits in the decoy database should be spurious. In detail, we want the decoy database to meet the following three conditions:

1. There is no overlap between the decoy database and the target database: That is, peptides in the decoy database are not in the target database, and vice versa.
2. The true peptide is not in the decoy database, so that any hit in the decoy database is a spurious hit.
3. A wrong hit in the target database is as probable as a hit in the decoy database.

In practice, it is not necessary that all three conditions are perfectly fulfilled: It is sufficient that the number of exceptions to these conditions is so small, that is does not interfere significantly with our calculations.

5.3 How to create a decoy database

Having talked so much about decoy databases, the first question that comes into mind, is: How do we build one? Different methods for creating a peptide decoy database have been proposed over the years. All start off from the target database either containing full protein sequences, or peptide sequences that have been digested *in silico*, see Sec. 4.1. The most commonly used methods to build a decoy databases are:

Inverted proteins. We invert all target proteins, that is, read them from right to left. Then, we do *in silico* digestion to create the peptide decoy database.

Inverted peptides. We invert all target peptides, generated from the target proteins by *in silico* digestion.

Pseudo-inverted peptides. We invert target peptides but keep the last character in place, so $s = s_1 \dots s_{l-1} s_l$ gets $s_{l-1} \dots s_1 s_l$.

Random iid. We use the target database to estimate the relative frequency of each amino acid. We create a decoy database by, for each peptide of the target database, a random peptide of the same length is created, randomly drawn with the amino acid frequencies estimated above. Each character is drawn independently and with identical distribution (i.i.d.).

Markov chain. Instead of drawing the letters independently, we can learn a Markov chain from the target database, and generate random peptides of identical length distribution as the target database using this Markov chain.

Random iid plus. We learn two distributions from the peptide target database: One for all letters but the last, one for only the last letter of each peptide. We then generate decoy peptides according to these two distributions.

Markov model plus. Similar to Markov chain and Random iid plus.

The "inverted proteins" method inverts each protein in the target protein database, then digests the resulting protein *in silico* to generate the decoy peptides. As we will see below, this method of generating a decoy database has certain shortcomings, and we consider it here merely to show that it is not adequate for what we have in mind. In contrast, the "inverted peptides" and the "pseudo-inverted peptides" methods consider the target *peptide* database, and for every peptide in there, we generate the corresponding decoy peptide. These first three methods are deterministic, as one target database corresponds to exactly one decoy database.

In contrast, the last four methods of building a decoy database are probabilistic: One target database will result in different decoy databases, if we do the computations repeatedly. For these probabilistic methods we first learn the stochastic model of amino acid distributions from the sequences in the database. Then, for every peptide in the target database, we generate a decoy peptide that has the same length as the target peptide, but is generated using the random model.

Two important observations are that target database and decoy database contain exactly the same number of peptides; and that the distribution of peptide lengths in the two databases is identical. In fact, this last observation is not true for the "inverted proteins" decoy database, see Exercise 5.1. Because of this, we will not further look into this method.

Why are the three assumptions that we posed at the beginning of this section, all realized for the six remaining methods? First, we take a look at Assumption 1. In application, this assumption is easy to check: Simply generate the decoy database, and search for overlap. But there are also some theoretical considerations telling us that this overlap can be neglected: We may assume that peptides we search for have some minimal length such as ten amino acids, as other peptides are rather uninformative in application. But there are about $20^{10} = 1.024 \cdot 10^{13}$ peptides of that length — ignoring that we cannot differentiate between leucine and isoleucine, plus that the last position of a peptide is more restricted. If peptides are generated by a random process, than the chance of a peptide being in both databases is negligible, even if there are thousands of peptides of length ten in the target and decoy database. This same arguments carry over to decoy databases made by reversing peptides or proteins, as there is no biological explanation of reversing an amino acid sequence and, so, these decoy databases are "close to random". For longer peptides, chances of an "overlap peptide" further decrease at an exponential rate.

What about Assumption 2? If the true peptide is in the decoy database then, by Assumption 1, it is not in the target database. This means that we have scored a lucky hit: We were searching in a database of chicken proteins and just by chance, the true peptide (which is not from chicken) happens to be in the decoy database. But the amino acid sequences in the decoy database are sort of random, so the chance to find exact the one we have in the sample is really low and can be ignored.

That this assumption holds, the databases have to be same size. We can check this by deleting the best hit in the target database out of the bag of spectra.

There is a problem with the stochastic methods for generating the decoy database, that is the larger, the smaller the target database: It is possible that we have never observed the amino acid, say, alanine in our target database. It is very unlikely that alanine should truly be absent from all proteins and peptides of the organism that we are looking at; it is much more likely that our database is simply "too small". This problem less pronounced for the "random iid" method, slightly pronounced for the "random iid plus" method, and strongly pronounced for the "markov chain" method, see also below.

Luckily, there is **[TODO: PASS OP!]**

While all but one methods for creating a decoy database are easily understandable, The last one is slightly more complex: How do you learn a Markov chain from the sequence database? We do not want to go into the details of Markov theory, but only recall the most important facts. A *Markov chain* is a series of random variables $X_0, X_1, X_2, \ldots$ with the Markov property, namely

$$\mathbb{P}\big(X_{n+1} = x_{n+1} : X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n\big) = \mathbb{P}\big(X_{n+1} = x_{n+1} : X_n = x_n\big)$$

In fact, the Markov chain that we want to come up with, is a particularly simple one: It is *time-homogeneous*, so

$$\mathbb{P}\big(X_{n+1} = y : X_n = x\big) = \mathbb{P}\big(X_1 = y : X_0 = x\big);$$

and, it is *irreducible* so that we can get from any state to any state. In addition, our state space is finite, namely the alphabet Σ of amino acids. Such a Markov model can be described via an *initial distribution* $\pi_0 : \Sigma \to [0,1]$ and a *transition matrix* $P = (p_{i,j})$ with $p_{i,j} = \mathbb{P}(X_1 = j : X_0 = i)$.

5.4 Using the decoy database: False Discovery Rates

We generate a grand database from the target database and the decoy database and look for the bag of measured spectra (at least 1000) in it. This is different from determining p-values, which can be identified for a single spectrum. Here we need a bag of spectra and we can only make a statement about the entirety of spectra.

For each spectrum we estimate the best hit in the grand database and sort the hits by their scores. We get hits from the real and from the decoy database.

Example 5.1. Search in the real database and the decoy database. The results are sorted by score.

peptide #	score	database	peptide #	score	database
37	128.1	target	18	92.0	target
124	122.8	target	69	90.7	decoy
12	121.2	target	72	89.9	target
950	103.1	target	174	87.3	decoy
730	102.3	target	111	86.5	decoy
217	96.4	target	750	86.4	target
918	94.8	target	828	84.2	target
333	94.3	decoy	830	82.3	target
212	93.5	target	13	82.2	target
4	93.4	target	522	80.9	decoy

Among the best $n = 10$ hits in Example 5.1 is only one false positive hit in the decoy database. With Assumption 3 there is also one wrong hit in the real database. So the number of false positives (FP) is two times the number of hits in the decoy database. So in the given Example 5.1 we expect $FP = 2$. The number of true positives (TP) is $TP = n - FP$.

