RESEARCH ARTICLE

Optimizing search conditions for the mass fingerprint-based identification of proteins

Elena Ossipova[1], David Fenyö[2] and Jan Eriksson[1]

[1] Department of Chemistry, Swedish University of Agricultural Sciences, Uppsala, Sweden
[2] The Rockefeller University, New York, USA

Received: June 28, 2005
Revised: September 29, 2005
Accepted: October 3, 2005

The two central problems in protein identification by searching a protein sequence collection with MS data are the optimal use of experimental information to allow for identification of low abundance proteins and the accurate assignment of the probability that a result is false. For comprehensive MS-based protein identification, it is necessary to choose an appropriate algorithm and optimal search conditions. We report a systematic study of the quality of PMF-based protein identifications under different sequence collection search conditions using the Probity algorithm, which assigns the statistical significance to each result. We employed 2244 PMFs from 2-DE-separated human blood plasma proteins, and performed identification under various search constraints: mass accuracy (0.01–0.3 Da), maximum number of missed cleavage sites (0–2), and size of the sequence collection searched (5.6×10^4–1.8×10^5). By counting the number of significant results (significance levels 0.05, 0.01, and 0.001) for each condition, we demonstrate the search condition impact on the successful outcome of proteome analysis experiments. A mass correction procedure utilizing mass deviations of albumin matching peptides was tested in an attempt to improve the statistical significance of identifications and iterative searching was employed for identification of multiple proteins from each PMF.

Keywords:
Bioinformatics / MS / Plasma proteome / Proteome analysis / Protein identification

1 Introduction

Accurate proteome analysis plays a leading role for understanding many physiological processes occurring in an organism. Proteome analysis employs several methods, among them is the identification of proteolytically digested proteins using MS and searching of a sequence collection [1–3]. To comprehensively identify proteins using MS data, it is necessary to choose an appropriate identification algorithm [4–11] and to find optimal search conditions. We here demonstrate the influence of different sequence collection search constraints on the statistical significance of PMF-based protein identification results.

A proteome is an intricate and complex system that represents all possible gene products and contains many thousands of different proteins that are present in a wide dynamic range [12–14]. The complexity of the proteome is amplified by the fact that several post-translationally modified forms of each gene expressed can be present in the proteome at the same time. Before the proteins are identified it is often necessary to reduce the complexity of the mixture. Protein separations technologies use various methods based on the differences in physical or chemical properties of different proteins [15, 16]. One of the most common techniques is the separation by 2-DE, which separates proteins by their respective p*I* and molecular weight [16]. After the visualization of the proteins in the gel, the spot of interest can be cut out and in-gel digested by an enzyme with high specificity (usually trypsin). The set of proteolytic peptides obtained in

Correspondence: Dr. Jan Eriksson, Department of Chemistry, Swedish University of Agricultural Sciences, Box 7015, 750 07, Uppsala, Sweden
E-mail: jan.eriksson@kemi.slu.se
Fax: +46-18-67-3476

Abbreviations: Δm, mass accuracy; u, maximum number of missed cleavage sites

this manner is unique for every protein and, hence, an MS analysis of the proteolytic peptides provides a fingerprint of each protein. The PMF map can be recognized when searching a collection of protein sequences [2, 4, 7, 8, 11, 17]. There is a risk of obtaining a false identification, because each mass determined by MS has an error, $\pm\Delta m$, and can match several proteolytic peptides of various proteins in the sequence collection [18, 19]. The potential for obtaining a true MS protein identification result depends on the choice of algorithm, as well as on experimental factors that influence the information content in the MS data. Current methods can never definitely prove that a result is true, but an appropriate choice of algorithm can provide a measure of the statistical risk that a result is false, *i.e.*, the statistical significance [20].

The two central problems in protein identification by searching a protein sequence collection with MS data are the optimal use of the experimental information to allow for identification of low abundance proteins and the accurate assignment of the probability that a result is a false positive. A potential consequence in the analysis of a low abundance protein is that it yields no or only weak MS signals. In the context of PMF, weak signals suggest that a lower number of peptides might be detected. A low number of peptides yields a more challenging situation for confident protein identification. Here we report the first systematic study using the Probity algorithm [20] to examine the quality of experimental protein identification results obtained under different sequence collection search conditions. The examination of results obtained from a large number of proteolytic peptide mass fingerprints from 2-DE-separated human blood plasma proteins demonstrates that the choice of search conditions impacts the successful outcome of proteome analysis experiments.

