De novo prediction of the elemental composition of peptides and proteins based on a single mass

De novo prediction of the elemental composition of peptides and proteins based on a single mass

Jürgen Claesen[1], Dirk Valkenborg[1], Tomasz Burzykowski[1]

[1] I-BioStat, Hasselt University, Belgium

Corresponding author: jurgen.claesen@uhasselt.be

Running title: pacMASS

Summary

Identification of peptides and proteins is a common task in mass spectrometry-based proteomics, but often fails to deliver a comprehensive list of identifications. Downstream analysis, quantitative or qualitative, depends on the outcome of this process. Despite continuous improvement of computational methods, a large fraction of the screened peptides and/or proteins remains unidentified.

We introduce here *pacMASS*, a method that *de novo* predicts the elemental composition of peptides and small proteins based on a single accurate mass, i.e., the observed monoisotopic or average mass. This novel approach returns in a fast and memory efficient manner a limited number of elemental compositions per queried peptide or protein.

1 Introduction

Protein and peptide identification is an important task in mass spectrometry-based proteomics. Towards this aim, tandem MS spectra are commonly used, either in combination with a database [1, 2, 3, 4] or as input for *de novo* peptide sequencing [5, 6]. The results of these approaches depend on the mass accuracy [4, 7, 8, 9], and can be inadequate when peptides and proteins contain amino acid substitutions or insertions or deletions, or are post-translationally modified. Additionally, database-driven identification methods depend on the completeness of the selected database.

Information from MS1 spectra can be used to complement the MS2 spectra. For instance, the aggregated isotope distribution, observed in a full scan spectrum, can be used to validate identified proteins or peptides [10, 11]. MS1 spectra can also be used to identify proteins and peptides. Several methods have been proposed to predict the elemental composition based on the aggregated isotope distribution [12] or based on the fine isotope distribution [13, 14, 15]. Note that the latter can only be retrieved by ultra-high resolution mass spectrometers such as FTICR-MS.

Another type of information available in MS1 spectra is the observed monoisotopic or average mass. In fact, in MS-based metabolomics, the elemental composition of a metabolite is often inferred

based on its accurately measured mass [16]. When applied to peptide-centric MS, the approach faces important limitations: the number of possible molecular formulae increases exponentially with increasing m/z-values [17] (see Figure 1) and, due to the discrete nature of biomolecules, it is impossible to assign a unique elemental composition for molecules with a mass above 126 daltons with a mass error of 1 ppm [18]. Hence, although a large number of candidate compostions can be excluded by applying the *Seven Golden Rules* [19] prediction of the elemental composition of peptides and proteins is considered to be impractical and/or useless given the huge number of candidate compositions.

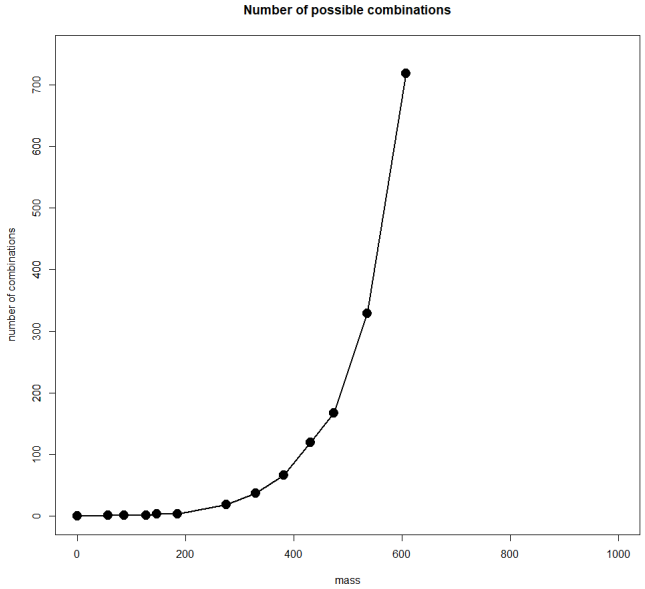

Figure 1: Number of all theoretically possible elemental compositions within a 10ppm wide mass-tolerance-window generated based upon the monoisotopic mass.

In this article, we take a critical look at the issue of prediction of elemental composition of a peptide from its mass. In particular, we show that, by using two simple constraining rules, we can generate in a time- and memory-efficient manner a manageable-size list of candidate compositions for peptides and small proteins with a mass up to 4000 daltons. Our method, **pacMASS** (**p**rediction of the **a**tomic **c**omposition based on a single accurate **mass** of peptides and small proteins), uses the monoisotopic or average mass observed in MS1. It does not require any organism-specific peptide or protein database, nor any information about the measured isotope distribution of a peptide.

Hence, it is very simple to apply.

2 Methodology

In this section, we introduce the workflow implemented in *pacMASS* to *de novo* predict the elemental composition of peptides and proteins with a mass up to 4000 daltons (Figure 2). The models used in the workflow have been trained on peptides and proteins with a monoisotopic mass in the range of 400 to 4000 daltons.

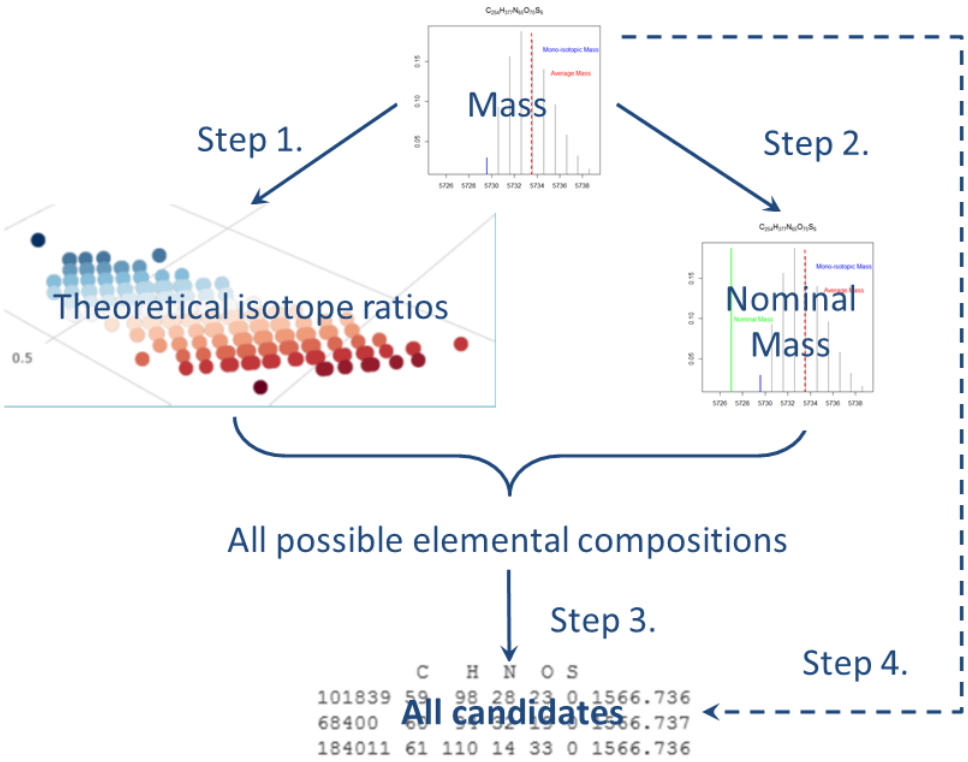

Figure 2: Illustration of the *pacMASS*-algorithm.

pacMASS mostly consists of three steps. In rare situations a fourth step is required. First, the range of C, H, N, and O-atoms for a given molecule is defined with the help of its predicted isotope ratios. The second step refines the proposed ranges for H- and N-atoms with the help of the nominal mass. In the third step, all possible elemental compositions are generated and

subsequently a mass-based filter is applied. Additional filtering (Step 4) is only done when the nominal mass could not be predicted precisely.

