Invited Keynote Talk: Computing P-Values for Peptide Identifications in Mass Spectrometry

Nikita Arnold[1,2], Tema Fridman[1], Robert M. Day[1], and Andrey A. Gorin[1]

[1] Computational Biology Institute, Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830
[2] Soft Matter Physics/Experimental Physics, J. Kepler University, Altenbergerstraße 69, A-4040, Linz, Austria

Abstract. Mass-spectrometry (MS) is a powerful experimental technology for "sequencing" proteins in complex biological mixtures. Computational methods are essential for the interpretation of MS data, and a number of theoretical questions remain unresolved due to intrinsic complexity of the related algorithms. Here we design an analytical approach to estimate the confidence values of peptide identification in so-called database search methods. The approach explores properties of mass tags — sequences of mass values (m_1 m_2 ... m_n), where individual mass values are distances between spectral lines. We define p-function — the probability of finding a random match between any given tag and a protein database — and verify the concept with extensive tag search experiments. We then discuss p-function properties, its applications for finding highly reliable matches in MS experiments, and a possibility to analytically evaluate properties of SEQUEST X-correlation function.

Keywords: mass-spectrometry, database search, confidence values.

1 Introduction

Mass-spectrometry based proteomics is the driving engine behind an increasingly rich variety of biological experiments: from a pull-down "hunt" of the protein complexes to whole cell protein expression profiles. The resulting information flow, while disparate in nature and usually huge in volume, often has a common structure of the underlying raw data — individual spectra of short peptides converted into sequences assigned to them by various algorithms.

In a typical experiment, cellular proteins are cut into relatively short peptides (10-20 amino acids), and each analyzed peptide results in an MS spectrum as presented in Fig 1. Peaks are footprints of smaller chemical fragments, where peak position reflects each fragment's mass-to-charge ratio that can be converted to a mass value. Ions corresponding to the breaking of a peptide bond (two highlighted peaks on the picture) are called b-ions and their complements to the full peptide are called y-ions. Typically these two types of ions have relatively high intensity as peptides break more

I. Măndoiu, R. Sunderraman, and A. Zelikovsky (Eds.): ISBRA 2008, LNBI 4983, pp. 100–109, 2008.

easily across peptide bonds. All identification methods utilize this property in some way, but other types of ions also saturate a spectrum (outnumbering "noble" b- and y-ions by a ratio of 20:1), and some of them can be very strong as well.

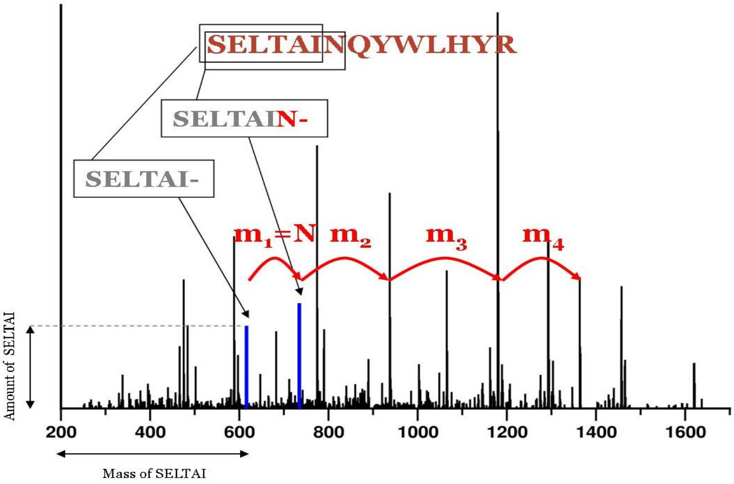

Fig. 1. MS experimental data. Highlighted peaks are formed by 2 b-ions; the distance between them is the mass of N residue at the right terminus of the partial peptide SELTAIN.

Peptide identification aims to infer peptide amino acid sequence from its spectrum. Database search methods [1-5] dominate the field, with an overwhelming majority of experiments using one of them. In database search methods, a peptide is assumed to belong to a known protein database (DB). The SEQUEST program (developed in John Yates group, [2]) uses the following algorithms (some details are simplified):

(1) The experimental spectrum is re-calibrated, so several of the strongest peaks are given an intensity of 1, and other peaks are rescaled accordingly.