2 Materials and methods

Human plasma (1 mL) was prefractionated by depleting albumin and IgG (Blue Sepharose HP and protein G-Sepharose HP, GE Healthcare, Uppsala, Sweden) and the remaining proteins were separated with anion exchange chromatography (Resource Q, GE Healthcare) using an ÄKTA FPLC system (GE Healthcare). Twenty-five protein fractions were collected and freeze dried before separation by 2-DE. Preparative 2-D DALT gels (GE Healthcare) were run and protein spots automatically handled using Ettan Spot handling workstation (GE Healthcare) and the proteolytic peptides were mass analyzed with Ettan MALDI-TOF Pro (GE Healthcare). Prior to the proteolytic peptide mass analysis, the mass spectrometer was externally calibrated using two peptides (angiotensin III, 897.5 Da, and residues 18–39 of human adrenocorticotropic hormone, 2465.2 Da). The proteolytic peptide mass spectra were automatically recalibrated using the trypsin autolysis peaks with masses 841 and 2211 Da when present. For the analysis of the impact of search conditions on the quality of protein identification results, 2244 PMFs from the experiment were employed. Monoisotopic mass values of peaks detected within a mass region between 800 and 4500 Da were extracted from each spectrum. The Probity algorithm [20] was applied for the protein identification. The identification was performed under a variety of search constraints: mass accuracy (Δm), maximum number of missed cleavage sites (u), and size of the sequence collection. Δm values were varied between 0.01 and 0.3 Da and u values were varied between 0 and 2. The number of proteins in the sequence collection searched was varied from 5.6×10^4 to 1.8×10^5. The smallest sequence collection contained human protein sequences only (IPI, 56 522 sequences) [21] and additional sequence collections were constructed using the IPI-sequence collection plus various combinations of the entire *Mus musculus* (78 059 sequences, NCBInr) and *Rattus norvegicus* (32 227 sequences, NCBInr) protein sequence collections. The number of significant results (at the 0.05, 0.01, and 0 001 significance levels, respectively) was counted for each set of search constraints tested. Only the highest ranked protein for each search was considered as an identification result.

The human plasma proteome is one of the most sampled proteomes and contains up to 55% albumin [12]. It has been observed that blood plasma samples yield a large number of 2-DE spots (and corresponding PMFs) that contain albumin [12]. Here this observation was utilized in a procedure of mass correction of the data subsequently to a first series of sequence collection searches. The mass error measured as the deviation between experimental and theoretical mass values for matching albumin peptide masses was examined for all highly significant albumin identification results when using Δm = 0.05–0.26 and u = 1. To potentially account for differences in the experimental mass accuracy in different regions of the mass scale (*e.g.*, due to imperfect calibration), we divided the mass scale into eight different mass regions [20, 22], and examined the mass errors of albumin-matching peptides in each region separately. The centroid of each distribution of albumin errors was measured and added to each mass value in each respective mass region in every spectrum of our data set. The assumption underlying this procedure is that the addition of the centroid value of the albumin error distribution would to some extent compensate for systematic errors in the initial calibration.

The observation that many 2-DE spots from plasma contain albumin, and the known function of albumin as a carrier of other proteins, suggest that many spots potentially contain more than one protein. Iteration has been demonstrated as a means for identifying the individual protein components of mixed protein fingerprints extracted from gels, using a simple identification algorithm [23], and, for very complex *in-silico* generated mixtures, using Probity [24]. Here the iterative identification procedure was employed for each respective PMF to test if more proteins could be significantly identified from each gel spot. Hence, in the first step all the mass

values were submitted to Probity. The first identification result was obtained, and masses corresponding with the protein identified were excluded from the data set. The remaining masses were resubmitted to Probity and again the masses corresponding with the protein identified were excluded. This process was repeated four times.

Scripts written in *Perl* were used for all the computations, which were performed on a Dell Optiplex GX280 (2.8 GHz Pentium(R) IV) and Dell (2.66 GHz Pentium IV) personal computers.

