research articles Journal of proteome research

Protein Identification in Complex Mixtures

Jan Eriksson[*,†] and David Fenyö[‡,§]

Department of Chemistry, Swedish University of Agricultural Sciences, Box 7015, SE-750 07, Uppsala, Sweden, GE Healthcare, 800 Centennial Avenue, Piscataway, New Jersey 08855, and The Rockefeller University, 1230 York Avenue, New York, New York 10021

Received October 13, 2004

This paper investigates the prospects of successful mass spectrometric protein identification based on mass data from proteolytic digests of complex protein mixtures. Sets of proteolytic peptide masses representing various numbers of digested proteins in a mixture were generated in silico. In each set, different proteins were selected from a protein sequence collection and for each protein the sequence coverage was randomly selected within a particular regime (15−30% or 30−60%). We demonstrate that the Probity algorithm, which is characterized by an optimal tolerance for random interference, employed in an iterative procedure can correctly identify >95% of proteins at a desired significance level in mixtures composed of hundreds of yeast proteins under realistic mass spectrometric experimental constraints. By using a model of the distribution of protein abundance, we demonstrate that the very high efficiency of identification of protein mixtures that can be achieved by appropriate choices of informatics procedures is hampered by limitations of the mass spectrometric dynamic range. The results stress the desire to choose carefully experimental protocols for comprehensive proteome analysis, focusing on truly critical issues such as the dynamic range, which potentially limits the possibilities of identifying low abundance proteins.

Keywords: protein identification • algorithm • bioinformatics • mass spectrometry • proteomics • protein • protein mixtures • complex mixtures • peptide • peptide mapping • significance testing • simulation

The identification of components in protein mixtures by proteolytic peptide mass fingerprinting has been demonstrated experimentally,[1,2] but a more thorough examination of the prospects of mass fingerprint based identification in mixtures has been lacking. We here present a simulation study that elucidates the great potential for the development of successful methods for the confident identification of the components of very complex protein mixtures using the information of proteolytic peptide masses. Our study also elucidates the reason underlying current limitations for mixture analysis.

A proteome is inherently complex in its nature and can contain many thousands of different proteins. The proteome complexity reaches beyond the number of genes of the organism, since coding regions of each gene can, for some organisms, be spliced into different mRNAs and each protein translated may undergo post-translational modification. Challenges of proteome analysis are due to this complexity but also due to the large range of expression levels of genes that is expected to be of the order of 10^6 or higher[3,4] and according to some hypotheses can be as large as of the order of 10^{10}.[5]

Proteome analysis often involves separation at the protein or the peptide level or at both levels. Separation can be obtained in different dimensions by utilizing the differences in physical or chemical properties of the molecules. The factor of reduced complexity can be as large as 50−100 for each dimension of electrophoresis or reversed-phase high-pressure liquid chromatography (RP−HPLC). The typical proteolytic peptide mass fingerprinting experiment begins with protein separation by 1- or 2-dimensional gel electrophoresis (1DE or 2DE).[6−8] Each gel-band or spot of interest is excised and the protein is subjected to in-gel digestion using an enzyme that has high digestion specificity (e.g., trypsin). The resulting proteolytic peptides are extracted and analyzed by mass spectrometry (MS). It is assumed that a set of proteolytic peptide masses measured by MS provides a “fingerprint” of a particular protein,[9] and it is assumed that the peptide mass fingerprint can be recognized when searching a collection of protein sequences[10] derived from genomic information.[11−15]

Comigration of proteins often occurs in gels and therefore the resulting mass fingerprint can contain contributions from several proteins.[1] To handle complex mass fingerprints resulting from imperfect separation by 2DE Jensen et al.[1] introduced the concept of searching the collection of sequences in an iterative manner, where the masses matching the protein identified in the first step were removed prior to a second step of searching etc. Using that procedure, they demonstrated the identification of up to five components of proteins unresolved by electrophoresis using matrix-assisted laser desorption ionization mass spectrometry at a mass accuracy of 30 ppm and the Pep-

[*] To whom correspondence should be addressed. E-mail: jan.eriksson@kemi.slu.se.
[†] Department of Chemistry, Swedish University of Agricultural Sciences.
[‡] GE Healthcare.
[§] The Rockefeller University.