There are two possibilities to measure the quality of our identification. The *precision*

$$precision = \frac{TP}{TP + FP} \tag{5.1}$$

and the *False Discovery Rate* (FDR)

$$FDR = \frac{FP}{TP + FP} \tag{5.2}$$

If we choose a *score threshold* of 93.4 in the given Example 5.1 with $n = 10$ the $FDR = 20\%$ and the $precision = 80\%$. If we choose a score threshold of 80.9 the $FDR = 50\%$.

In practice we choose the FDR at first (e.g. 5%) and look for the minimum score threshold (the maximum n) with a FDR lower the given FDR threshold. We accept all hits in the real database with score ≥ score threshold, so we get a list of reliable identifications with only for example 5% of the identifications in this list are probably wrong.

5.5 Individual False Discovery Rates: q-values and relatives

As mentioned in Sec. 5.4 we only make a statement about the list of reliable identifications, but we want to know the quality of each single hit.

There are three possible solutions: *q-values*, *Posterior Error Probability* (PEP) and *p-values*.

The q-value for a single hit "spectrum ↔ peptide" is the smallest FDR with the hit in the list of reliable identifications. Note that the q-value of a hit depends also on the other identifications. With this definition to assign the q-values, the values get inexact for small $q \sim 0.1\%$.

The PEP is the probability of the incorrectness of a hit. This estimation is very extensive, the parametric distribution of the scores is needed as model and much statistic and stochastic has to be done.

The p-values got a disadvantage: If there are many spectra in a grand database, some hits get small p-values accidentally. But the p-values are really estimated for each measured spectrum and they do not change if some measured spectra are removed.

5.6 Further reading and other approaches

Our presentation of decoy databases follows [68]. Regarding generating the decoy database, the authors do not consider the idea of learning the last letter of the peptide individually, for the random and Markov Model decoy database.

Sashimi project hosts the Trans-Proteomic Pipeline (TPP), see Keller *et al*. [129]. **[ToDo: Citation correct?]** ProHits Liu *et al*. [149].

5.7 Exercises

5.1 Assume that we build a decoy database using the "inverted proteins" method. Explain why we cannot guarantee that the decoy database contain exactly the same number of peptides; or, that the distribution of peptide lengths in the two databases is identical. One protein does the trick.

5.2 Given a target database of proteins

$$\{\mathsf{TVKQDEGHRWTL}, \mathsf{YPPNKCRRDHIKVRAA}, \mathsf{DDCDKPKMN}, \mathsf{FIKTTSRQPRVYYC}, \mathsf{MNMQKWAWAKFIFIRVW}\},$$

build the corresponding peptide decoy databases for methods "inverted proteins", "inverted peptides", and "pseudo-inverted peptides".

5.3 For the target database from the previous exercise, build the "random iid" and "random iid plus" models with pseudocounts.

5.4 For the target database from Exercise 5.2, build the "markov chain" model of order 2 with pseudocounts.

6 Significances, p-values, and E-values

> "The grand assertion is that you must see the world through probability and that probability is the only guide you need." (Dennis Lindley)

THERE is some text missing here. **[ToDo: Pass op!]**

6.1 Introduction and data

[ToDo: What about [74]?]

6.2 A naïve approach for estimating p-values

There exist two direct approaches for assigning a p-value to a score (the "true" score): These are based either on randomizing the data (bootstrapping, resampling), or on randomizing the reference. We will go for the second possibility, as there is no reasonable method known to randomize mass spectrometry data.

Assume that our measured spectrum $\mathcal{M}'$ (the data) was scored highest against reference spectrum $\mathcal{M}^*$ from the database, and reached score score$(\mathcal{M}^*, \mathcal{M}')$ (the true score). To randomize the reference, we have to sample a large number of random reference objects, score each random object against the data, and count the number of times this score is larger or equal to the true score. In detail, let Ω be the space of reference spectra. Randomly choose a reference spectrum $\mathcal{M} \in \Omega$ and score the reference spectrum against the measured spectrum $\mathcal{M}'$, computing the score score$(\mathcal{M}, \mathcal{M}')$. Repeat this between 1000 and 1000000 times, to get reasonable p-values. Count the number of random reference spectra $\mathcal{M}$ with score$(\mathcal{M}, \mathcal{M}') \geq$ score$(\mathcal{M}^*, \mathcal{M}')$. Divide by the number of repetitions, to compute an empirical p-value.

What is a reasonable background model, that is, a reasonable set Ω of reference spectra to chose from? In the old days of computational mass spectrometry, some people proposed to use mass spectra with random peak masses as Ω: Simply draw peak masses at random, for example, uniformly distributed the interval $[0, M]$ where M is the parent mass of the measured spectrum. Here, the number of peaks may be chosen as the average number of peaks of a reference spectra database. Unfortunately, this is a very bad background model: Due to the experimental setup, most of the measured mass spectra will actually correspond to *some* peptide, even though it might not be recorded in the database. Peptide fragmentation spectra have a particular structure that is not covered using randomized peaks. Even if our database hit is spurious, it might share some peaks with the measured spectrum, possibly because a few amino acids at the start or end of the peptide agree with the measured peptide we are searching for. In contrast, randomizing peak masses will make it unlikely to find any peaks that actually match. In total, we will grossly overestimate the actual p-value. This stays true if peak masses are drawn with respect to some empirical distribution computed from, say, a reference database: Peak masses in a peptide fragmentation spectrum are highly correlated, and independently drawing peaks neglects these dependencies.

In MS/MS a peptide has a parent mass M, so peptide strings with the same parent mass are chosen and the reference spectra are generated from them.

Randomly draw a string with mass M: Problem is very similar to counting compomers. **[ToDo: Pass op!]**

But we are interested in very small significances, so there is a huge difference between 10^{-5} and 10^{-10}.

6.3 Parametric Distribution

In this approach it has to be established why the scores follow a known distribution and their parameters have to be estimated.

BLAST (Basic Local Alignment Search Tool) is used to compute heuristically a local alignment of DNA or protein sequences and the score x is converted to a significance. What is the likelihood to get a $score \geq x$ with a random sequence in a random database (of given length) and how many sequences in the database are expected with $score \geq x$ (expectation value E)? The score of the best alignment of two random sequences follows the extreme value distribution, but this so called Karlin-Altschul statistic works not for sequences with gaps. But extensive simulations show, that the distribution of the score of alignments with gaps nearly equals the distribution of the score of alignments without gaps. The parameters of the extreme value distribution were also determined by simulations.

Note that scores are normally distributed in *none* of the cases relevant here. This means that we cannot evaluate results by reporting the "number of standard deviations above the mean," as this implicitly assumes scores to be normally distributed. As the true distribution of scores is usually highly skewed and asymmetric, assuming a normal distribution will result in misleading or usually even wrong conclusions drawn from the data.

The distribution of the scores for the detected spectrum versus 1000 reference spectra has to be identified for 100 measured spectra and compared to known distributions. Now the parameter of the distribution have to be estimated. These result from the moments of the distribution: expectation value, variance, skew, ...