3 Results

3.1 Analysis of the data

A detailed analysis of the impact of search conditions on the outcome of protein identification results was performed using a plasma proteome sample based on 2244 PFMs and the Probity algorithm. Figure 1 displays the significance value for all the 2244 identification results assuming search conditions often employed as default (u = 1, Δm = 0.2 for externally calibrated MALDI data). It is seen in Fig. 1 that 40% of the results were significant at the 0.05 significance level, 31% of the proteins were identified at the 0.01 significance level, and 22% of the results were significant at the 0.001 significance level. The fraction of identifications that yielded significant results was monitored for these three significance levels for the various search conditions investigated. The number of unique and statistically significant protein identification results under these conditions was 74 (significance level 0.05).

3.2 Mass correction

The examination of deviations between theoretical and measured values of proteolytic peptide masses for highly significant albumin results (0.001 level) was performed over a broad range of Δm values. Centroid values for the distribution of deviations were determined in different mass regions [20] for each Δm employed. We observed that below

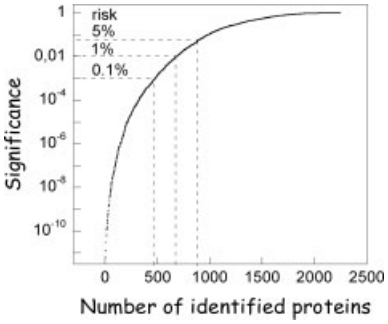

Figure 1. Statistical significance for all proteins identified in our data set assuming Δm = 0.2 and u = 1. The dashed lines indicate the three significance levels monitored.

Δm = 0.2 the centroid values scattered between different Δm values, whereas for Δm values of 0.2, 0.23, and 0.26, the centroid values remained approximately constant. Therefore, the centroid values determined for Δm = 0.2 were employed as our mass correction values added to all PMFs in our data set. The use of the mass correction procedure improved the fraction of the results considered significant for all search conditions tested (see *e.g.* Fig. 2, left).

3.3 Effect of different Δm values

To find conditions that make the optimum use of the information content in mass spectra, we investigated the dependence of the number of significant results on different values of Δm employed in the sequence collection search. The searching with different values of Δm was tested for the entire set of PMFs prior to, as well as after, the mass correction procedure described above. It is seen in Fig. 2 (left) that the number of significantly identified proteins is higher for the mass corrected data over a broad range of Δm values. Figure 2 (left) shows that the maximum number of significantly identified proteins for the mass corrected data is achieved at $0.1 < \Delta m < 0.17$, whereas for the data without mass correction the maximum number of significantly identified proteins is achieved at $0.17 < \Delta m < 0.2$ for all the

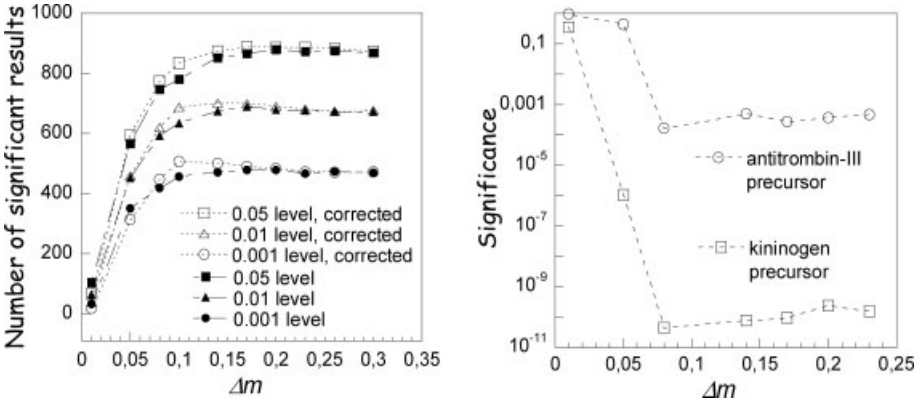

Figure 2. Left: The number or proteins identified by Probity at different significance levels as a function of the Δm value employed in the sequence collection search before and after a mass correction procedure utilizing the error distribution of albumin (see text for details). Right: The influence of Δm on the statistical significance computed for two different peptide mass fingerprints randomly chosen from the mass corrected data set.

three significance levels investigated. The result displayed in the left panel of Fig. 2 indicates that the significance values of many protein identification results vary with Δm. Of course, the precise influence of different Δm-values can vary between different PMFs. Two specific protein examples of how the significance is influenced by Δm are shown in the right panel of Fig. 2. These two examples display similar trends but also demonstrate that the magnitude of the impact of Δm on the significance can vary broadly between different PMFs.