The workflow assumes that the number of S-atoms is specified. If a range of possible S-atoms is defined, *pacMASS* is applied for each potential S-atom. In theory, it is possible to determine the correct number of sulphur atoms from the isotope intensities (see Figure 3). We do not consider this approach due to the limited accuracy of the measured isotope distribution intensities.

A similar approach can be used to identify post-translational modifications which consist of other chemical elements than C, H, N, O, or S. For example, in case of phosphorylation, one can define a range of the possible number of P-atoms. Other approaches to account for post-translational-modifications are possible, and are a topic for further research.

2.1 Step 1. Determining the ranges of C-, H-, N-, and O-atoms

The isotope distribution of a molecule is a function of the elemental composition. The aggregated isotope distribution can be represented as a set of isotope ratios, i.e., the ratios between the probability of occurrence of the $(i+1)^{th}$ isotopologue (isotope variant) and the i^{th} isotopologue. As can been seen from Figure 3, the theoretical isotope ratios, R_i, can be used to specify the number of C-, H-, N-, and O-atoms. A given combination of isotope ratios corresponds to a specific number of C-, H-, N-, and O-atoms. For example, the "isotope ratio"-vector (R_1=0.465, R_2=0.322, R_3=0.241) (Figure 3, arrow) indicates a molecule with 36 C-atoms, 52 H-atoms, 22 O-atoms, and 12 N-atoms.

By comparing the theoretical isotope ratios with the isotope ratios derived from the measured isotope-variant intensities, the elemental composition of a given peptide or protein can be determined. As mentioned earlier, due to the limited accuracy of the measured isotope-variant intensities, it is unlikely that the correct elemental composition would be found without acknowledging this uncertainty. Therefore, instead of using the observed isotope ratios, we propose to estimate the isotope ratios based upon the observed monoisotopic or average mass by applying a 4^{th} order

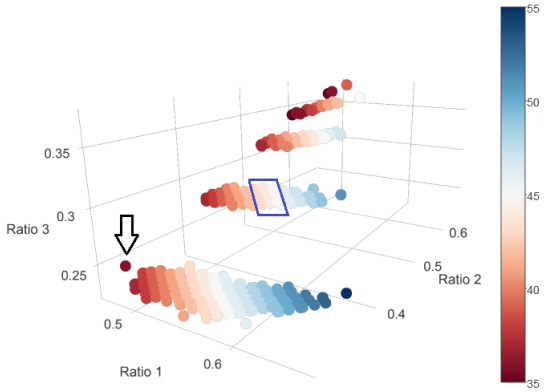

Figure 3: Theoretical isotope ratios of all peptides with a monoisotopic mass between 1000 and 1005Da. Each peptide is colored according to its number of carbons. The five different clouds of points correspond to the number of sulphur-atoms (ranging from 0 (bottom) to 4 (top)).The black arrow points at a molecule with atomic composition $C_{36}H_{52}N_{12}O_{22}$, and isotope ratios R_1=0.465, R_2=0.322, R_3=0.241. The blue square illustrates the proposed range of C-atoms. This range is based on the predicted isotope ratios.

polynomial regression model proposed by [20]:

$$R_i = \beta_{0,i} + \beta_{1,i} \times m/1000 + \beta_{2,i} \times (m/1000)^2 + \beta_{3,i} \times (m/1000)^3 + \beta_{4,i} \times (m/1000)^4, \quad (1)$$

where R_i is the i^{th} isotope ratio and m is the observed monoisotopic or average mass of the peptide.

The coefficients of (1) have been estimated based on the theoretical isotope distributions of an *in-silico* digest of the human proteome (UniProtKB 9606, keyword 181, Release 2011-11) (Table S1 and S2).

Based on the observed monoisotopic or average mass of an unknown peptide, m_x, the four isotope ratios (R_1, R_2, R_3, and R_4) are estimated with (1). Their 95% prediction intervals are calculated as follows:

$$\hat{R}_i \pm 1.96 \times \sqrt{MSE_i \times (1 + 1/n + \frac{(m_x/1000 - \bar{m}/1000)^2}{\sum_j (m_j/1000 - \bar{m}/1000)^2})} \quad (2)$$

where MSE_i is the mean squared error of model (1) for the i^{th} isotope ratio, n is the total number of peptides used to fit model (1), and $\bar{m}$ is the mean mass of the peptides. The prediction intervals are used to define the minimum and maximum number of the atoms of the selected peptide to be checked in the theoretical isotope ratio lookup table created from the human proteome (UniProtKB

9606) (Figure 3, blue square). The determined ranges of C, H, N and O are used in the subsequent steps.

2.2 Step 2. Generating theoretically possible numbers for C-, H-, N-, and O-atoms

Based on the predicted isotope ratios and the database of theoretical isotope ratios, the minimum and maximum number of C-, H-, N-, and O-atoms is determined (Step 1). For carbon and oxygen, each number within the determined range is in theory possible. However, this is not the case for nitrogen- and hydrogen-atoms, as stated by the respective nitrogen-rule [19] and the related hydrogen-rule. These rules are based on the nominal mass of a molecule, i.e., the sum of the integer masses of the most abundant isotopes of each constituent element.

The nitrogen rule is commonly known and poses that the number of nitrogen atoms is even when the nominal mass is even. A similar rule can be formulated for the hydrogen atoms of peptides and proteins, i.e., a peptide or protein with an odd nominal mass has an odd number of H-atoms. Figure 4 illustrates both rules. Based on Figure 4 (right), a refined hydrogen-rule can be stated: peptides with a nominal mass that is divisible by four (e.g., 2040 Da) have an even number of H-atoms that is divisible by four (e.g., 120, 124, ..., 168). We are currently investigating if this refined rule is true for all peptides. Therefore, we have chosen not to incorporate this "refined" hydrogen-rule in *pacMASS*.