(2) The program forms a large list of candidate peptides selected from the target protein DB.

(3) Theoretical spectra are generated for each of the candidate peptides. They usually include only b- and y-ions, and only ion positions are important, as currently there is no reliable way to model relative peak intensity.

(4) Theoretical constructs are matched against the experimental spectrum to compute a matching score. The X-corr reflects the total intensity of experimental peaks that were matched within experimental precision of theoretical positions. The candidate peptide with the highest X-corr value is selected as the output solution.

A typical proteomics experiment incorporates millions of individual peptide identifications, and the reliability of individual assignments is crucially important. We have described the SEQUEST algorithm, because (1) it is hugely popular (probably ~ 50% of the market); and (2) many other search methods were inspired by SEQUEST and work in a similar fashion. The description is also instructive in regards to algorithm complexity and challenges that one needs to overcome to estimate the reliability of the answers. Every spectrum will be assigned some candidate peptide, but what cutoff of the X-corr values would guarantee, for example, that 95% of the assignments are correct?

The standard way of addressing this problem is by introducing an artificial negative control into the experiment [6,7,8]. The identification procedure is run against a database with two parts: a "true" DB of all protein sequences, which actually were present in the source sample, and "false" one, containing negative controls (also called decoy DB). The decoy database contains proteins that cannot be possibly matched by the sample in question. Several research groups extensively investigated the best approaches to create negative control DB and use them for learning reliable values of the X-corr [9,10,11,12,13,14]. The matches to the decoy part are incorrect by design of the experiment, and X-corr cutoffs can be mapped to the sensitivity values by assuming, for example, that the total number of false matches was twice as high as observed in the false DB.

However, this approach has numerous drawbacks. The X-corr values recorded for a given spectrum depend on many factors: size of the database (in a non-trivial way that is hard to figure out), the particular type of MS device, the type of precursor ion, contaminations (MS experiments are ultra sensitive), and even on the organism, that was the source of the tested sample. On an intuitive level it is clear that the X-corr cutoff should be determined by parameters of a particular spectrum. But this road has insurmountable difficulties for an empirical approach, as there is no obvious way to divide spectra into classes.

Strictly speaking the mapping of the X-corr to the probability has to be done for each modification of the experimental system, but it is not an easy demand. The practical approach is to take a "high enough" X-corr cutoff and hope that the fraction of correct matches will not fall too low. Usually only a small fraction of spectra passes the required cutoff (10-20%), and dissatisfaction goes both ways: it is often a rather small "crop", and it is still not obvious how reliable the obtained matches are.

We propose to explore a different approach to the problem by examining database matches of somewhat simpler objects, which we call mass tags. The idea of "tags" was pioneered by Mann's group [15] and further developed in [16]. We derive an analytical expression for the probability of tag random match and explore the properties of the corresponding function. We also propose a database search model that gives an analytical estimate for the fraction of correct matches and outline how SEQUEST X-corr function can be evaluated in the same mathematical framework.

2 Probability Function of Peptide Mass

We define mass tag as a sequence of mass values (m_1 m_2 ... m_l), where individual values (called connector masses or simply connectors) are distances between spectral lines in a specific peak subset. One such subset and the corresponding mass tag of length 4 are illustrated for the spectrum in Fig. 1. We define a match between tag and database in a different way than match between spectrum and database is usually defined. A DB entry *seq* matches a tag (m_1 m_2 … m_n) if it contains *n* consecutive protein sequences (seq_1 seq_2 … seq_n), where each sequence seq_k has a mass m_k within experimental precision of the MS device (for our purposes we assume it to be 0.5 Da).

2.1 *P*-Function for a Single Connector

To explore properties of tag matches we will introduce another definition, which is central for all subsequent developments. We define peptide probability mass function — *p*-function — as probability to observe a peptide in the window (*m-dm, m+dm*) starting at an arbitrary point of the protein database. This probability is a function of both *m* (mass of the desired peptide) and *dm* (detection precision), but it does not depend on the size of the database. For example, if *m* is the mass of amino acid Ala and *dm*=0.5 Da, then the probability is equal to Ala frequency in the database.