Distribution	parameters	mean	variance	skew
Normal	μ, σ	μ	σ^2	0
Exponential	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	2
Gamma	$k > 0, \theta > 0$	$k\theta$	$k\theta^2$	$2/\sqrt{k}$
Extreme value	μ, $\beta > 0$	$\mu + 0.577\,\beta$	$\frac{1}{6}\pi^2\beta^2$	-1.140

Table 6.1: Mean and central moments of parametric distributions

These moments can be estimated Let $x_1, \ldots, x_n$ be the scores for a detected spectrum versus n reference spectra. The estimator for the expectation value is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6.1}$$

Given a measured spectrum with parent mass M. 100 to 1000 reference sequences with parent mass M have to be generated. For each random reference sequence the reference spectrum has to be generated and aligned with the detected spectrum to get a score. From these

100 to 1000 score values the moments can be estimated and the parameters of the distribution can be calculated. Now for each reference sequence in the database a score can be computed, by simulating the reference spectrum and aligning with the detected spectrum. Let S be the score of the best database hit. The likelihood that this score results accidentally for the calculated parametric distribution can be computed (by *erf* for the normal distribution). Note that the parameters of the distribution have to be calculated only ones for each measured spectrum.

Here are some examples of parametric distributions that have been proposed, for the scoring introduced in the respective papers, over the last years:

year	tool	reference	proposed distribution
2003	PepProbe	Sadygov and Yates III [199]	Hypergeometric distribution
2003	X!Tandem	Fenyö and Beavis [74]	Gumbel distribution
2004	OMSSA	Geer *et al.* [90]	Poisson distribution
2007	RAId_DbS	Alves *et al.* [2]	**[ToDo: Custom?]** distribution
2008	Crux	Park *et al.* [183]	Weibull distribution

6.4 Exact computations using generating functions

We now turn to a method for exact computation of p-values, which has been suggested by Kim, Gupta, and Pevzner [133]. The authors present their method using the mathematical formalism of generating functions. Generating functions allow us to do involved mathematical tricks such as multiplication, division, or taking the derivative of the functions, which usually are infinite series. None of this is required here, so we will use a simpler mathematical formalism based on random variables and the convolution of distributions.

Assume that you are given an ideal die. You will model this stochastically using a discrete random variable $X : \Omega \to \{1,\dots,6\}$ where Ω denotes the sample space (everything that might happen). The probability that a particular value $x \in \{1,\dots,6\}$ is reached, is $\mathbb{P}(X = x) = \frac{1}{6}$, and zero everywhere else. Assume that we have a second die with random variable Y, and we want to model the sum of these two dice. One can easily see that the sum of the dice, $X + Y$, has distribution

$$\mathbb{P}(X+Y=x) = \sum_{y=1,\dots,6} \mathbb{P}(X=x-y)\cdot\mathbb{P}(Y=y). \tag{6.2}$$

This can be generalized beyond dice: For two random variables $X, Y : \Omega \to \mathbb{N}$ we have

$$\mathbb{P}(X+Y=x) = \sum_{y=0,\dots,x} \mathbb{P}(X=x-y)\cdot\mathbb{P}(Y=y). \tag{6.3}$$

and if both random variables have finite support (that is, only a finite set of numbers has probability strictly greater than zero) then this is actually a finite sum.

It is now simple to actually compute these probabilities for $X + Y$: Let $P_X[0\dots x_{\max}]$ be the array with $P_X[x] = \mathbb{P}(X = x)$ and $\sum_{x=0,\dots,x_{\max}} P_X[x] = 1$, and $P_Y[0\dots y_{\max}]$ analogously. Then, we can compute $P_{Y+X}[0\dots x_{\max}+y_{\max}]$ as

$$P_{Y+X}[x] \leftarrow \sum_{x=0,\dots,y_{\max}} P_X[x-y]\cdot P_Y[y] \tag{6.4}$$

where we assume $P_X[x] = 0$ for $x < 0$ and $x > x_{\max}$.

This is all the mathematics that we need in this section. We again over-simplify our problem slightly, to improve readability. To this end, assume that

6.5 Posterior error probabilities

PeptideProphet: Keller et al., Anal. Chem., 2002; Choi et al., J. Proteome Res., 2008

Compute a discriminant score for each PSM (reported by another tool) using multiple features. Bundle PSMs and draw the histogram of scores. Fit the histogram into two distributions (one for false and one for true) using Expectation-Maximization algorithm.

Pros Returns more PSMs Can use of extra features unavailable to database search tools (e.g. distribution of correct PSMs).

Cons Unclear how to determine the two distributions Different distributions are used depending on the database search tools. Discriminant scores are not perfectly normalized. Cannot be used as a stand- alone tool. Requires large number of PSMs Inappropriate low-throughput experiments.

6.6 Further reading and other approaches

The problem of wrongly assuming a score distribution to be normal by reporting the "number of standard deviation above the mean," has already been pointed out by Waterman and Vingron [231] for pairwise sequence alignments.

Sampling random strings of a fixed parent mass was proposed by Lu and Chen [151].

Bibliography

[1] A. Aaant. I need a title, quick. **[TODO: REPLACE WITH A REAL CITATION]**, 2101.

[2] G. Alves, A. Y. Ogurtsov and Y.-K. Yu. RAId_DbS: peptide identification using database searches with realistic statistics. *Biol. Direct.*, 2:25, 2007.

[3] S. Andreotti, G. W. Klau and K. Reinert. Antilope – a lagrangian relaxation approach to the *de novo* peptide sequencing problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2011. To appear, doi:10.1109/TCBB.2011.59.

[4] R. Apweiler, H. Hermjakob and N. Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, 1473(1): 4–8, 1999.

[5] G. Audi, A. Wapstra and C. Thibault. The AME2003 atomic mass evaluation (ii): Tables, graphs, and references. *Nucl. Phys. A*, 729:129–336, 2003.

[6] J.-M. Autebert, J. Berstel and L. Boasson. Context-free languages and pushdown automata. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 111–174. Springer, 1997.

[7] V. Bafna and N. Edwards. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17:S13–S21, 2001.

[8] D. A. Barkauskas and D. M. Rocke. A general-purpose baseline estimation algorithm for spectroscopic data. *Anal. Chim. Acta*, 657(2):191–197, 2010.

[9] C. Bartels. Fast algorithm for peptide sequencing by mass spectrometry. *Biomed. Environ. Mass Spectrom.*, 19:363–368, 1990.

[10] J. M. S. Bartlett and D. Stirling. A short history of the polymerase chain reaction. *Methods Mol. Biol.*, 226:3–6, 2003.

[11] C. Bauer, R. Cramer and J. Schuchhardt. Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods Mol. Biol.*, 696:341–352, 2011.

[12] M. Beck, I. M. Gessel and T. Komatsu. The polynomial part of a restricted partition function related to the Frobenius problem. *Electron. J. Comb.*, 8(1):N7, 2001.

[13] D. E. Beihoffer, J. Hendry, A. Nijenhuis and S. Wagon. Faster algorithms for Frobenius numbers. *Electron. J. Comb.*, 12:R27, 2005.