In order to be able to apply the hydrogen- and nitrogen-rules as constraints for elemental composition prediction, the nominal mass has to be known. We propose to estimate the nominal mass and its 95% prediction interval by using the following linear model:

$$m_N = \beta_0 + \beta_1 \times m + \varepsilon, \tag{3}$$

with m_N denoting the nominal mass, m denoting the observed monoisotopic or average mass of a peptide, and $\varepsilon \sim N(0, \sigma^2)$. The coefficients of (3) are estimated by using the nominal masses of

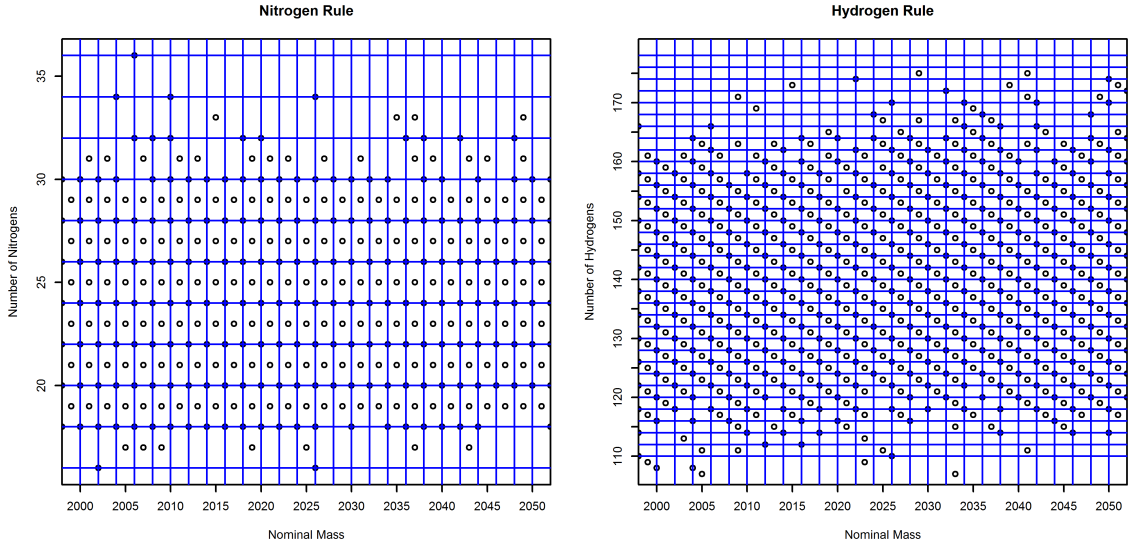

Figure 4: Illustration of the nitrogen- and hydrogen-rule. The vertical lines correspond to even nominal masses, the horizontal lines correspond to an even number of N- or H-atoms.

the human proteome considered in Step 1 (Table S3). The 95% prediction interval of a peptide with mass m_x is calculated as follows:

$$\hat{m}_N \pm 1.96 \times \sqrt{MSE \times (1 + 1/n + \frac{(m_x - \bar{m})^2}{\sum_j (m_j - \bar{m})^2})} \tag{4}$$

where MSE is the mean squared error of the model (3), n is the total number of peptides used to fit model (3), and $\bar{m}$ is the mean of mass of the peptides.

Whenever the rounded upper and lower limit of the predicted nominal mass are identical, the hydrogen- and nitrogen-rules are applied to generate all theoretically possible numbers of H and N. For example, for the peptide "TGGLADK" with monoisotopic mass 660.3443 Da, the lower and upper limit of the 95% prediction interval is, respectively, 659.8780 and 660.1064 Da. Rounding these limits to the nearest integer leads to identical values, and thus the hydrogen- and nitrogen-rules can be applied. When the rounded prediction limits are not identical, every integer between the minimum and maximum, determined in Step 1, is considered to be possible. In this case, an additional filtering step is applied after generating all possible elemental compositions (Step 4).

2.3 Step 3. Generating and filtering all possible elemental compositions

With the lists of potential carbon-, hydrogen-, nitrogen-, and oxygen-atoms, a long list of possible elemental compositions is generated. A small fraction of these elemental compositions have a mass similar to the observed mass. A mass-based-filter is used to retain the candidate elemental compositions within a predefined mass tolerance window. The tolerance is chosen in function of the reported mass accuracy of the mass spectrometer.

2.4 Step 4. Filtering the candidate elemental compositions

For the rare cases when the hydrogen- and nitrogen-rules cannot be applied, all numbers of nitrogens and hydrogens within the specified range are used to generate all possible elemental compositions. As a consequence, the list of possible elemental compositions becomes approximately four times larger than when Step 2 can be used. Applying the mass filter of Step 3 reduces the total number of elemental compositions, but not to the same extent as when the hydrogen- and nitrogen-rules are applied. Therefore, we propose an extra filter based upon the first condition of Senior's theorem [21, 22]. The condition states that the sum of valences or the total number of atoms having odd valences is even. Combining the first condition of Senior's theorem together with the mass filter results in exactly the same list of candidate elemental compositions as when using the nitrogen- and hydrogen rules. Although using Step 4 returns the same outcome as when using Step 2, it should not be the preferred choice when the hydrogen- and nitrogen-rules are applicable, as Step 2 is more efficient with respect to memory usage and computation time.

3 Data

We illustrate the performance of *pacMASS* using three datasets. The first dataset is an *in-silico* tryptic digest of bovine cytochrome C and six proteins and peptides commonly used as internal standards (Table 1). We did not add measurement errors to the theoretical monoisotopic and average masses. This dataset is added as a proof-of-concept to illustrate the potential of *pacMASS*

in an ideal setting.

Table 1: Bovine cytochrome C tryptic digest and internal standards

Peptide	Amino acid sequence	Atomic composition	monoisotopic mass	Average mass
1	IFVQK	$C_{31}H_{51}N_7O_7$	633.38500	633.78083
2	YIPGTK	$C_{32}H_{51}N_7O_9$	677.37483	677.79037
3	MIFAGIK	$C_{37}H_{62}N_8O_8S$	778.44113	779.00482
4	KYIPGTK	$C_{38}H_{63}N_9O_{10}$	805.46979	805.96297
5	EDLIAYLK	$C_{45}H_{73}N_9O_{14}$	963.52770	964.11515
6	TGPNLHGLFGR	$C_{52}H_{81}N_{17}O_{14}$	1167.61489	1168.30777
7	GEREDLIAYLKK	$C_{64}H_{107}N_{17}O_{20}$	1433.78783	1434.63949
8	TGQAPGFSYTDANK	$C_{63}H_{93}N_{17}O_{23}$	1455.66302	1456.51580
9	KTGQAPGFSYTDANK	$C_{69}H_{105}N_{19}O_{24}$	1583.75798	1584.68839
10	IFVQKCAQCHTVEK	$C_{71}H_{116}N_{20}O_{20}S_2$	1632.81162	1633.93851
11	GITWGEETLMEYLENPK	$C_{90}H_{136}N_{20}O_{30}S$	2008.94519	2010.22927
12	GITWGEETLMEYLENPKK	$C_{96}H_{148}N_{22}O_{31}S$	2137.04016	2138.40186
13	RPPGF	$C_{27}H_{40}N_8O_6$	572.30708	572.65787
14	DRVYIHPF	$C_{50}H_{71}N_{13}O_{12}$	1045.53451	1046.18111
15	QLYENKPRRPYIL	$C_{78}H_{124}N_{22}O_{20}$	1688.93622	1689.95850
16	ELYENKPRRPYIL	$C_{78}H_{123}N_{21}O_{21}$	1689.92024	1690.94322
17	RPVKVYPNGAEDESAEAFPLEF	$C_{112}H_{165}N_{27}O_{36}$	2464.19105	2465.67328
18	FVNQHLCGSHLVEALYLVCGERGFFYTPKA	$C_{157}H_{232}N_{40}O_{41}S_2$	3397.67401	3399.90528

The second dataset is a tryptic digest of bovine serum albumin A. One vial of bovine serum albumin (BSA) digest (Bruker part number 8217498) was taken in 500 μl of 2% ACN, with 0.1 FA in water. From this 1pmol/μl solution 1μl was loaded onto a reverse phase C18 column. The HPLC system was directly coupled to an Impact II ESI-Q-TOF system. Peptides eluting from the reverse phase chromatography were measured and fragmented.