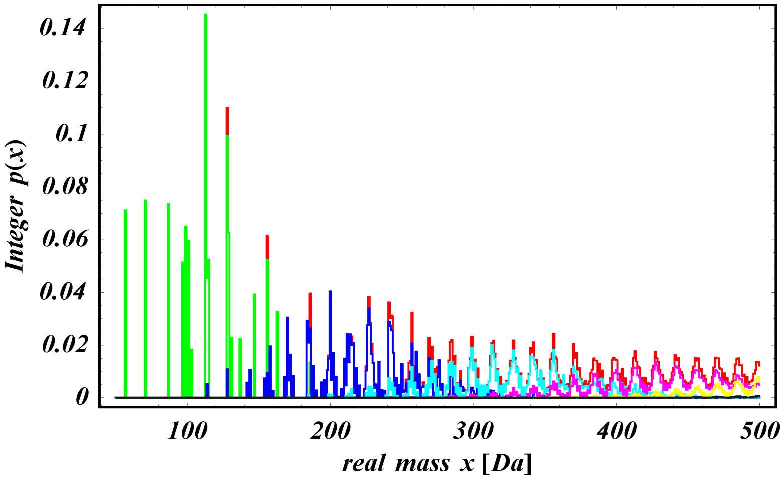

Fig. 2. The contribution of different steps into p-function distribution: green curve - single, blue –double, cyan - triple, magenta - 4 residue steps, and yellow - 5. The overall value of the theoretical *p*-function is shown in red.

The *p*-function dependence from *dm* is an interesting topic that deserves a separate discussion. For the rest of the paper we assume *dm*=0.5 Da. This value is a good choice for two reasons: (1) 0.5 Da is the precision of the most common mass-spectrometers used in proteomics research; (2) peptide masses are naturally concentrated to the centers of so-called Mann bins [17,18,19], which are separated by ~1Da distance on the mass axis. In this sense masses of all peptides, derived ions, and distances between ions are nearly integer (Mann's bin mass is ~1.0005 Da). Tags have connectors that could be expressed as an integer number of Mann bins, and *p*-function can be computed for *m* values centered on such bins with the *dm*=0.5. However it is worth to note that our methodology will straightforwardly accommodate any value of *dm*.

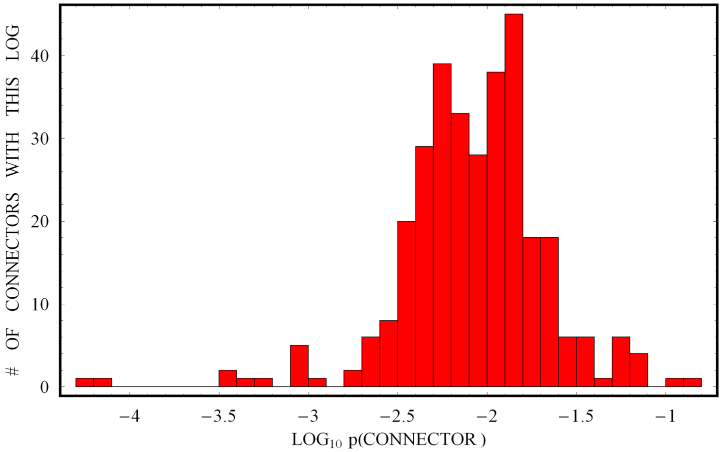

Fig. 3. Distribution of *p*-function values for 500 mass connectors on $\log_{10}$ scale. The Y-axis bin size was 0.1 with bin centers at 0.05, -0.05, -0.15, etc. The distribution contains data for 321 connectors with non-zero *p*-function values.

Remarkably the values of *p*-function are almost independent of all protein DB properties, except frequencies of 20 standard amino acids. We have demonstrated this independence by computing *p*-function recursively without keeping track of particular amino acid combinations leading to a given mass to avoid combinatorial explosion. The process is known as a renewal process [20], as was also pointed out in [16]. The calculations start from bins filled with single amino acids and continue until bin number of 2000.