3.4 Influence of the chosen *u* value

The assumed number of missed cleavage sites, *u*, can influence results of protein identification since the number of potentially matching mass values when searching the sequence collections varies with *u*. It is seen in the left panel of Fig. 3a that when including results from all PMFs the value $u = 1$ yielded the best performance of the search for all significance levels monitored. The result displayed in the left panel of Fig. 3 shows that the significance values of many protein identification results vary with *u*. However, the exact influence of different *u* values can vary between different PMFs. It is seen in the left panel of Fig. 3b that, when excluding the substantial fraction of the PMFs employed that yielded albumin as the identification result, the overall dependence on *u* is changed dramatically. With albumin excluded, $u = 0$ is the most favorable condition. Hence, albumin displayed a different dependence on *u* than many other proteins identified. This observation indicates that the exact influence of different *u* values can vary broadly between different PMFs. This phenomenon is further illustrated in the right panel of Fig. 3, which displays distinctly different results from three different PMFs randomly chosen from the data set.

The observation of a broad variation in the optimum setting of *u* between different PMFs suggests that the use of a fixed *u* value for a data set is a sub-optimized condition. When using the *u* value that yielded the best significance value for each respective PMF instead of a fixed setting of *u* for all PMFs, the number of significant results from our data set was enhanced by 29% (40%), 23% (50%), and 28% (63%) for the 0.001, 0.01, and 0.05 significance levels, respectively, as compared with the best result for the fixed *u* value setting (results including albumin in parenthesis). Using this optimization procedure increased the number of unique and statistically significant protein identification results (significance level 0.05) from 74 to 103.

3.5 Sequence collection size

The influence of the number of sequences searched on the statistical significance was investigated. It is seen in Fig. 4 that the number of significant results decreased as a function of an increasing number of sequences searched.

3.6 Iteration

It has been shown [23] that using an iterative search procedure can reduce the complexity of the PMFs, and hence decrease the problem of random matching and the risk of obtaining false results. We here examined our data set with and without employing iteration. Figure 5 displays the improvement of the number of significant results when the iterative procedure is used. It is seen that the iteration clearly yielded an improvement for the 0.05 and 0.01 significance levels, whereas for the 0.001 level the same number of results are significant as when not using iteration. Nineteen unique identifications (significance 0.05) were gained by employing the iterative procedure.

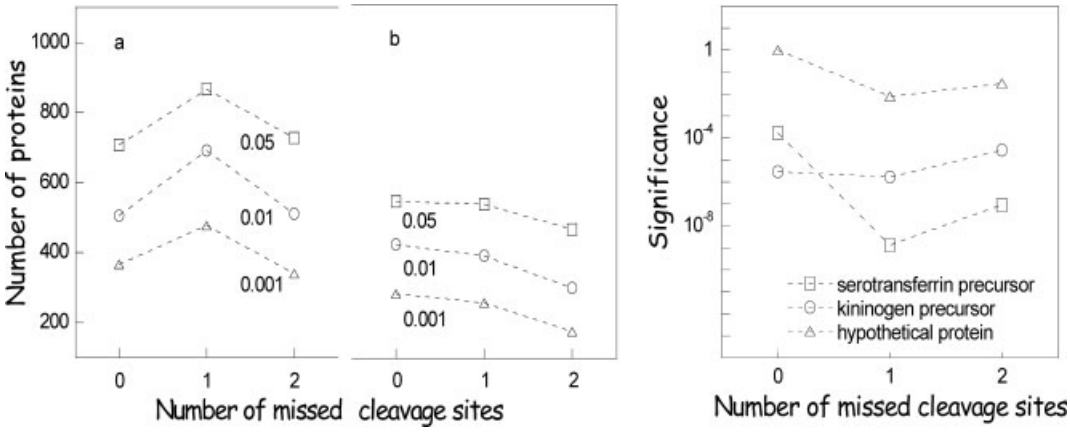

Figure 3. Left panel: The influence of different *u* values on the number of significant results for three different significance levels. (a) All significant results from the data set were counted. (b) All significant results except albumin results were counted. Right panel: The statistical significance of the protein identified displayed as a function of *u* for three randomly chosen PMFs.