[14] C. Benecke, T. Grüner, A. Kerber, R. Laue and T. Wieland. MOLecular Structure GENeration with MOLGEN, new features and future developments. *Anal. Chim. Acta*, 314:141–147, 1995.

[15] G. Benson. Composition alignment. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2003)*, volume 2812 of *Lect. Notes Comput. Sc.*, pages 447–461. Springer, 2003.

[16] M. W. Bern and D. Goldberg. EigenMS: De novo analysis of peptide tandem mass spectra by spectral graph partitioning. In *Proc. of Research in Computational Molecular Biology (RECOMB 2005)*, volume 3500 of *Lect. Notes Comput. Sc.*, pages 357–372. Springer, 2005.

[17] M. W. Bern and D. Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J. Comput. Biol.*, 13(2):364–378, 2006.

[18] A. Bertsch, A. Leinenbach, A. Pervukhin, M. Lubeck, R. Hartmer, C. Baessmann, Y. A. Elnakady, R. Müller, S. Böcker, C. G. Huber, and O. Kohlbacher. De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, 30(21):3736–3747, 2009.

[19] K. Biemann, C. Cone and B. R. Webster. Computer-aided interpretation of high-resolution mass spectra. II. Amino acid sequence of peptides. *J. Am. Chem. Soc.*, 88(11):2597–2598, 1966.

[20] K. Biemann, C. Cone, B. R. Webster and G. P. Arsenault. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.*, 88(23):5598–5606, 1966.

[21] A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. Fourier meets Möbius: fast subset convolution. In *Proc. of ACM Symposium on Theory of Computing (STOC 2007)*, pages 67–74. ACM Press New York, 2007.

[22] N. Blow. Glycobiology: A spoonful of sugar. *Nature*, 457(7229):617–620, 2009.

[23] S. Böcker. Sequencing from compomers: Using mass spectrometry for DNA de-novo sequencing of 200+ nt. *J. Comput. Biol.*, 11(6):1110–1134, 2004.

[24] S. Böcker and Zs. Lipták. A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, 48(4):413–432, 2007.

[25] S. Böcker and V. Mäkinen. Combinatorial approaches for mass spectra recalibration. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5(1):91–100, 2008.

[26] S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[27] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lect. Notes Comput. Sc.*, pages 12–23. Springer, 2006.

[28] S. Böcker, B. Kehr and F. Rasche. Determination of glycan structure from tandem mass spectra. In *Proc. of Computing and Combinatorics Conference (COCOON 2009)*, volume 5609 of *Lect. Notes Comput. Sc.*, pages 258–267. Springer, 2009.

[29] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[30] S. Böcker, F. Rasche and T. Steijger. Annotating fragmentation patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2009)*, volume 5724 of *Lect. Notes Comput. Sc.*, pages 13–24. Springer, 2009.

[31] A. Brauer and J. E. Shockley. On a problem of Frobenius. *J. Reine Angew. Math.*, 211: 215–220, 1962.

[32] R. Breitling, A. R. Pitt and M. P. Barrett. Precision mapping of the metabolome. *Trends Biotechnol.*, 24(12):543–548, 2006.

[33] K. Q. Brown. *Geometric transforms for fast geometric algorithms*. Report cmucs-80-101, Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, USA, 1980.

[34] S. Cappadona, P. Nanni, M. Benevento, F. Levander, P. Versura, A. Roda, S. Cerutti, and L. Pattini. Improved label-free LC-MS analysis by wavelet-based noise rejection. *J Biomed Biotechnol*, 2010:131505, 2010.

[35] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell and S. M. Haslam. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.*, 7 (4):1650–1659, 2008.

[36] D. C. Chamrad, G. Körting, K. Stühler, H. E. Meyer, J. Klose and M. Blüggel. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4:619–628, 2004.

[37] S. Chattopadhyay and P. Das. The K-dense corridor problems. *Pattern Recogn. Lett.*, 11 (7):463–469, 1990.

[38] E. Check. Proteomics and cancer: Running before we can walk? *Nature*, 429:496–497, 2004.

[39] T. Chen, M.-Y. Kao, M. Tepel, J. Rush and G. M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, 8(3):325–337, 2001. Preliminary version in *Proc. of Symposium on Discrete Algorithms (SODA 2000)*, Association for Computing Machinery, 2000, 389–398.

[40] W. L. Chen. Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.*, 46(6): 2230–2255, 2006.

[41] F. Y. Chin, C. A. Wang and F. L. Wang. Maximum stabbing line in 2D plane. In *Proc. of Conf. on Computing and Combinatorics (COCOON 1999)*, volume 1627 of *Lect. Notes Comput. Sc.*, pages 379–388. Springer, 1999.

[42] H. H. Chou, H. Takematsu, S. Diaz, J. Iber, E. Nickerson, K. L. Wright, E. A. Muchmore, D. L. Nelson, S. T. Warren, and A. Varki. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl. Acad. Sci. U. S. A.*, 95(20):11751–11756, 1998.

[43] Y. Chu and T. Liu. On the shortest arborescence of a directed graph. *Sci. Sinica*, 14: 1396–1400, 1965.

[44] K. R. Clauser, P. Baker and A. L. Burlingame. Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, 71(14):2871–2882, 1999.

[45] C. A. Cooper, E. Gasteiger and N. H. Packer. GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, 1(2):340–349, 2001.

[46] C. A. Cooper, H. J. Joshi, M. J. Harrison, M. R. Wilkins and N. H. Packer. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.*, 31(1):511–513, 2003.

[47] R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[48] V. Dančík, T. A. Addona, K. R. Clauser, J. E. Vath and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry: A graph-theoretical approach. *J. Comput. Biol.*, 6(3/4):327–342, 1999. Preliminary version in *Proc. of Research in Computational Molecular Biology (RECOMB 1999)*, 135–144.

[49] C. Dass. *Principles and practice of biological mass spectrometry*. John Wiley and Sons, 2001.

[50] R. Datta and M. W. Bern. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. *J. Comput. Biol.*, 16(8):1169–1182, 2009.

[51] J. L. Davison. On the linear diophantine problem of Frobenius. *J. Number Theory*, 48(3): 353–363, 1994.

[52] M. de Berg, M. van Kreveld, M. Overmars and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, second edition, 2000.

[53] E. de Hoffmann and V. Stroobant. *Mass Spectrometry: Principles and Applications*. Wiley-Interscience, third edition, 2007.

[54] J. R. de Laeter, J. K. Böhlke, P. D. Biévre, H. Hidaka, H. S. Peiser, K. J. R. Rosman and P. D. P. Taylor. Atomic weights of the elements. Review 2000 (IUPAC technical report). *Pure Appl. Chem.*, 75(6):683–800, 2003.

[55] E. W. Deutsch, H. Lam and R. Aebersold. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*, 33:18–25, 2008.

[56] P. A. DiMaggio and C. A. Floudas. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal. Chem.*, 79(4):1433–1446, 2007.

[57] B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312:212–217, 2006.

[58] B. Domon and C. E. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate J.*, 5:397–409, 1988.