10.1021/pr049816f CCC: $30.25

Published on Web 02/09/2005

tideSearch algorithm. The iterative approach reduces the problem of random mass matching[16,17]—a general problem that is particularly severe when analyzing mixtures and using algorithms that rank the proteins of the sequence collection strictly by their number of theoretical proteolytic peptide masses matching the masses in the data (some algorithms however utilize more sophisticated ranking methods, see e.g., Refs 18−20).

The work of Jensen et al.[1] is an example of enhancing the sophistication of the processing of the information present in the mass data in order to handle an undesired imperfection of the separation part of the experiment. A different experimental approach is to deliberately go for a fast experimental procedure that yields highly complex data potentially containing information on all the proteins of the proteome analyzed. An example of this latter approach was presented by Ramström et al.,[2] who performed protein identification using data acquired by Fourier transform ion cyclotron resonance MS of on-line-RP−HPLC separated and electrosprayed ions from a complex mixture of tryptic peptides from proteins in human cerebrospinal fluid. 6551 monoisotopic masses with an accuracy of 5 ppm were entered in the searching of a sequence collection and resulted in the identification of 39 proteins. The prospects of judging the general applicability of this exciting approach is hampered by the fact that they searched a sequence collection composed of only 150 proteins already identified as being associated with the human body fluids.

To examine the prospects of mixture identification in a more general way, we here performed simulations of protein identifications using synthetic but realistic data. The Probity algorithm,[20] which is characterized by a very good tolerance against random background, was employed in an iterative manner. The Probity algorithm accurately assigns the statistical significance, i.e., the risk that the result is false, to each result, which allowed us to monitor the quality of the results as a function of the number of iterations.

A necessary condition for the successful analysis of a proteome is of course that peptides of the various proteins are detectable by MS. Each step of separation potentially introduces losses of proteins or peptides that can result in insufficient amounts of molecules for detection by MS. When considering the analysis of mixtures, it is critical that a mass spectrometer can detect simultaneously ions resulting from peptide species originating from different proteins present in different amounts in the proteome. Typical dynamic ranges of mass spectrometers are in sharp contrast with the hypothesized ranges of levels of expressed genes. Therefore, we also simulated the influence of a limited mass spectrometric dynamic range on the possibilities of detecting ions originating from complex protein mixtures.

The inherent large scale of comprehensive proteome analysis makes systems for high-throughput and automation highly desirable. This paper demonstrates that a parallel handling of experimental mass information is possible provided that successful efforts are made to minimize limitations due to the mass spectrometric dynamic range.

Materials and Methods

Simulation A. Generation of Synthetic Mass Data. Sets of proteolytic peptide masses representing various numbers of digested proteins in a mixture were generated. In each set, the *proteins* (MW < 100 000) were *randomly* selected from a protein sequence collection (*Saccharomyces cerevisiae*, 6403 ORFs, NCBI, May 2000 release). Protein digestion was performed in silico assuming exposure to trypsin (cleaves with high specificity at the carboxyl side of lysine and arginine residues). For each protein in a data set, the *sequence coverage* (expressed as a fraction of the total number of proteolytic peptides with monoisotopic masses between 800 and 4500 Da that a protein sequence can yield) was *randomly* selected within a particular regime (15−30% or 30−60%). *Peptides* (with mass values between 800 and 4500 Da) were selected *randomly* from each randomly selected protein sequence until the sequence coverage randomly selected was reached. Experimental mass errors were simulated by altering the mass of each peptide selected by adding a number randomly selected from a Gaussian distribution (standard deviation = $0.5\Delta m$).[18]

Protein Identification. Simulation of protein identification was performed using the Probity algorithm[20] to search the *S. cerevisiae* sequence collection assuming no missed cleavage sites and a maximum mass deviation of $\pm\Delta m$. Here, the Probity algorithm was employed in an iterative manner. The statistical significance, i.e., −the statistical risk that a result is false (random),[16,17,20−22] of the highest ranked protein was determined in each iteration step. In each new step, the masses matching the highest ranked protein in the previous step were removed from the set of masses. The iterative procedure was compared with the procedure of using a single-step search and monitoring the ranking list. The results were compared with the list of proteins present in the data to check whether results were true or false.