One can also compute an "experimental" *p*-function for a given database. One just has to generate all peptides, compute their masses, and figure out occupancy frequencies for all [*m-dm, m+dm*] windows of interest. For example, for human genome DB the *p*-function can be obtained as a histogram of approximately ~2×10^8 values (approximate number of peptides in range mass 0 to 2000), which are distributed over 2000 bins. After normalizing histogram by the DB length, we obtained p-function that was almost indistinguishable from the theoretically computed (shown in Fig. 2). Some of the bins are empty. They correspond to mass connectors that could never be observed for a true protein tag. There are many such bins at masses below 200, as this region is occupied mostly by short amino acids combinations.

It is instructive to understand why the model that assumes total independence of the consecutive amino acids provides such a good approximation to reality, while it is known that the real protein text has short and long range sequence correlations. The reason is a "combinatorial elimination" of the correlation artifacts. For example, though combinations like AAAA, QQQQ, and similar ones are much more frequent than it would be expected from uncorrelated model, their contribution changes the *p*-function values only a little, because there are much more other combinations in the same mass bin, which do not show any statistical bias.

The distribution of $\log_{10}$ p values for all 321 non-empty bins found in the interval of m values between 0 and 500 is presented in Fig 3.

2.2 Probability of a Random Database Match

Calculated p functions allow us to compute the filtering power of an arbitrary tag (x_1 x_2 ... x_l), where l is the length of the tag. As we assume that there are no correlations between adjacent connectors (highly reasonable assumption for almost all connectors), the total probability of a random realization for a general tag is:

$$p_\Pi = \prod_{i=1}^{l} p(x_i)$$

When the tag p-function is calculated, it is straightforward to estimate the probability of a random match for a given database. As the match can start from any of N_D amino acids in the database, the average expected number of matches is $r = N_D p_\Pi$. The distribution over the observed number of m matches is binomial with the probability p_Π and the number of Bernoulli trials N_D (size of the DB), but since p_Π is small, and N_D is large, it is possible use its Poisson approximation:

$$p(m) = \frac{r^m}{m!} e^{-r}$$

The probability to observe at least one random match is:

$$p_{m>0} = 1 - p(0) = 1 - e^{-r} \approx N_D p_\Pi$$

The last expression is an approximation for the case of strong filtering $N_D p_\Pi << 1$. The same result can be obtained from the Binomial distribution. $p_{m>0}$ or a complementary quantity $p(0) = e^{-r} \approx 1 - N_D p_\Pi$ is a natural measure of database match significance. For a single tag, the match can be considered as non-random with the confidence level $p(0)$.

We have tested our theory by conducting tag match experiments for 10^4 mass tags that were generated in the following way:

(1) Random tag length l is chosen from values 2,3,4 and 5.
(2) Random integer numbers are chosen from the interval 57 and 2000, and converted to real values by random selection from centers of the corresponding Mann bins with Gaussian accuracy σ=0.15 (corresponds to the observations on 0.5 Da accuracy devices). These numbers form the testing tag.
(3) Tag probability p_Π is calculated. If p_Π=0, the tag is discarded, and we return to the step 2.

The results of the tag matches against a large DB (N_D~12×10^6) are presented on Fig. 4. Here, the p-function was constructed from the experimental statistics, but the theoretical function yields similar results. Both the expected and the observed number of matches for the tag of given length span more than 2 orders of magnitude (red, green dots), but, nevertheless, there is a very high correlation between the two.