[59] R. Dondi, G. Fertin and S. Vialette. Complexity issues in vertex-colored graph pattern matching. *J. Discrete Algorithms*, 2010. In press, doi:10.1016/j.jda.2010.09.002.

[60] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.

[61] S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1(3):195–207, 1972.

[62] M. Dyer. Approximate counting by dynamic programming. In *Proc. of Symposium on Theory of Computing (STOC 2003)*, pages 693–699, 2003.

[63] S. R. Eddy. "antedisciplinary" science. *PLoS Comput. Biol.*, 1(1):e6, 2005.

[64] P. Edman. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.*, 4:283–293, 1950.

[65] J. Edmonds. Optimum branchings. *J. Res. Nat. Bur. Stand.*, 71B:233–240, 1967.

[66] M. Ehrich, S. Böcker and D. van den Boom. Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. *Nucleic Acids Res.*, 33(4):e38, 2005.

[67] D. Einstein, D. Lichtblau, A. Strzebonski and S. Wagon. Frobenius numbers by lattice point enumeration. *INTEGERS*, 7(1):#A15, 2007.

[68] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, 2007.

[69] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth and S. P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.*, 22(2):214–219, 2004.

[70] J. K. Eng, A. L. McCormack and J. R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectr.*, 5:976–989, 1994.

[71] M. Ethier, J. A. Saba, M. Spearman, O. Krokhin, M. Butler, W. Ens, K. G. Standing, and H. Perreault. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(24):2713–2720, 2003.

[72] M. Fellows, G. Fertin, D. Hermelin and S. Vialette. Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *Proc. of International Colloquium on Automata, Languages and Programming (ICALP 2007)*, volume 4596 of *Lect. Notes Comput. Sc.*, pages 340–351. Springer, 2007.

[73] J. Fenn, M. Mann, C. Meng, S. Wong and C. Whitehouse. Electrospray ionisation for mass spectrometry of large biomolecules. *Science*, 246:64–71, 1989.

[74] D. Fenyö and R. C. Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, 75(4):768–774, 2003.

[75] J. Fernández-de-Cossío, L. J. Gonzalez and V. Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Comput. Appl. Biosci.*, 11(4):427–434, 1995.

[76] J. Fernández-de-Cossío, J. Gonzalez, T. Takao, Y. Shimonishi, G. Padron and V. Besada. A software program for the rapid sequence analysis of unknown peptides involving modifications, based on MS/MS data. In *ASMS Conf. on Mass Spectrometry and Allied Topics, Slot 074*, 1997.

[77] J. Fernández-de-Cossío, L. J. Gonzalez, Y. Satomi, L. Betancourt, Y. Ramos, V. Huerta, A. Amaro, V. Besada, G. Padron, N. Minamino, and T. Takao. Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. *Nucleic Acids Res.*, 32(Web Server issue):W674–W678, 2004.

[78] A. R. Fernie, R. N. Trethewey, A. J. Krotzky and L. Willmitzer. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, 5(9):763–769, 2004.

[79] H. I. Field, D. Fenyö and R. C. Beavis. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, 2(1):36–47, 2002.

[80] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.*, 77(22):7265–7273, 2005.

[81] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. Freely available from `http://algo.inria.fr/flajolet/Publications/book.pdf`.

[82] A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 15:964–973, 2005.

[83] A. M. Frank, M. M. Savitski, M. N. Nielsen, R. A. Zubarev and P. A. Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.*, 6(1):114–123, 2007.

[84] A. Fürst, J.-T. Clerc and E. Pretsch. A computer program for the computation of the molecular formula. *Chemom. Intell. Lab. Syst.*, 5:329–334, 1989.

[85] V. A. Fusaro, D. R. Mani, J. P. Mesirov and S. A. Carr. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.*, 27(2):190–198, 2009.

[86] H. Gabow, Z. Galil, T. Spencer and R. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6:109–122, 1986.

[87] M. R. Garey and D. S. Johnson. *Computers and Intractability (A Guide to Theory of NP-Completeness)*. Freeman, New York, 1979.

[88] J. Gasteiger, W. Hanebeck and K.-P. Schulz. Prediction of mass spectra from structural information. *J. Chem. Inf. Comput. Sci.*, 32(4):264–271, 1992.

[89] S. P. Gaucher, J. Morrow and J. A. Leary. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal. Chem.*, 72(11):2331–2336, 2000.

[90] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3:958–964, 2004.

[91] P. Gilmore and R. Gomory. Multi-stage cutting stock problems of two and more dimensions. *Oper. Res.*, 13(1):94–120, 1965.

[92] D. Goldberg, M. Sutton-Smith, J. Paulson and A. Dell. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics*, 5(4):865–875, 2005.

[93] D. Goldberg, M. W. Bern, B. Li and C. B. Lebrilla. Automatic determination of O-glycan structure from fragmentation spectra. *J. Proteome Res.*, 5(6):1429–1434, 2006.

[94] D. Goldberg, M. W. Bern, S. Parry, M. Sutton-Smith, M. Panico, H. R. Morris and A. Dell. Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res.*, 6(10):3995–4005, 2007.

[95] D. Goldberg, M. W. Bern, S. J. North, S. M. Haslam and A. Dell. Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics*, 25(3):365–371, 2009.

[96] A. H. Grange, M. C. Zumwalt and G. W. Sovocool. Determination of ion and neutral loss compositions and deconvolution of product ion mass spectra using an orthogonal acceleration time-of-flight mass spectrometer and an ion correlation program. *Rapid Commun. Mass Spectrom.*, 20(2):89–102, 2006.

[97] N. A. Gray. Applications of artificial intelligence for organic chemistry: Analysis of C-13 spectra. *Artificial Intelligence*, 22(1):1–21, 1984.

[98] N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White, and L. Creary. Computerized mass spectrum prediction and ranking. *Anal. Chem.*, 52(7):1095–1102, 1980.

[99] N. A. B. Gray, A. Buchs, D. H. Smith and C. Djerassi. Computer assisted structural interpretation of mass spectral data. *Helv. Chim. Acta*, 64(2):458–470, 1981.

[100] H. Greenberg. Solution to a linear diophantine equation for nonnegative integers. *J. Algorithms*, 9(3):343–353, 1988.

[101] D. H. Greene and D. E. Knuth. *Mathematics for the Analysis of Algorithms*, volume 1 of *Progress in Computer Science and Applied Logic (PCS)*. Birkhäuser Boston, 1990.

[102] J. Gross. *Mass Spectrometry: A textbook*. Springer, Berlin, 2004.

[103] K. Grützmann, S. Böcker and S. Schuster. Combinatorics of aliphatic amino acids. *Naturwissenschaften*, 98(1):79–86, 2011.

[104] M. Guilhaus. Principles and instrumentation in time-of-flight mass spectrometry. *J. Mass Spectrom.*, 30:1519–1532, 1995.

[105] S. Guillemot and F. Sikora. Finding and counting vertex-colored subtrees. In *Proc. of Symposium on Mathematical Foundations of Computer Science (MFCS 2010)*, volume 6281 of *Lect. Notes Comput. Sc.*, pages 405–416. Springer, 2010.