Simulation B. A necessary condition for MS-based protein identification is that proteolytic peptide ions can be detected. Limitations of detection sensitivity and dynamic range of MS can potentially hinder the detection of various peptides in a sample subjected to MS. The dynamic range is of particular interest when considering protein identification in mixtures, since the amount of peptide molecules from each respective protein can be very different. We therefore investigated the prospects of *detecting* ions originating from protein mixtures by MS. The investigations were based on various hypothetical assumptions about the dynamic ranges of gene expression and mass spectrometers—i.e., the distribution of the abundance of various proteins in a sample and the ability of a mass spectrometer to simultaneously detect ion signals from proteins of different abundances in a sample.

Model of Protein Abundances. A general knowledge of the frequency, f, of the various protein abundances, x, in a proteome is lacking. However, the very thorough examination by Ghaemmaghami et al.[3] of the amount of various proteins in the *S. cerevisiae* proteome indicate a range of abundances of 10^6 and a symmetric Gaussian-like distribution. Although the general validity of this result is unknown, we here chose to employ a Gaussian distribution—i.e.

$$f = 2^{-1/2}\cdot\pi^{-1/2}\cdot\sigma^{-1}\cdot\exp(-(x' - 3)^2/2\cdot\sigma^2) \tag{1}$$

with $x' = {}^{10}\log x$ and the standard deviation $\sigma = 1$. Hence, f is centered at 10^3 and with 3 standard deviations located at 1 and 10^6.

Part (i) Simulation of Limited MS-Dynamic Range—No Peptide Separation. The hypothetical mass spectrometric dynamic range, D, was varied from 10^2 to 10^6. The simulations were performed by assuming different numbers of proteins in a sample and by randomly selecting the respective protein abundance, x_i, from the Gaussian distribution f (eq 1). Mass

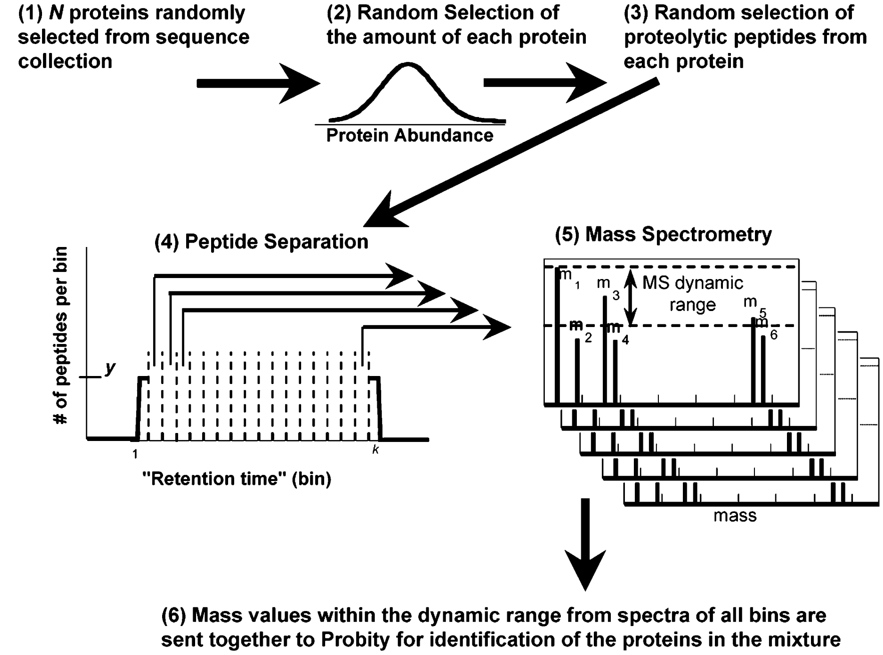
(1) *N* proteins randomly selected from sequence collection
(2) Random Selection of the amount of each protein
Protein Abundance
(3) Random selection of proteolytic peptides from each protein
(4) Peptide Separation
# of peptides per bin
y
1
k
"Retention time" (bin)
(5) Mass Spectrometry
m1
m3
MS dynamic range
m5
m2
m4
m6
mass
(6) Mass values within the dynamic range from spectra of all bins are sent together to Probity for identification of the proteins in the mixture