We use this dataset to illustrate the potential of using *pacMASS* to improve the level of identification obtained by using database search-engines such as MASCOT. In particular, we use the results of a MASCOT [3] search that was performed on 3122 precursor ions with a 10ppm peptide mass tolerance on the monoisotopic mass and a 0.05Da fragment mass tolerance against the SwissProt database 2017.05, allowing for at most one missed cleavage, and returning at most ten possible identifications per precursor ion. Two modifications have been included in the search, i.e., carbamidomethylation of cysteine and oxidation of methionine. Sixty three peptides have been identified with MASCOT version 1.0 (Table 2). The average masses of these peptides were calculated from their measured aggregated isotope distributions: the masses of the isotopologues were multiplied by their respective peak heights, and this product was divided by the sum of the considered peak heights.

Finally, we analyzed a publicly available HeLa cell tryptic digest-dataset (PXD001592, [23]) mea-

Table 2: Bovine serum albumin A tryptic digest. The theoretical monoisotopic and average masses are calculated without accounting for the reported post-translational modifications.

Peptide	Amino-acid sequence	Atomic Composition	m_0^{theor}	m_{avg}^{theor}	Post-translational modifications	m_0^{obs}	m_{avg}^{obs}
1	AFDEK	$C_{27}H_{40}N_6O_{10}$	608.28059	608.64201		608.28036	608.67836
2	CASIQK	$C_{26}H_{48}N_8O_9S$	648.32650	648.77496	Carbamidomethylation	705.34715	705.82109
3	IETMR	$C_{26}H_{48}N_8O_9S$	648.32650	648.77496		648.32522	648.77178
4	QEPER	$C_{26}H_{43}N_9O_{11}$	657.30820	657.67473		657.30475	657.64297
5	TPVSEK	$C_{28}H_{49}N_7O_{11}$	659.34901	659.73036		659.34877	659.75798
6	KFWGK	$C_{34}H_{48}N_8O_6$	664.36968	664.79655		664.37032	664.77259
7	AWSVAR	$C_{31}H_{48}N_{10}O_8$	688.36566	688.77664		688.36526	688.78194
8	GACLLPK	$C_{31}H_{56}N_8O_8S$	700.39418	700.89276	Carbamidomethylation	757.41590	757.93043
9	SEIAHR	$C_{29}H_{49}N_{11}O_{10}$	711.36639	711.76866		711.36423	711.84129
10	CCAADDK	$C_{26}H_{44}N_8O_{12}S_2$	724.25201	724.80750	Carbamidomethylation (2x)	838.29508	838.83065
11	NYQEAK	$C_{32}H_{49}N_9O_{12}$	751.35007	751.78619		751.35000	751.79511
12	LVTDLTK	$C_{35}H_{64}N_8O_{12}$	788.46437	788.93077		788.46506	788.96524
13	ATEEQLK	$C_{34}H_{59}N_9O_{14}$	817.41815	817.88588		817.41848	817.89818
14	LCVLHEK	$C_{37}H_{64}N_{10}O_{10}S$	840.45276	841.03300	Carbamidomethylation	897.47456	898.07933
15	LSQKFPK	$C_{40}H_{66}N_{10}O_{10}$	846.49634	847.01501		846.49808	846.97161
16	DDSPDLPK	$C_{37}H_{59}N_9O_{16}$	885.40798	885.91690		885.40803	885.92888
17	AEFVEVTK	$C_{42}H_{67}N_9O_{14}$	921.48075	922.03529		921.48240	922.08071
18	YLYEIAR	$C_{44}H_{66}N_{10}O_{12}$	926.48617	927.05676		926.48610	927.06178
19	DLGEEHFK	$C_{43}H_{63}N_{11}O_{15}$	973.45051	974.02716		973.45162	974.02875
20	LVVSTQTALA	$C_{44}H_{79}N_{11}O_{15}$	1001.57571	1002.16495		1001.57656	1002.16940
21	QNCDQFEK	$C_{41}H_{62}N_{12}O_{16}S$	1010.41274	1011.06998	Carbamidomethylation	1067.43527	1068.10797
22	QTALVELLK	$C_{46}H_{83}N_{11}O_{14}$	1013.61210	1014.21878		1013.61274	1014.23427
23	SHCIAEVEK	$C_{42}H_{70}N_{12}O_{15}S$	1014.48043	1015.14484	Carbamidomethylation	1071.50185	1072.20142
24	CCTESLVNR	$C_{39}H_{69}N_{13}O_{15}S_2$	1023.44775	1024.17752	Carbamidomethylation (2x)	1137.49310	1138.22224
25	EACFAVEGPK	$C_{46}H_{71}N_{11}O_{15}S$	1049.48518	1050.18898	Carbamidomethylation	1106.50873	1107.26694
26	CCTKPESER	$C_{40}H_{69}N_{13}O_{16}S_2$	1051.44266	1052.18766	Carbamidomethylation (2x)	1165.48737	1166.33718
27	KQTALVELLK	$C_{52}H_{95}N_{13}O_{15}$	1141.70706	1142.39137		1141.70914	1142.38990
28	LVNELTEFAK	$C_{53}H_{86}N_{12}O_{17}$	1162.62339	1163.32271		1162.62306	1163.32623
29	ECCDKPLLEK	$C_{49}H_{84}N_{12}O_{17}S_2$	1176.55188	1177.39605	Carbamidomethylation (2x)	1290.59769	1291.42123
30	FKDLGEEHFK	$C_{58}H_{84}N_{14}O_{17}$	1248.61389	1249.37399		1248.61458	1249.34814
31	HPEYAVSVLLR	$C_{59}H_{94}N_{16}O_{16}$	1282.70337	1283.47822		1282.70230	1283.49637
32	HLVDEPQNLIK	$C_{58}H_{96}N_{16}O_{18}$	1304.70885	1305.48217		1304.70959	1305.49243
33	TCVADESHAGCEK	$C_{52}H_{84}N_{16}O_{22}S_2$	1348.53875	1349.45226	Carbamidomethylation (2x)	1462.58533	1463.46962
34	SLHTLFGDELCK	$C_{60}H_{95}N_{15}O_{19}S$	1361.66494	1362.55445	Carbamidomethylation	1418.68667	1419.56772