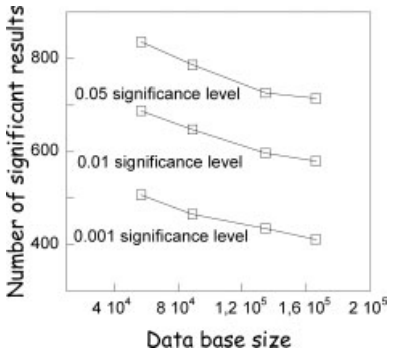

Figure 4. The number significant protein identification results as a function of different number of protein sequences searched.

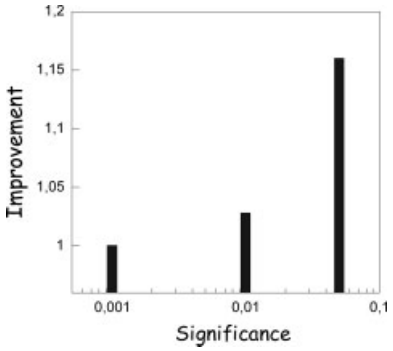

Figure 5. The ratio between the total number of identified proteins using an iterative search procedure and the number of identified proteins without iteration for three different significance levels.

4 Discussion

4.1 Mass correction and the choice of Δm

The procedure for calibrating MALDI-TOF spectra influences the resulting accuracy of proteolytic peptide mass values [25]. The use of internal calibration, *i.e.*, one or several substances with known masses are present in the analyte sample, typically yields a better accuracy than external calibration that relies on performing the calibration with a separate sample of known substances. Here, internal calibration was employed for the spectra that yielded signal for one or two trypsin autolysis peptide mass values, whereas the remaining spectra were externally calibrated. It is possible that search conditions (Δm values) could be adjusted according to whether an internal calibration was based on 0, 1, or 2 autolysis peaks. Such a procedure may, however, be suboptimal since the trypsin autolysis peaks can be overlapped by peaks from the digested protein, leading to imperfect internal calibration. Furthermore, it is often observed that the closer the mass value is to a mass employed in the calibration the better is the mass accuracy. Hence, even internally calibrated spectra in the present data set are expected to display differences in the accuracy for different regions of the mass range covered by the PMFs. In addition, peak detection algorithms that assign centroid values to the monoisotopic peaks detected are often imperfect in the sense that the accuracy of the assignment can depend on the individual peak intensity and shape. The concept of detailed evaluation of mass deviation distributions of matching tryptic peptides for each respective PMF has been described by Egelhofer *et al.* [26] as a means for ranking protein sequences in the sequence collection, and by Magnin *et al.* [27] for the purpose of data filtering. The simplified and related procedure employed here, *i.e.*, examination of deviations between theoretical and measured values of proteolytic peptide masses for highly significant results (albumin) followed by a correction to all mass values in all PMFs based on the centroid value of the distribution of mass deviations for many peaks over the entire mass scale, could potentially compensate for any systematic error in the mass calibration caused by the reasons stated above. The successful outcome of the use of the correction procedure indicates that systematic errors are reduced.

The examination of mass deviation distributions for albumin results also provides information on what Δm values could be reasonable for optimizing the number of significant results. We found that the average SD of the mass deviation distributions in the different mass regions was 0.05. The mass deviation distribution is bell shaped. Hence, if using a Gaussian approximation, about 95% of the mass values in the data would be detected within a mass deviation region of ± 0.1 Da and about 99% of the measured masses would be detected within ± 0.15 Da. This information *per se* suggests that optimum Δm values for searching the sequence collection would approximately be $0.1 < \Delta m < 0.15$. This optimum of Δm is supported by the systematic study of the influence of Δm as displayed in Fig. 2.

The intrinsic compensation in Probity for the risk of obtaining false results as the random matching increases with increasing Δm is expected to yield a decline in the number of significant results, as the Δm value is increased to unnecessarily high values. The results displayed in Fig. 2 do not follow this expected trend for high values of Δm. There are two plausible reasons for this observation. The first is that random matching occurs in a narrow region around each nominal mass value referred to as a tryptic peptide mass peak distribution [18]. The width of the tryptic peptide mass peak distribution increases with the nominal mass. The Δm values around 0.2 are already high enough to account for almost the entire theoretical tryptic peptide mass peak distributions [18] below nominal masses of 2500 Da, and hence only a minor increase of the risk of random matching occurs as Δm is increased to 0.3 Da [19]. The second is that the dominating protein identified in our data set is albumin (>300 significant results from 2244 PMFs). The systematic study of albumin peptide mass deviations indicated that there were peaks related to albumin that displayed