Figure 1. Schematic description of Simulation B, part (ii): (1) *N* proteins are selected randomly from a sequence collection. (2) The amount of each protein is selected randomly from a Gaussian distribution (see text for details). (3) Proteolytic peptides are selected randomly. The amount of each individual proteolytic peptide is assumed to be the same as the abundance of the protein from which it originates. (4) The peptides are *separated randomly* into *k* different bins. The number of peptides per bin is assumed to be a constant value *y*. (5) Only a fraction of the *y* peptides in each bin can be detected by MS due to the limitation of the MS dynamic range. (6) Detectable mass values from *all* bins are submitted *together* to Probity for searching the sequence collection and to identify the proteins in the mixture.

spectrometric signal was assumed to be detectable within the chosen mass spectrometric dynamic range measured relative to the highest randomly chosen protein abundance, $x_{\max}$, in the sample—i.e., the protein, i, is detectable if $x_i > x_{\max} - D$. The simulations assumed that a mixture is composed of whole proteins or of one or several peptides originating from one protein mixed with one or several peptides originating from another protein etc. Hence, the simulation corresponds with experimental situations that do not involve peptide separation: e.g., the proteolytic peptides are mass analyzed directly from a digest solution or from an extract of unresolved proteins digested in a gel. We investigated 100−300 samples for each degree of mixture complexity (10−100 proteins in each sample). Differences in ionization probabilities between different peptides and limitations due to lack of detection sensitivity were not considered.

Part (ii) Simulation of Limited MS-Dynamic Range—Utilizing Peptide Separation. If a proteome analysis experiment utilizes *separation* of the proteolytic *peptides*, then the prospects of mixture analysis must be examined in a somewhat different manner than that for no peptide separation (*i*) described above. A RP−HPLC separation of peptides originating from a protein mixture results in *randomization* with respect to an individual protein. If we assume that *y* peptides are *eluting simultaneously* from a column loaded with *eN* peptides (*N* is the number of proteins, and *e* is the average number of proteolytic peptides per protein), the *y* peptides represent a sample of the distribution of protein abundance. A fraction of *y* is not detectable by MS due to the limitations in dynamic range. The loss of a peptide signal due to a limited dynamic range is primarily a loss of *sequence coverage* of a protein and *not* a loss of the whole protein. The *protein* is lost when the sequence coverage becomes too low to allow significant identification. To model this phenomenon we performed a simulation that combines features of the Simulation A with limitations introduced by the protein abundance distribution, *f*, and limitations of the MS dynamic range. Synthetic data composed of proteolytic peptide masses corresponding to different randomly selected sequence coverages of *N* randomly selected proteins were generated as in Simulation A. The abundance of each protein was randomly selected from the Gaussian distribution (eq 1). The peptide separation was assumed to result in a uniform distribution of the number of peptides eluting simultaneously. The retention time scale was modeled as a series of *k* different bins. We assumed $k \approx 100$ (mimicking, e.g., a 33.3 min gradient elution with an average peptide elution duration of 20 s) and hence $y \approx eN/k$. Each bin was "filled" by selecting *y* mass values in random order from the list of proteolytic peptide masses generated as in Simulation A. The protein abundance corresponding with each peptide in each respective bin was analyzed and mass spectrometric signal was assumed to be detectable from each peptide having an abundance within the chosen mass spectrometric dynamic range measured relative to the peptide of highest abundance in the bin. Mass values corresponding with abundances outside the dynamic range were *discarded.* Protein identification was performed as in Simulation A—i.e., mass values from *all* bins were employed together in the sequence collection search, but using only the mass values that were considered as detectable within the dynamic range in each bin. A schematic representation of the simulation utilizing separation of peptides is displayed in Figure 1. Differences in ionization probabilities between different peptides and limitations due to lack of detection sensitivity were not considered.