35	ETYGDMADCCEK	$C_{53}H_{81}N_{13}O_{23}S_3$	1363.47304	1364.48443	Carbamidomethylation (2x) and Oxidation	1493.51104	1494.63324
36	ETYGDMADCCEK	$C_{53}H_{81}N_{13}O_{23}S_3$	1363.47304	1364.48443	Carbamidomethylation (2x)	1477.51741	1478.42509
37	YICDNQDTISSK	$C_{57}H_{91}N_{15}O_{23}S$	1385.61329	1386.48810	Carbamidomethylation	1442.63727	1443.50166
38	EYEATLEECCAK	$C_{57}H_{89}N_{13}O_{23}S_2$	1387.56357	1388.52482	Carbamidomethylation (2x)	1501.60806	1502.54542
39	TVMENFVAFVDK	$C_{64}H_{98}N_{14}O_{19}S$	1398.68534	1399.61447	Oxidation	1414.68049	1415.52420
40	TVMENFVAFVDK	$C_{64}H_{98}N_{14}O_{19}S$	1398.68534	1399.61447		1398.68570	1399.59726
41	RHPEYAVSVLLR	$C_{65}H_{106}N_{20}O_{17}$	1438.80448	1439.66430		1438.80590	1439.61026
42	LGEYGFQNALIVR	$C_{68}H_{106}N_{18}O_{19}$	1478.78816	1479.68183		1478.79056	1479.67732
43	DDPHACYSTVFDK	$C_{65}H_{92}N_{16}O_{23}S$	1496.62419	1497.58867	Carbamidomethylation	1553.64934	1554.57771
44	VPQVSTPTLVEVSR	$C_{66}H_{114}N_{18}O_{22}$	1510.83551	1511.72210		1510.83674	1511.72899
45	LKPDPNTLCDEFK	$C_{67}H_{106}N_{16}O_{22}S$	1518.73883	1519.72191	Carbamidomethylation	1575.76256	1576.68842
46	DAFLGSFLYEYSR	$C_{74}H_{102}N_{16}O_{22}$	1566.73546	1567.69921		1566.73710	1567.65519
47	ECCHGDLLECADDR	$C_{60}H_{95}N_{19}O_{25}S_3$	1577.59086	1578.71002	Carbamidomethylation (3x)	1748.65780	1749.70773
48	QEPERNECFLSHK	$C_{68}H_{105}N_{21}O_{23}S$	1615.74129	1616.75782	Carbamidomethylation	1672.76391	1673.61538
49	YNGVFQECCQAEDK	$C_{68}H_{100}N_{18}O_{25}S_2$	1632.65484	1633.76278	Carbamidomethylation (2x)	1746.70183	1747.71956
50	KVPQVSTPTLVEVSR	$C_{72}H_{126}N_{20}O_{23}$	1638.93047	1639.89469		1638.93429	1639.84518
51	PCFSALTPDETYVPK	$C_{76}H_{114}N_{16}O_{24}S$	1666.79126	1667.88087	Carbamidomethylation	1723.81572	1724.90578
52	MPCTEDYLSLILNR	$C_{72}H_{118}N_{18}O_{23}S_2$	1666.80586	1667.94985	Carbamidomethylation and Oxidation	1739.82500	1740.93782
53	MPCTEDYLSLILNR	$C_{72}H_{118}N_{18}O_{23}S_2$	1666.80586	1667.94985	Carbamidomethylation	1723.83023	1724.87285
54	CCAADDKEACFAVEGPK	$C_{72}H_{113}N_{19}O_{26}S_3$	1755.72663	1756.98119	Carbamidomethylation (3x)	1926.79487	1927.94128
55	RPCFSALTPDETYVPK	$C_{82}H_{126}N_{20}O_{25}S$	1822.89237	1824.06695	Carbamidomethylation	1879.91688	1881.00649
56	NECFLSHKDDSPDLPK	$C_{79}H_{121}N_{21}O_{28}S$	1843.84106	1844.99999	Carbamidomethylation	1900.86783	1901.94629
57	LFTFHADICTLPDTEK	$C_{84}H_{127}N_{19}O_{26}S$	1849.89204	1851.08902	Carbamidomethylation	1906.91929	1908.01754
58	HPYFYAPELLYYANK	$C_{94}H_{125}N_{19}O_{23}$	1887.91957	1889.11620		1887.92499	1889.03375
59	RHPYFYAPELLYYANK	$C_{100}H_{137}N_{23}O_{24}$	2044.02068	2045.30228		2044.02160	2045.17196
60	ECCHGDLLECADDRADLAK	$C_{82}H_{133}N_{25}O_{32}S_3$	2075.87106	2077.28425	Carbamidomethylation (3x)	2246.94033	2248.17853
61	YNGVFQECCQAEDKGACLLPK	$C_{99}H_{154}N_{26}O_{32}S_3$	2315.03846	2316.64026	Carbamidomethylation (3x)	2486.11151	2487.33361
62	DAIPENLPPLTADFAEDKDVCK	$C_{105}H_{165}N_{25}O_{37}S$	2400.15189	2401.65014	Carbamidomethylation	2457.18080	2458.43924
63	GLVLIAFSQYLQQCPFDEHVK	$C_{113}H_{171}N_{27}O_{31}S$	2434.23550	2435.80073	Carbamidomethylation	2491.26461	2492.59891

sured with an Impact II ESI-Q-TOF. We use the dataset to investigate the question whether, based on a spectrum for a complex peptide mixture, the lists of candidate compositions provided by *pacMASS* do include the compositions of putative amino acid sequences.

4 Results

4.1 Bovine cytochrome C

pacMASS was applied to the 18 peptides and internal standards. We used the theoretical monoisotopic mass and average mass as input. We defined, for each peptide, a range of S-atoms, i.e., 0, 1, 2, and 3. We also set the mass tolerance of *pacMASS* for the monoisotopic and the average mass equal to 5ppm. A relatively small number of elemental compositions was returned by *pacMASS* with the monoisotopic and average mass as input (Table 3). For example, for peptide #18, with monoisotopic mass 3397.67Da, 354 and 361 potential atomic compositions were found for the monoisotopic and average mass, respectively. For each peptide, the list of candidate elemental compositions always included the correct one.

Table 3: *pacMASS* results for bovine cytochrome C tryptic digest and internal standards, with 5ppm as tolerance. The numbers in brackets are the numbers of combinations when the correct number of S-atoms is specified.