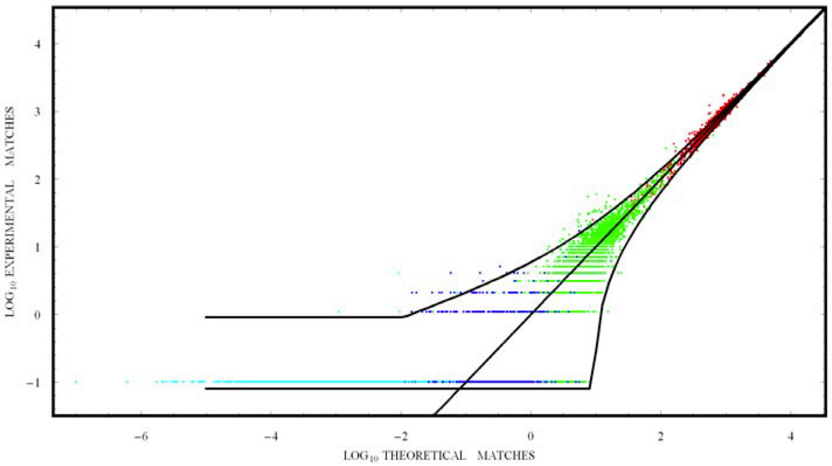

Fig. 4. Log-Log plot of experimental vs. theoretical number of matches. Human proteome DB (N_D=12×10^6). Tag's lengths are shown in color: red (2), green (3), blue (4), cyan (5). Solid lines show the diagonal and the confidence ranges.

Nevertheless, the figure reveals appreciable number of overmatched tags. The confidence ranges at both ends (black curves) are at 10^{-4} level, while the total number of tags was 10^4, so that all points beyond this range are statistically significant. An appreciable amount of such tags is of length 3 (green) or 4 and even 5 (blue, cyan). In some typical cases, the tag of length 3 with the expected number of matches of about 6 demonstrates about 30 of them, which is of course highly improbable. Such cases were investigated and all of them appeared to match identical peptides, resulting from homologies present in any real database.

Fig. 4 contains several unexpected lessons for mass spectrometry identification methods. For green points the number of expected and observed matches concentrates around log-value 1 (ten matches). The green points correspond to tags of 3 mass connectors or, in other words, containing just 2 real ions; and DB size here is larger than in a typical MS experiment. It means that in a real experiments it may be sufficient (in many cases) to find just two true ions to uniquely identify underlying peptide.

2.3 P-Value for High throughput Identification

The calculated probability of random tag matches provides an immediate opportunity to compute the "p-value" of peptide identification in high throughput tag search. Let us consider an asset of tags T={t_i} generated by some tag-selecting algorithm. For each tag we will consider 3 outcomes:

(1) No match: there is no place in the database, where all imposed constraints are satisfied simultaneously

(2) Correct match: tag has matched the correct peptide. By definition it is a single match, and it always happens if the database is complete and does not include mistakes (as we assume everywhere in this paper)
(3) Random match(es): one or more matches that do satisfy all connector conditions. This outcome is not mutually exclusive with outcome (2). Generally, it is possible to have a correct match together with several random ones.

Assuming that for all tags $r_i << 1$, we will get a number of random matches for the whole set $N_{random} = \Sigma r_i$. The sum must be computed through all tags, including those that were never matched. Combining this sum with the observed total number of matches M, we can write a formula for the algorithm sensitivity C, i.e. fraction of correct matches over the total number of matches:

$$C = \frac{M - \sum r_i}{M}$$

This estimate will work for *any* tag-generating algorithm and for any database. It does not require machine learning procedures or an adaptation to a particular database. It also accounts for the precision of ion detections, as well as other possible constraints on peptides (such as tryptic or nontryptic parent peptide). The only requirement is a complete separation between the process of tag construction and tag matching. The part responsible for tag generation should not have "backdoor" access to the database and use only information contained in the spectra itself to generate the tag.

Tag based approaches open new algorithmic possibilities for analysis of the proteomics spectra. Our analysis shows that many different strategies can be pursued, but one has to take into account that informational value of tags differ by 4 (!) orders of magnitude, and it is true even inside the group of tags of length 5. It also seems unwise to consider very few tags in the searches, as the number of random matches can be tightly controlled.

2.4 Computing SEQUEST X-Corr Values

Now we can propose a possible way to calculate analytically "black box" of the SEQUEST X-corr function. We aim to estimate the following: for a given spectrum S and a given database DB, which is a decoy database for the spectrum, find probability to obtain an X-corr value above of a given cutoff CT. The calculation can be accomplished by the following algorithm:

(1) Recalibrate spectrum by the usual SEQUEST procedure;
(2) Determine all groups of peaks that have sum of the recalibrated intensities above CT;
(3) Calculate p_i — p-function value for tags formed by each of those groups. Each subgroup forms a single tag. In addition to all real peaks, two "pseudo" peaks positioned at the zero mass and at the parent mass are added.
(4) The sum Σp_i provides a very good estimate that one of those scenario will realize, and the search procedure will detect an X-corr above CT. More precise estimation can be obtained by considering dependencies between overlapping tags.