rather large mass deviations. In Fig. 2, albumin is included and, hence, for high values of Δm, albumin obtains additional matches that somewhat improve the significance and compensate for the increased risk of random matching. The almost constant number of significant results observed for $0.2 < \Delta m < 0.3$ is mainly due to the albumin results. When albumin is excluded from the data set, the number of significant results observed decreases between $\Delta m = 0.2$ and $\Delta m = 0.3$ (data not shown).

4.2 Number of missed cleavage sites, *u*

The demonstrated benefit for searching the optimum *u* value for each PMF could be related to the fact that different protein sequences might be differently susceptible to enzymatic cleavage. However, if a low concentration of incompletely digested peptides exists for each protein, these peptides would, due to a limited detection sensitivity, be detected only if the protein was initially present in a high concentration. Our data analysis suggests that the concentration of an individual protein in a sample can influence the distribution of the number of missed cleavage sites in the peptides detected. The number of matches with peptides that contained one or two missed cleavage sites was pronounced for albumin (Fig. 3) and also for the highly abundant fibrinogen protein.

The concept of searching for the optimum *u* value for each individual PMF leads to a strong improvement of the number of significant results. A necessary condition for this procedure to be sound is that the probability for a protein to obtain a particular number of matches by chance is computed accurately. As the value of *u* is increased, the number of matches with an individual protein sequence is either constant or increasing. The statistical risk that additional matches are random is computed by Probity. In the examples of the right panel of Fig. 3, two to three additional matches were obtained when changing *u* from 0 to 1, which reduced the overall risk that the results are false (*i.e.*, improved the statistical significance of the results), whereas changing the *u* value from one to two yielded higher risks of random matching due to no additional matches but an increased number of theoretical mass values that can match randomly for each protein sequence. The procedure of searching with a series of *u* values for each PMF increases the computation time. If computation time is considered a critical issue, an alternative approach could be to use a test data set for each experimental protocol to search for the *u* value that yields the best overall performance. We tested the potential of this approach by randomly selecting 300 PMFs out of the 2244 PMFs employed in the present work. The randomly selected PMFs and the entire data set displayed very similar dependences of the number of significant results as a function of *u*, which suggests that the use of a test data set could be a rapid means for enhancing the final number of significant results.

4.3 Sequence collection size

An increasing number of sequences searched leads to a reduction of the number of significant results (Fig. 4). This is due to the fact that several sequences can match randomly the experimental peptide mass fingerprint, and hence the risk to obtain a false result is increased. The results displayed here stress the desire to keep the best possible taxonomic precision in the search.

4.4 Iteration

The benefit of iteration is due to the fact that in each step of the iteration procedure the complexity of the respective peptide mass fingerprint is reduced and, hence, the problem of random matching is reduced. Here, the use of iteration yielded 57 new significant results that allowed us to identify 19 unique proteins that would have been missed without iteration. If the protein separation is excellent there is of course no need for the iterative search procedure. Although 2-DE is often assumed to display excellent separation, proteome analysis experiments typically reveal that many PMFs contain peptides from more than one protein. For plasma proteome analysis the risk that PMFs contain peptides from several proteins is potentially influenced by the presence of albumin, which can easily form complexes with other proteins. Albumin was identified in 350 PMFs in the first iteration step (significance 0.05). Detailed analysis of the significant identification results subsequently to the first iteration step showed that 28% of the identified proteins originated from PMFs where albumin was first identified.

4.5 Unique protein identifications

The 2244 PMFs employed in this study is a sample of the plasma proteome taken for the purpose of demonstrating the impact of the search conditions on the successful outcome of protein identification. The number of unique proteins identified was increased by >60% by employing various optimization efforts as compared with typical default conditions. None of the proteins uniquely identified after optimization belongs to the group of 22 highly abundant plasma proteins [28]. This suggests that the procedure of optimizing search conditions has the potential to benefit the identification of low abundance proteins.