Peptide		pacMASS m_0 Candidates	pacMASS m_{avg} Candidates
1	$C_{31}H_{51}N_7O_7$	5 (3)	16 (10)
2	$C_{32}H_{51}N_7O_9$	7 (3)	17 (14)
3	$C_{37}H_{62}N_8O_8S$	13 (4)	17 (12)
4	$C_{38}H_{63}N_9O_{10}$	17 (7)	22 (14)
5	$C_{45}H_{73}N_9O_{14}$	23 (9)	33 (14)
6	$C_{52}H_{81}N_{17}O_{14}$	41 (12)	40 (20)
7	$C_{64}H_{107}N_{17}O_{20}$	71 (25)	69 (29)
8	$C_{63}H_{93}N_{17}O_{23}$	79 (23)	53 (27)
9	$C_{69}H_{105}N_{19}O_{24}$	82 (23)	74 (27)
10	$C_{71}H_{116}N_{20}O_{20}S_2$	98 (19)	94 (25)
11	$C_{90}H_{136}N_{20}O_{30}S$	132 (37)	42 (42)
12	$C_{96}H_{148}N_{22}O_{31}S$	152 (40)	156 (42)
13	$C_{27}H_{40}N_8O_6$	4 (2)	8 (6)
14	$C_{50}H_{71}N_{13}O_{12}$	33 (13)	32 (20)
15	$C_{78}H_{124}N_{22}O_{20}$	88 (28)	100 (34)
16	$C_{78}H_{123}N_{21}O_{21}$	86 (26)	97 (35)
17	$C_{112}H_{165}N_{27}O_{36}$	195 (54)	196 (53)
18	$C_{157}H_{232}N_{40}O_{41}S_2$	354 (78)	361 (73)

4.2 Bovine serum albumin A

We selected all 2988 precursor ions with a monoisotopic mass between 400 and 4000 daltons. We used *pacMASS* to predict, based on the observed monoisotopic mass with a tolerance of 5ppm, elemental compositions for the precursor ions. Separate predictions were obtained by assuming the presence of 0, 1, 2, or 3 S-atoms. For 2906 ions, the use of *pacMASS* resulted in multiple candidate compositions. As expected, the number of candidate formulae increased when the monoisotopic mass increased (Figure 5).

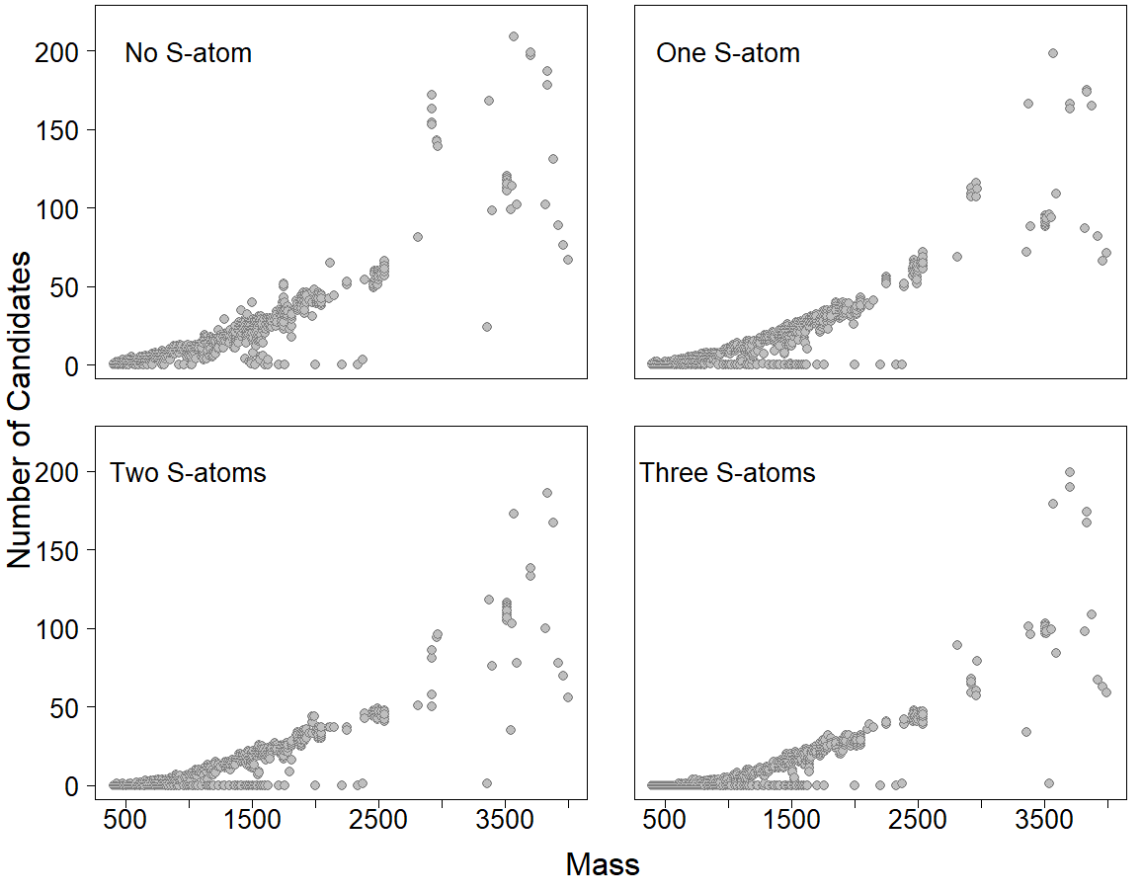

Figure 5: Number of possible elemental compositions for the bovine serum albumin A precursor ions within a 10ppm wide mass-tolerance-window.

For 1735 ions out of the 2988 selected precursor ions, the MASCOT search returned 6979 candidate amino acid sequences, with ion scores ranging from 0.0 to 145.5. We compared the amino acid sequences suggested by MASCOT with the atomic formulae predicted with *pacMASS*. For

1136 precursor ions, we found an overlap between the MASCOT-generated candidate amino acid sequences and the predicted elemental compositions of *pacMASS*. In particular, for those precursor ions, *pacMASS* suggested 3263 putative amino acid sequences. By intersecting *pacMASS* suggestions with MASCOT identifications, the total number of candidate sequences could be considerably limited. In particular, the fraction of unique identifications increased from 32.2% to 47.4%, i.e., by 15.2% (Table 4). For example, for an ion with a monoisotopic mass of 664.371648 daltons, ten candidate MASCOT identifications were reduced, based on *pacMASS* results, to just one amino acid sequence, i.e., KFWGK with $C_{34}H_{48}N_8O_6$ as the atomic composition.

We expect that the number of matches between MASCOT and *pacMASS* would further increase if we increased the mass tolerance, or if we could check for carbamidomethylation as modification. However, the latter information could not be extracted from the available MASCOT .dat file.

Table 4: Candidate amino acid sequences for bovine serum albumin. The numbers at the bottom indicate the total number of precursor ions with one or more Aa sequences, and the total number of candidate Aa sequences (in brackets).

	MASCOT	MASCOT and *pacMASS*
# potential Aa sequences	# precursor ions	# precursor ions
1	559 (32.2%)	538 (47.4%)
2	318 (18.3%)	179 (15.8%)
3	182 (10.5%)	106 (9.3%)
4	104 (6.0%)	78 (6.9%)
5	95 (5.5%)	57 (5.0%)
6	43 (2.5%)	48 (4.2%)
7	47 (2.7%)	25 (2.2%)
8	38 (2.2%)	20 (1.8%)
9	34 (2.0%)	21 (1.8%)
10	315 (18.2%)	64 (5.6%)
total	1735 (6979)	1136 (3263)

The computation time of *pacMASS* with the monoisotopic mass as input, for 2988 peptides within the range of 400 to 4000 Da while repeating Step 1 to Step 4 for S=0, S=1, and S=2, was around 9.5 minutes. The computations were executed in R (version 3.4.4) on a laptop with Windows10, an i7-7700HQ processor, and 16GB of RAM.