3 Conclusions

We presented a rigorous mathematical formalism quantifying the probability of random DB matches for arbitrary tags extracted from the tandem MS spectra. It is shown that the tags consisting of ions separated by some hundreds of Da are in many cases more advantageous then tags consisting of shorter connectors. For example 3-ion tags (and in some cases even 2-ion tags) may suffice for an unambiguous identification in the non-redundant human DB. Developed approach allows a reliable quantification of the expected probability. The random match probabilities for the tags of similar length may differ by several orders of magnitude and are log-normal distributed.

The observed number of random DB matches obeys the Poisson distribution with the mean value calculated as the product of probability of realization for the given mass tag and the database size. This holds even for the tags that differ by several orders of magnitude in the random match probability and observed number of matches. The deviations from this law are shown to be almost exclusively due to homologies present even in the curated non-redundant databases.

Possible extensions of suggested approach include generalizations to arbitrary experimental accuracy, sequence correlations, consideration of database errors, as well as theoretical estimates for background values in many scoring functions currently existing in the field (including SEQUEST X-corr function).

Acknowledgments. This work was funded by a Biopilot project from the DOE Office of Advanced Scientific Computing Research. We also wish to thank Max Fridman for help with the manuscript.

References

1. Hirosawa, M., Hoshida, M., Ishikawa, M., Toya, T.: MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. Comput. Appl. Biosci. 9, 161–167 (1993)
2. Eng, J.K., McCormack, A.L., Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry 5, 976–989 (1994)
3. Yates III, J.R., Eng, J.K., McCormack, A.L.: Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. 67, 3202–3210 (1995)
4. Tabb, D.L., McDonald, W.H., Yates III, J.R.: DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. 1, 21–26 (2002)
5. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567 (1999)
6. Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392 (2002)
7. Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658 (2003)

8. Kapp, E.A., Schutz, F., Connolly, L.M., Chakel, J.A., Meza, J.E., Miller, C.A., Fenyo, D., Eng, J.K., Adkins, J.N., Omenn, G.S., Simpson, R.J.: An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5, 3475–3490 (2005)
9. Higdon, R., Hogan, J.M., Van Belle, G., Kolker, E.: Randomized sequence databases for tandem mass spectrometry peptide and protein identification. Omics 9, 364–379 (2005)
10. Higdon, R., Hogan, J.M., Kolker, N., van Belle, G., Kolker, E.: Experiment-specific estimation of peptide identification probabilities using a randomized database. Omics 11, 351–365 (2007)
11. Huttlin, E.L., Hegeman, A.D., Harms, A.C., Sussman, M.R.: Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. J. Proteome Res. 6, 392–398 (2007)
12. Qian, W.J., Liu, T., Monroe, M.E., Strittmatter, E.F., Jacobs, J.M., Kangas, L.J., Petritis, K., Camp II, D.G., Smith, R.D.: Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. J. Proteome Res. 4, 53–62 (2005)
13. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214 (2007)
14. Choi, H., Ghosh, D., Nesvizhskii, A.I.: Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J. Proteome Res. 7, 286–292 (2008)
15. Mann, M., Wilm, M.: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. 66, 4390–4399 (1994)
16. Sunyaev, S., Liska, A.J., Golod, A., Shevchenko, A., Shevchenko, A.: MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal. Chem. 75, 1307–1315 (2003)
17. Frahm, J.L., Howard, B.E., Heber, S., Muddiman, D.C.: Accessible proteomics space and its implications for peak capacity for zero-, one- and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. J. Mass Spectrom 41, 281–288 (2006)
18. Mann, M.: Useful tables of possible and probable peptide masses. In: 43rd ASMS Conference on Mass Spectrometry and Allied Topics, Am. Soc. Mass Spectr., Atlanta (1995)
19. Zubarev, R.A., Hakansson, P., Sundqvist, B.: Accuracy Requirements for Peptide Characterization by Monoisotopic Molecular Mass Measurements. Anal. Chem. 68, 4060–4063 (1996)
20. Kampen, N.G.v.: Stochastic processes in physics and chemistry. North-Holland, Amsterdam, New York (1992)