5 Conclusions

The Probity algorithm, which computes accurately the risk that a result is false under any given search condition, was employed for the first time on a large experimental data set. This allowed us to demonstrate quantitatively how the choice of search conditions impacts the successful outcome of proteome analysis experiments. The analysis revealed that optimizing the number of missed cleavage sites, *u*, for each re-

spective peptide mass fingerprint has a strong influence on the number of significant results. A mass correction procedure based on mass deviation information of albumin-matching peptides obtained from a first pass identification yielded a clear-cut improvement of the results in a second pass identification. The optimum choice of the mass accuracy in the search, Δm, as well as the optimum u values were somewhat stricter than that usually expected in default assumptions. The results of the present study confirm previous simulation results showing that: (i) iterated searching with Probity can enhance the number of significant results, and (ii) maximized taxonomic precision in the sequence collection search maximizes the number of significant results. The optimization procedures increased the number of unique and statistically significant (0.05) protein identification results by >60% as compared with search conditions often assumed as default in this type of experiment.

We are grateful to Reid Asbury, Phil Beckett, Rama Bhikhabhai, Erik Forsberg, Anneli Jorsback, Helena Nordvarg, Raf Lemmens, and Stacey Loyland for providing the experimental data used in this study. This work was supported by a grant from the AgriFunGen program at SLU, Sweden.

6 References

[1] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[2] Beavis, R., Fenyo, D., *Proteomics: A Trends Guide*, Elsevier, London 2000, pp. 22–27.

[3] Krutchinsky, A. N., Kalkum, M., Chait, B. T., *Anal. Chem.* 2001, *73*, 5066–5077.

[4] Clauser, K. R., Baker, P., Burlingame, A. L., *Anal. Chem.* 1999, *71*, 2871–2882.

[5] Fenyo, D., Beavis, R. C., *Anal. Chem.* 2003, *75*, 768–774.

[6] Griffin, P. R., MacCoss, M. J., Eng, J. K., Blevins, R. A. *et al.*, *Rapid Commun. Mass Spectrom.* 1995, *9*, 1546–1551.

[7] Pappin, D. J. C., Hojrup, P., Bleasby, A., *Curr. Biol.* 1993, *3*, 327–332.

[8] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[9] Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C. *et al.*, *Methods Mol. Biol.* 1999, *112*, 531–552.

[10] Yates, J. R. D., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, *67*, 1426–1436.

[11] Zhang, W., Chait, B. T., *Anal. Chem.* 2000, *72*, 2482–2489.

[12] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[13] Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W. *et al.*, *Nature* 2003, *425*, 737–741.

[14] Tyers, M., Mann, M., *Nature* 2003, *422*, 193–197.

[15] Bjellqvist, B., Basse, B., Olsen, E., Celis, J. E., *Electrophoresis* 1994, *15*, 529–539.

[16] Fey, S. J., Larsen, P. M., *Curr. Opin. Chem. Biol.* 2001, *5*, 26–33.

[17] James, P., Quadroni, M., Carafoli, E., Gonnet, G., *Biochem. Biophys. Res. Commun.* 1993, *195*, 58–64.

[18] Eriksson, J., Fenyo, D., *Proteomics* 2002, *2*, 262–270.

[19] Fenyo, D., Qin, J., Chait, B. T., *Electrophoresis* 1998, *19*, 998–1005.

[20] Eriksson, J., Fenyo, D., *J. Proteome Res.* 2004, *3*, 32–36.

[21] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[22] Eriksson, J., Fenyo, D., *J. Proteome Res.* 2004, *3*, 979–982.

[23] Jensen, O. N., Mortensen, P., Vorm, O., Mann, M., *Anal. Chem.* 1997, *69*, 4741–4750.

[24] Eriksson, J., Fenyo, D., *J. Proteome Res.* 2005, *4*, 387–393.

[25] Gobom, J., Mueller, M., Egelhofer, V., Theiss, D. *et al.*, *Anal. Chem.* 2002, *74*, 3915–3923.

[26] Egelhofer, V., Gobom, J., Seitz, H., Giavalisco, P. *et al.*, *Anal. Chem.* 2002, *74*, 1760–1771.

[27] Magnin, J., Masselot, A., Menzel, C., Colinge, J., *J. Proteome Res.* 2004, *3*, 55–60.

[28] Haab, B. B., Geierstanger, B. H., Michailidis, G., Vitzthum, F. *et al.*, *Proteomics* 2005, *5*, 3278–3291.