[106] C. Hamm, W. Wilson and D. Harvan. Peptide sequencing program. *Comput. Appl. Biosci.*, 2:115–118, 1986.

[107] F. Harary, R. W. Robinson and A. J. Schwenk. Twenty-step algorithm for determining the asymptotic number of trees of various species. *J. Austral. Math. Soc.*, 20(Series A): 483–503, 1975.

[108] M. Havilio, Y. Haddad and Z. Smilansky. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.*, 75:435–444, 2003.

[109] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola and J. Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun. Mass Spectrom.*, 22(19):3043–3052, 2008.

[110] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman and D. F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.*, 80(14):5574–5582, 2008.

[111] W. M. Hines, A. M. Falick, A. L. Burlingame and B. W. Gibson. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *J. Am. Soc. Mass Spectrom.*, 3(4):326 – 336, 1992.

[112] C. A. R. Hoare. FIND (algorithm 65). *Communications of the ACM*, 4:321–322, 1961.

[113] D. H. Horn, R. A. Zubarev and F. W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectr.*, 11:320–332, 2000.

[114] C. S. Hsu. Diophantine approach to isotopic abundance calculations. *Anal. Chem.*, 56(8): 1356–1361, 1984.

[115] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman and R. G. Cooks. The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.*, 40(4):430–443, 2005.

[116] R. Hussong and A. Hildebrandt. Signal processing in proteomics. *Methods Mol. Biol.*, 604: 145–161, 2010.

[117] N. Jaitly, M. E. Monroe, V. A. Petyuk, T. R. W. Clauss, J. N. Adkins and R. D. Smith. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.*, 78(21):7397–7409, 2006.

[118] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, 2005.

[119] R. S. Johnson and J. A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Methods Mol. Biol.*, 146:41–61, 2000.

[120] R. S. Johnson and J. A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol. Biotechnol.*, 22(3):301–315, 2002.

[121] P. Jones, R. G. Côté, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob, and R. Apweiler. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, 34(Database-Issue):659–663, 2006.

[122] H. J. Joshi, M. J. Harrison, B. L. Schulz, C. A. Cooper, N. H. Packer and N. G. Karlsson. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics*, 4(6):1650–1664, 2004.

[123] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11): 923–925, 2007.

[124] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34:D354–D357, 2006.

[125] R. Kannan. Lattice translates of a polytope and the Frobenius problem. *Combinatorica*, 12:161–177, 1991.

[126] E. A. Kapp, F. Schütz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn, and R. J. Simpson. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics*, 5:3475–3490, 2005.

[127] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Anal. Chem.*, 60:2299–2301, 1988.

[128] A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, 74(20):5383–5392, 2002.

[129] A. Keller, J. Eng, N. Zhang, X.-J. Li and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, 1:2005.0017, 2005.

[130] E. Kendrick. A mass scale based on CH2 = 14.0000 for high resolution mass spectrometry of organic compounds. *Anal. Chem.*, 35(13):2146–2154, 1963.

[131] A. Kerber, R. Laue and D. Moser. Ein Strukturgenerator für molekulare Graphen. *Anal. Chim. Acta*, 235:221 – 228, 1990.

[132] A. Kerber, R. Laue, M. Meringer and C. Rücker. Molecules in silico: The generation of structural formulae and its applications. *J. Comput. Chem. Japan*, 3(3):85–96, 2004.

[133] S. Kim, N. Gupta and P. A. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, 7(8):3354–3363, 2008.

[134] S. Kim, N. Bandeira and P. A. Pevzner. Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Mol. Cell. Proteomics*, 8(6):1391–1400, 2009.

[135] S. Kim, N. Gupta, N. Bandeira and P. A. Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics*, 8(1):53–69, 2009.

[136] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7(1): 234, 2006.

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

[138] H. Kubinyi. Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem. *Anal. Chim. Acta*, 247:107–119, 1991.

[139] K.-S. Kwok, R. Venkataraghavan and F. W. McLafferty. Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system. *J. Am. Chem. Soc.*, 95 (13):4185–4194, 1973.

[140] V. Lacroix, C. G. Fernandes, and M.-F. Sagot. Motif search in graphs: Application to metabolic networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(4):360–368, 2006.

[141] A. J. Lapadula, P. J. Hatcher, A. J. Hanneman, D. J. Ashline, H. Zhang and V. N. Reinhold. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS^n data. *Anal. Chem.*, 77(19):6271–6279, 2005.

[142] R. L. Last, A. D. Jones and Y. Shachar-Hill. Towards the plant metabolome and beyond. *Nat. Rev. Mol. Cell Biol.*, 8:167–174, 2007.

[143] A. Lavanchy, T. Varkony, D. H. Smith, N. A. B. Gray, W. C. White, R. E. Carhart, B. G. Buchanan, and C. Djerassi. Rule-based mass spectrum prediction and ranking: Applications to structure elucidation of novel marine sterols. *Org. Mass Spectrom.*, 15(7): 355–366, 1980.

[144] J. Lederberg. Topological mapping of organic molecules. *Proc. Natl. Acad. Sci. U. S. A.*, 53 (1):134–139, 1965.

[145] J. Lederberg. How DENDRAL was conceived and born. In *ACM Conference on the History of Medical Informatics, History of Medical Informatics archive*, pages 5–19, 1987. Available from `http://doi.acm.org/10.1145/41526.41528`.

[146] T. A. Lee. *A Beginner's Guide to Mass Spectral Interpretation*. Wiley, 1998.

[147] M. Lefmann, C. Honisch, S. Boecker, N. Storm, F. von Wintzingerode, C. Schloetelburg, A. Moter, D. van den Boom, and U. B. Goebel. A novel mass spectrometry based tool for genotypic identification of mycobacteria. *J. Clin. Microbiol.*, 42(1):339–346, 2004.

[148] G. Li and F. Ruskey. The advantages of forward thinking in generating rooted and free trees. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms (SODA 1999)*, pages 939–940, Philadelphia, PA, USA, 1999. Society for Industrial and Applied Mathematics.

[149] G. Liu, J. Zhang, B. Larsen, C. Stark, A. Breitkreutz, Z.-Y. Lin, B.-J. Breitkreutz, Y. Ding, K. Colwill, A. Pasculescu, T. Pawson, J. L. Wrana, A. I. Nesvizhskii, B. Raught, M. Tyers, and A.-C. Gingras. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat. Biotechnol.*, 28(10):1015–1017, 2010.

[150] K. K. Lohmann and C.-W. von der Lieth. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.*, 32(Web Server issue):W261–W266, 2004.

[151] B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra: Applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, 19(Suppl 2):ii113–ii121, 2003. Proc. of *European Conference on Computational Biology* (ECCB 2003).

[152] A. Luedemann, K. Strassburg, A. Erban and J. Kopka. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737, 2008.

[153] G. S. Lueker. Two NP-complete problems in nonnegative integer programming. Technical Report TR-178, Department of Electrical Engineering, Princeton University, 1975.