Sixty three peptides were identified with high confidence by MASCOT. For these peptides, we extracted the isotope distributions from the corresponding MS1 spectra. These distributions were used to determine the average mass. A mass tolerance of 50ppm was specified. The reason for this

raised tolerance is the increased inaccuracy of the observed isotope intensities and, consequently, an increased uncertainty about the average mass.

Similarly to bovine cytochrome C, a relatively small number of potential elemental compositions was found for each peptide when using *pacMASS* with the monoisotopic mass as input (Table 5). The list of candidate elemental compositions contained the correct atomic formula for every peptide except of peptide #4. For this peptide, the mass tolerance was not sufficient, as the difference between the observed and the theoretical monoisotopic mass is 5.25ppm. Increasing the mass tolerance to 7.5ppm for this peptide resulted in 15 candidates, including the correct one.

When using the average mass as input, the number of potential atomic compositions increased due to the inflated tolerance. However, as illustrated by Figure S1, the tolerance was not high enough to identify all peptides, with 40 out of 63 being correctly identified. Raising the tolerance increased the number of identifications, but at the cost of an increased number of candidate elemental compositions (results not shown).

4.3 HeLa cell tryptic digest

The HeLa cell tryptic digest dataset contains 48,355 uniquely identified peptide sequences. The identification search included carbamidomethylation of cysteine as fixed modifications, and N-terminal protein acetylation and methionine oxidation as variable modifications. Prior to identification, the m/z-values were recalibrated by MaxQuant [24].

We selected all 48,033 identified peptides with a monoisotopic mass between 400 and 4000 daltons. The recalibrated monoisotopic m/z-values were used as input for *pacMASS*. The allowed mass tolerance was set to 5ppm, and a range of S-atoms from 0 to 5 was chosen.

For 47,630 peptides (99.2%) we found a match between the identified amino acid sequence and the candidate elemental compositions of *pacMASS*. Five out of the 403 peptides, for which no matching atomic composition was found, had more than five S-atoms. Increasing the mass tolerance had no effect on the number of matches. Changing the 95% prediction interval of the polynomial model (1)

Table 5: *pacMASS* results for bovine serum albumin A. The checkmark (✓) indicates that the list of candidates contains the correct atomic composition.

Peptide		*pacMASS* m_0 Candidates	pacMASS m_{avg} Candidates
1	$C_{27}H_{40}N_6O_{10}$	7 ✓	65
2	$C_{26}H_{48}N_8O_9S + C_2H_3NO$	9 ✓	120 ✓
3	$C_{26}H_{48}N_8O_9S$	8 ✓	101 ✓
4	$C_{26}H_{43}N_9O_{11}$	9	41 ✓
5	$C_{28}H_{49}N_7O_{11}$	6 ✓	92 ✓
6	$C_{34}H_{48}N_8O_6$	8 ✓	93 ✓
7	$C_{31}H_{48}N_{10}O_8$	7 ✓	99 ✓
8	$C_{31}H_{56}N_8O_8S + C_2H_3NO$	14 ✓	153 ✓
9	$C_{29}H_{49}N_{11}O_{10}$	13 ✓	126
10	$C_{26}H_{44}N_8O_{12}S_2 + 2 \times C_2H_3NO$	12 ✓	103
11	$C_{32}H_{49}N_9O_{12}$	12 ✓	115 ✓
12	$C_{35}H_{64}N_8O_{12}$	12 ✓	164 ✓
13	$C_{34}H_{59}N_9O_{14}$	19 ✓	166 ✓
14	$C_{37}H_{64}N_{10}O_{10}S + C_2H_3NO$	23 ✓	250 ✓
15	$C_{40}H_{66}N_{10}O_{10}$	17 ✓	198
16	$C_{37}H_{59}N_9O_{16}$	23 ✓	178 ✓
17	$C_{42}H_{67}N_9O_{14}$	23 ✓	260 ✓
18	$C_{44}H_{66}N_{10}O_{12}$	27 ✓	256 ✓
19	$C_{43}H_{63}N_{11}O_{15}$	28 ✓	236 ✓
20	$C_{44}H_{79}N_{11}O_{15}$	24 ✓	322 ✓
21	$C_{41}H_{62}N_{12}O_{16}S + C_2H_3NO$	39 ✓	283 ✓
22	$C_{46}H_{83}N_{11}O_{14}$	25 ✓	323 ✓
23	$C_{42}H_{70}N_{12}O_{15}S + C_2H_3NO$	41 ✓	374 ✓
24	$C_{39}H_{69}N_{13}O_{15}S_2 + 2 \times C_2H_3NO$	45 ✓	325 ✓
25	$C_{46}H_{71}N_{11}O_{15}S + C_2H_3NO$	40 ✓	426 ✓
26	$C_{40}H_{69}N_{13}O_{16}S_2 + 2 \times C_2H_3NO$	46 ✓	438 ✓
27	$C_{52}H_{95}N_{13}O_{15}$	27 ✓	401 ✓
28	$C_{53}H_{86}N_{12}O_{17}$	48 ✓	475 ✓
29	$C_{49}H_{84}N_{12}O_{17}S_2 + 2 \times C_2H_3NO$	58 ✓	512
30	$C_{58}H_{84}N_{14}O_{17}$	52 ✓	444 ✓
31	$C_{59}H_{94}N_{16}O_{16}$	51 ✓	560 ✓
32	$C_{58}H_{96}N_{16}O_{18}$	54 ✓	571 ✓
33	$C_{52}H_{84}N_{16}O_{22}S_2 + 2 \times C_2H_3NO$	74 ✓	416
34	$C_{60}H_{95}N_{15}O_{19}S + C_2H_3NO$	73 ✓	715 ✓
35	$C_{53}H_{81}N_{13}O_{23}S_3 + 3 \times C_2H_3NO + O$	42 ✓	788 ✓
36	$C_{53}H_{81}N_{13}O_{23}S_3 + 2 \times C_2H_3NO$	45 ✓	226
37	$C_{57}H_{91}N_{15}O_{23}S + C_2H_3NO$	78 ✓	586 ✓
38	$C_{57}H_{89}N_{13}O_{23}S_2 + 2 \times C_2H_3NO$	84 ✓	584
39	$C_{64}H_{98}N_{14}O_{19}S + O$	72 ✓	685
40	$C_{64}H_{98}N_{14}O_{19}S$	73 ✓	746 ✓
41	$C_{65}H_{106}N_{20}O_{17}$	75 ✓	757 ✓
42	$C_{68}H_{106}N_{18}O_{19}$	69 ✓	740 ✓
43	$C_{65}H_{92}N_{16}O_{23}S + C_2H_3NO$	88 ✓	555 ✓
44	$C_{66}H_{114}N_{18}O_{22}$	71 ✓	822 ✓
45	$C_{67}H_{106}N_{16}O_{22} + C_2H_3NOS$	82 ✓	707
46	$C_{74}H_{102}N_{16}O_{22}$	88 ✓	717 ✓
47	$C_{60}H_{95}N_{19}O_{25}S_3 + 3 \times C_2H_3NO$	92 ✓	437
48	$C_{68}H_{105}N_{21}O_{23}S + C_2H_3NO$	101 ✓	332
49	$C_{68}H_{100}N_{18}O_{25}S_2 + 2 \times C_2H_3NO$	112 ✓	505
50	$C_{72}H_{126}N_{20}O_{23}$	77 ✓	968 ✓
51	$C_{76}H_{114}N_{16}O_{24}S + C_2H_3NO$	100 ✓	982 ✓
52	$C_{72}H_{118}N_{18}O_{23}S_2 + C_2H_3NO + O$	108 ✓	1076 ✓
53	$C_{72}H_{118}N_{18}O_{23}S_2 + C_2H_3NO$	99 ✓	942
54	$C_{72}H_{113}N_{19}O_{26}S_3 + 3 \times C_2H_3NO$	135 ✓	776
55	$C_{82}H_{126}N_{20}O_{25}S + C_2H_3NO$	118 ✓	1084
56	$C_{79}H_{121}N_{21}O_{28}S + C_2H_3NO$	129 ✓	879
57	$C_{84}H_{127}N_{19}O_{26}S + C_2H_3NO$	131 ✓	1117
58	$C_{94}H_{125}N_{19}O_{23}$	125 ✓	1135 ✓
59	$C_{100}H_{137}N_{23}O_{24}$	130 ✓	1237
60	$C_{82}H_{133}N_{25}O_{32}S_3 + 3 \times C_2H_3NO$	177 ✓	714
61	$C_{99}H_{154}N_{26}O_{32}S_3 + 3 \times C_2H_3NO$	196 ✓	433
62	$C_{105}H_{165}N_{25}O_{37}S + C_2H_3NO$	199 ✓	1113
63	$C_{113}H_{171}N_{27}O_{31}S + C_2H_3NO$	199 ✓	1679