[154] Y.-R. Luo. *Handbook of Bond Dissociation Energies in Organic Compounds*. CRC Press, Boca Raton, 2003.

[155] B. Ma and G. Lajoie. Improving the de novo sequencing accuracy by combining two independent scoring functions in peaks software. Poster at the ASMS Conference on Mass Spectrometry and Allied Topics, 2005.

[156] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(20):2337–2342, 2003.

[157] B. Ma, K. Zhang and C. Liang. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *J. Comput. Syst. Sci.*, 70:418–430, 2005.

[158] K. Maass, R. Ranzinger, H. Geyer, C.-W. von der Lieth and R. Geyer. "Glyco-peakfinder" – de novo composition analysis of glycoconjugates. *Proteomics*, 7(24):4435–4444, 2007.

[159] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.*, 25(1):125–131, 2007.

[160] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, 66(24):4390–4399, 1994.

[161] S. Martello and P. Toth. An exact algorithm for large unbounded knapsack problems. *Oper. Res. Lett.*, 9(1):15–20, 1990.

[162] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Chichester, 1990.

[163] R. Matthiesen, J. Bunkenborg, A. Stensballe, O. N. Jensen, K. G. Welinder and G. Bauw. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics*, 4(9):2583–2593, 2004.

[164] R. Matthiesen, M. B. Trelle, P. Hojrup, J. Bunkenborg and O. N. Jensen. VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.*, 4(6):2338–2347, 2005.

[165] L. McHugh and J. W. Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS Comput. Biol.*, 4(2):e12, 2008.

[166] P. E. Miller and M. B. Denton. The quadrupole mass filter: Basic operating concepts. *J. Chem. Educ.*, 63:617–622, 1986.

[167] L. Mo, D. Dutta, Y. Wan and T. Chen. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal. Chem.*, 79(13):4870–4878, 2007.

[168] E. Mostacci, C. Truntzer, H. Cardot and P. Ducoroy. Multivariate denoising methods combining wavelets and principal component analysis for mass spectrometry data. *Proteomics*, 10(14):2564–2572, 2010.

[169] I. K. Mun and F. W. Mclafferty. Computer methods of molecular structure elucidation from unknown mass spectra. In *Supercomputers in Chemistry*, ACS Symposium Series, chapter 9, pages 117–124. American Chemical Society, 1981.

[170] S. Na, J. Jeong, H. Park, K.-J. Lee and E. Paek. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Mol. Cell. Proteomics*, 7(12): 2452–2463, 2008.

[171] S. Neumann and S. Böcker. Computational mass spectrometry for metabolomics – a review. *Anal. Bioanal. Chem.*, 398(7):2779–2788, 2010.

[172] N. Nguyen, H. Huang, S. Oraintara and A. Vo. Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics*, 26(18): i659–i665, 2010.

[173] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[174] J. A. November. *Digitizing life: the introduction of computers to biology and medicine*. PhD thesis, Princeton University, Princeton, USA, 2006.

[175] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. results of an austrian multicenter study. *J. Mass Spectrom.*, 44(4):485–493, 2009.

[176] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. optimization and characterization of the search algorithm. *J. Mass Spectrom.*, 44(4):494–502, 2009.

[177] S. Orchard, L. Montechi-Palazzi, E. W. Deutsch, P.-A. Binz, A. R. Jones, N. Paton, A. Pizarro, D. M. Creasy, J. Wojcik, and H. Hermjakob. Five years of progress in the standardization of proteomics data: 4th annual spring workshop of the HUPO-proteomics standards initiative. *Proteomics*, 7:3436–3440, 2007.

[178] R. Otter. The number of trees. *The Annals of Mathematics*, 49(3):583–599, 1948.

[179] K. G. Owens. Application of correlation analysis techniques to mass spectral data. *Appl. Spectrosc. Rev.*, 27(1):1–49, 1992.

[180] N. H. Packer, C.-W. von der Lieth, K. F. Aoki-Kinoshita, C. B. Lebrilla, J. C. Paulson, R. Raman, P. Rudd, R. Sasisekharan, N. Taniguchi, and W. S. York. Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics*, 8(1):8–20, 2008.

[181] G. Palmisano, D. Antonacci and M. R. Larsen. Glycoproteomic profile in wine: a 'sweet' molecular renaissance. *J. Proteome Res.*, 9(12):6148–6159, 2010.

[182] D. J. Pappin, P. Hojrup and A. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3(6):327–332, 1993.

[183] C. Y. Park, A. A. Klammer, L. Käll, M. J. MacCoss and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.*, 7(7):3022–3027, 2008.

[184] W. E. Parkins. The uranium bomb, the calutron, and the space-charge problem. *Physics Today*, 58(5):45–51, 2005.

[185] V. Pellegrin. Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. *J. Chem. Educ.*, 60(8):626–633, 1983.

[186] D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[187] R. H. Perry, R. G. Cooks and R. J. Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev.*, 27(6):661–699, 2008.

[188] G. Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Mathematica*, 68(1):145–254, 1937.

[189] S. C. Pomerantz, J. A. Kowalak and J. A. McCloskey. Determination of oligonucleotide composition from mass spectromectrically measured molecular weight. *J. Am. Soc. Mass Spectrom.*, 4:204–209, 1993.

[190] R. Raman, S. Raguram, G. Venkataraman, J. C. Paulson and R. Sasisekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat. Methods*, 2(11):817–824, 2005.

[191] J. L. Ramírez-Alfonsín. *The Diophantine Frobenius Problem*. Oxford University Press, 2005.

[192] J. L. Ramírez-Alfonsín. Complexity of the Frobenius problem. *Combinatorica*, 16(1):143–147, 1996.

[193] I. Rauf, F. Rasche and S. Böcker. Computing maximum colorful subtrees in practice. Manuscript. **[TODO: REMOVE OR UPDATE]**, 2011.

[194] A. L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom.*, 17(3):415–419, 2006.

[195] A. L. Rockwood, M. M. Kushnir and G. J. Nelson. Dissociation of individual isotopic peaks: Predicting isotopic distributions of product ions in MS^n. *J. Am. Soc. Mass Spectr.*, 14:311–322, 2003.

[196] A. L. Rockwood, J. R. Van Orman and D. V. Dearden. Isotopic compositions and accurate masses of single isotopic peaks. *J. Am. Soc. Mass Spectr.*, 15:12–21, 2004.

[197] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601, 1984.

[198] S. Rogers, R. A. Scheltema, M. Girolami and R. Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518, 2009.

[199] R. G. Sadygov and J. R. Yates III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.*, 75(15):3792–3798, 2003.

[200] R. G. Sadygov, D. Cociorva and J. R. Yates III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods*, 1(3): 195–202, 2004.

[201] T. Sakurai, T. Matsuo, H. Matsuda and I. Katakuse. PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed. Mass Spectrom.*, 11(8):396–399, 1984.

[202] A. Salomaa. Counting (scattered) subwords. *B. Euro. Assoc. Theo. Comp. Sci.*, 81:165–179, 2003.

[203] F. Sanger, S. Nicklen and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, 1977.