to 99% decreased the number of mismatches from 403 to 190 at the cost of obtaining more lengthy lists of candidate compositions.

5 Conclusion

pacMASS is a memory-efficient and computationally fast *de novo* predictor of the elemental composition of peptides and small proteins based upon the observed mass. It uses constraints based on theoretical isotope ratios and the nitrogen- and hydrogen-rules to limit the number of possible atomic formulae. When using high-accuracy masses, a limited list of candidate molecular formulas includes the correct one.

While *pacMASS* cannot (yet) serve to obtain unique identifications in every case, it can be applied in validation of protein- and peptide-identification methods, or to reduce the search space of database-driven identification tools. In this case, the predicted atomic compositions can be used to reduce the number of the amino acid sequences in the protein- or peptide-database suggested by the identification tool. This reduction may lead to a unique identification.

References

[1] Eng, J.K., McCormack, A.L., Yates, J.R. (1994). An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J Am Soc Mass Spectrom.*, 5, 976–989.

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

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

[4] Nesvizhskii, A.I. (2007) Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching. *Methods Mol Biol*, 367, 87-119.

[5] Taylor, J.A., Johnson, R.S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.*, 11(9), 1067-1075.

[6] Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A. (1999) De Novo Peptide Sequencing via Tandem Mass Spectrometry. *J. Comput. Biol*, 6, 327–342.

[7] Mann, M. (1995) Useful Tables of Possible and Probable Peptide Masses. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, GA.*

[8] Nesvizhskii, A.I. (2010) A survey of computational methods and error rate estimation procedures for peptides and protein identification in shotgun proteomics. *J. Proteomics*, 73, 2092-2123.

[9] Spengler, B. (2004) De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 15(5), 703-714.

[10] Zamanzad-Ghavidel, F., Mertens, I., Baggerman, G., Laukens, K., Burzykowski, T., and Valkenborg, D. (2014) The use of the isotopic distribution as a complementary quality metric to assess tandem mass spectra results. *J. Proteom.*, 98, 150-158.

[11] Zamanzad-Ghavidel, F, Claesen, J., Burzykowski, T., and Valkenborg, D. (2014) Comparison of the Mahalanobis distance and Pearson's χ^2-statistic as measures of similarity of isotope patterns. *J. Am. Soc. Mass Spectrom.* 25, 293-296.

[12] Roussis S.G., and Proulx, R. (2003) Reduction of Chemical Formulas from the Isotopic Peak Distributions of High-Resolution Mass Spectra. *Anal. Chem.*, 75, 1470-1482

[13] Stoll N., Schmidt E., and Thurow, K. (2006) Isotope Pattern Evaluation for the Reduction of Elemental Compositions Assigned to High-Resolution Mass Spectral Data from Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *J. Am. Soc. Mass Spectrom.*, 17, 1692-1699.

[14] Miura, D., Tsuji, Y., Takahashi, K., Wariishi, H., and Saito, K. (2010) A strategy for the determination of the elemental composition by fourier transform ion cyclotron resonance mass spectrometry based on isotopic peak ratios., *Anal. Chem.*, 82, 5887-5891.

[15] Miladinovic, S.M., Kozhinov, A.N., Gorshkov, M.V., and Tsybin, Y.O. (2012) On the Utility of Isotopic Fine Structure Mass Spectrometry in Protein Identification. *Anal. Chem.*, 84, 4042-4051.

[16] Scheubert, K., Hufsky, F., and B$\acute{}$ocker (2013) Computational mass spectrometry for small molecules. *J. Chemoinformatics*, 5(1), 12

[17] Grange, A.H., Genicola, F.A., and Sovocool, G.W. (2002) Utility of three types of mass spectrometers for determining elemental compositions of ions formed from chromatographically separated compounds. *Rapid Commun. Mass Spectrom.*, 16, 2356-2369.

[18] Kind, T. and Fiehn, O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 28, 234-243

[19] Kind, T. and Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8, 105-124.

[20] Valkenborg, D., Jansen, I., and Burzykowski, T. (2008) A model-based method for the prediction of the isotopic distribution of peptides. *J. Am. Soc. Mass Spectrom.*, 19, 703-712.

[21] Senior, J.K. (1951) Unimerism, *J. Chem. Phys.*, 19, 865-873.

[22] Senior, J.K. (1951) Partitions and their representative graphs,. *Amer. J. Math.*, 73, 663-689.

[23] Beck, S., Michalski, A., Raether, O., Lubeck, M., Kaspar, S., Goedecke, N., Baessmann, C., Hornburg, D., Meier, F., Paron, I., Kulak, N.A., Cox, J., and Mann, T. (2015) The Impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics *Mol Cell Proteomics*, 14, 2014-2029.

[24] Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26, 1367-1372.

Acknowledgements

The authors would like to thank Romano Hebeler and Frank Aerts from Bruker for providing us with the bovine serum albumin A data.