[204] M. M. Savitski, M. L. Nielsen, F. Kjeldsen and R. A. Zubarev. Proteomics-grade de novo sequencing approach. *J. Proteome Res.*, 4:2348–2354, 2005.

[205] K. Scheubert, F. Hufsky, F. Rasche and S. Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. In *Proc. of Research in Computational Molecular Biology (RECOMB 2011)*, volume 6577 of *Lect. Notes Comput. Sc.*, pages 377–391. Springer, 2011.

[206] J. Seidler, N. Zinn, M. E. Boehm and W. D. Lehmann. De novo sequencing of peptides by MS/MS. *Proteomics*, 10(4):634–649, 2010.

[207] J. Senior. Partitions and their representative graphs. *Am. J. Math.*, 73(3):663–689, 1951.

[208] B. Shan, B. Ma, K. Zhang and G. Lajoie. Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J. Bioinformatics and Computational Biology*, 6(1): 77–91, 2008.

[209] Q. Sheng, Y. Mechref, Y. Li, M. V. Novotny and H. Tang. A computational approach to characterizing bond linkages of glycan isomers using matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, 22(22):3561–3569, 2008.

[210] I. V. Shilov, S. L. Seymour, A. A. Patel, A. Loboda, W. H. Tang, S. P. Keating, C. L. Hunter, L. M. Nuwaysir, and D. A. Schaeffer. The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics*, 6(9):1638–1655, 2007.

[211] H. Shin, M. P. Sampat, J. M. Koomen and M. K. Markey. Wavelet-based adaptive denoising and baseline correction for MALDI TOF MS. *OMICS*, 14(3):283–295, 2010.

[212] F. Sikora. An (almost complete) state of the art around the graph motif problem. Technical report, Université Paris-Est, France, 2010. Available from `http://www-igm.univ-mlv.fr/~fsikora/pub/GraphMotif-Resume.pdf`.

[213] R. M. Silverstein, F. X. Webster and D. Kiemle. *Spectrometric Identification of Organic Compounds*. Wiley, 7th edition, 2005.

[214] G. Siuzdak. *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, second edition, 2006.

[215] D. H. Smith, N. A. Gray, J. G. Nourse and C. W. Crandell. The DENDRAL project: recent advances in computer-assisted structure elucidation. *Anal. Chim. Acta*, 133(4):471 – 497, 1981.

[216] R. K. Snider. Efficient calculation of exact mass isotopic distributions. *J. Am. Soc. Mass Spectrom.*, 18(8):1511–1515, 2007.

[217] H. M. Sobell. Actinomycin and DNA transcription. *Proc. Natl. Acad. Sci. U. S. A.*, 82(16): 5328–5331, 1985.

[218] H. Steen and M. Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature Rev.*, 5: 699–711, 2004.

[219] M. T. Sykes and J. R. Williamson. Envelope: interactive software for modeling and fitting complex isotope distributions. *BMC Bioinformatics*, 9:446, 2008.

[220] J. J. Sylvester and W. J. Curran Sharp. Problem 7382. *Educational Times*, 37:26, 1884.

[221] D. L. Tabb, M. J. MacCoss, C. C. Wu, S. D. Anderson and J. R. Yates. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.*, 75(10):2470–2477, 2003.

[222] H. Tang, Y. Mechref and M. V. Novotny. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics*, 21 Suppl 1:i431–i439, 2005. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2005).

[223] S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. Inspect: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77:4626–4639, 2005.

[224] J. A. Taylor and R. S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.*, 73(11):2594–2604, 2001.

[225] J. A. Taylor and R. S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 11:1067–1075, 1997.

[226] J. van Lint and R. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2001.

[227] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler, editors. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, second edition, 2009. Freely available from `http://www.ncbi.nlm.nih.gov/books/NBK1908/`.

[228] R. Venkataraghavan, F. W. McLafferty and G. E. van Lear. Computer-aided interpretation of mass spectra. *Org. Mass Spectrom.*, 2(1):1–15, 1969.

[229] C.-W. von der Lieth, A. Bohne-Lang, K. K. Lohmann and M. Frank. Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief. Bioinform.*, 5(2):164–178, 2004.

[230] S. A. Waksman and H. B. Woodruff. Bacteriostatic and bacteriocidal substances produced by soil actinomycetes. *Proc. Soc. Exper. Biol.*, 45:609–614, 1940.

[231] M. S. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. U. S. A.*, 91(11):4625–4628, 1994.

[232] J. T. Watson and O. D. Sparkman. *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*. Wiley, 2007.

[233] M. E. Wieser. Atomic weights of the elements 2005 (IUPAC technical report). *Pure Appl. Chem.*, 78(11):2051–2066, 2006.

[234] H. Wilf. *generatingfunctionology*. Academic Press, second edition, 1994. Freely available from `http://www.math.upenn.edu/~wilf/DownldGF.html`.

[235] S. Wolf, S. Schmidt, M. Müller-Hannemann and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.

[236] W. E. Wolski, M. Lalowski, P. Jungblut and K. Reinert. Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC Bioinformatics*, 6:203, 2005.

[237] J. W. Wong, G. Cagney and H. M. Cartwright. SpecAlign—processing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.

[238] L.-C. Wu, H.-H. Chen, J.-T. Horng, C. Lin, N. E. Huang, Y.-C. Cheng and K.-F. Cheng. A novel preprocessing method using Hilbert Huang transform for MALDI-TOF and SELDI-TOF mass spectrometry data. *PLoS One*, 5(8):e12493, 2010.

[239] Y. Wu, Y. Mechref, I. Klouckova, M. V. Novotny and H. Tang. A computational approach for the identification of site-specific protein glycosylations through ion-trap mass spectrometry. In *Proc. of RECOMB 2006 satellite workshop on Systems biology and computational proteomics*, volume 4532 of *Lect. Notes Comput. Sc.*, pages 96–107. Springer, 2007.

[240] C. Xu and B. Ma. Complexity and scoring function of MS/MS peptide de novo sequencing. In *Proc. of Computational Systems Bioinformatics Conference (CSB 2006)*, volume 4 of *Series on Advances in Bioinformatics and Computational Biology*, pages 361–369. Imperial College Press, 2006.

[241] J. Yates, P. Griffin, L. Hood and J. Zhou. Computer aided interpretation of low energy MS/MS mass spectra of peptides. In J. Villafranca, editor, *Techniques in Protein Chemistry II*, pages 477–485. Academic Press, San Diego, 1991.

[242] J. A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys.*, 52(2–3):337–349, 1983.

[243] J. Zaia. Mass spectrometry of oligosaccharides. *Mass Spectrom. Rev.*, 23(3):161–227, 2004.

[244] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins and Y. Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genomics*, 10(6):388–401, 2009.

[245] J. Zhang, D. Xu, W. Gao, G. Lin and S. He. Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Commun. Mass Spectrom.*, 23(21):3448–3456, 2009.

[246] N. Zhang, R. Aebersold and B. Schwikowski. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2(10):1406–1412, 2002.

[247] W. Zhang and B. T. Chait. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.*, 72(11):2482–2489, 2000.

[248] R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics.*, 6(3):377–381, 2007.