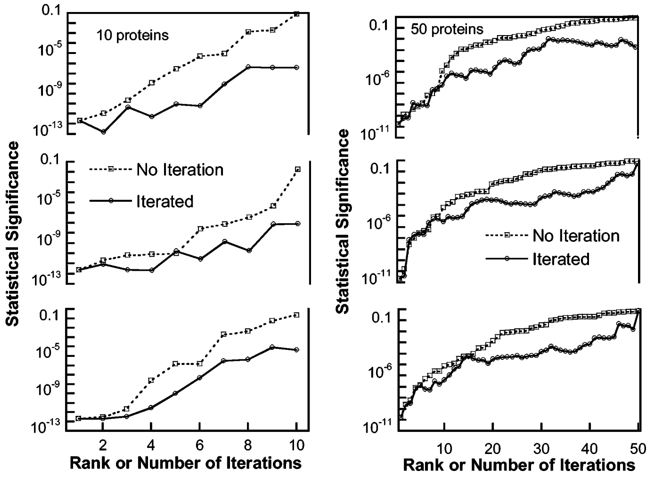

Figure 2. Consistent improvement of statistical significance (the risk that the result is false) when using an iterative search procedure compared with a single-step search and monitoring the ranking list. The mass accuracy was 0.03 Da and the sequence coverages of the proteins in the synthetic data were 30−60% with a minimum of 3 peptides per protein. *Left:* Three examples with 10 proteins in the mixtures. *Right:* Three examples with 50 proteins in the mixtures.

Scripts written in Perl were employed for all the simulations, which were performed on a Dell (2.66 GHz Pentium IV) personal computer.

Results

Simulation A. The prospects of identifying the components of complex mixtures were investigated as a function of different search strategies (iterative/noniterative) and as a function of various experimental conditions (mass accuracy, sequence coverage, and mixture complexity).

Effect of Iteration. Series of protein identifications from mixtures of various degrees of complexities were investigated with and without iteration. Figure 2 elucidates the consistent improvement of the statistical significance (the risk that the result is false) when using the iterative approach—leading to a substantial enhancement of the number of proteins in the mixtures identified at a desired[23] significance level of 0.001 (Figure 3) The reason underlying the improvement is simply that the iteration causes a successive reduction of the mixture complexity, which reduces the problem of random matching between mass values in the data and mass values corresponding to peptides of the proteins in the sequence collection. Random mass matching is the sole cause of false results[16,17] and Probity inherently accounts for the reduced risk of obtaining a false result as the number of mass values in the data is reduced in each iteration step.[20]

Influence of Mass Accuracy. The influence of the mass accuracy on the prospects of identifying proteins in complex mixtures was investigated. Figure 4 displays an example of the different performance for the two levels of mass accuracy (±0.03 Da and ±0.003 Da) investigated for a mixture of 50 proteins using the iterative Probity-based procedure. It is evident that a good mass accuracy improves the prospects for significant identification of the components of complex mixtures. This observation is due to the fact that the problem of random matching is reduced when the accuracy is improved.[17,24] In the example of 50 proteins in a mixture displayed in Figure 4, an accuracy of ±0.03 Da led to 38 results significant at the 0.1% level, whereas the corresponding number for the accuracy of ±0.003 Da was 49.

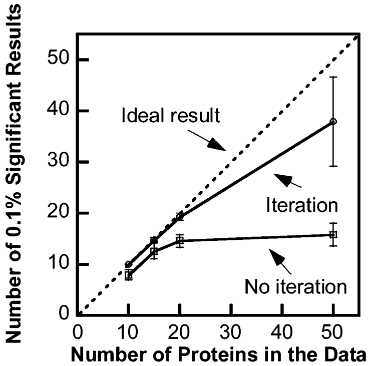

Figure 3. Number of proteins identified at the 0.1% significance level (results with at most 0.1% risk of being false) as a function of the number of proteins in the mixture when using iteration and no iteration, respectively. The dashed line indicates the ideal result that all proteins in all mixtures yield highly significant (0.001) results. Each data-point represents the average number of 0.1% significant results from 10 different data-sets and the bars represent the standard deviation.

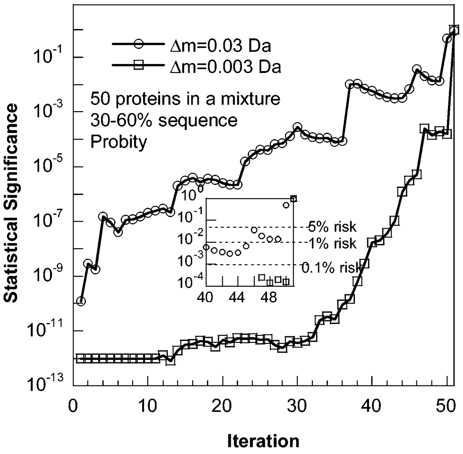

Figure 4. Statistical significance as a function of the number of iterations (identifications) for two different levels of the mass accuracy of data and in the identification process. The inset displays the statistical significance for the 40th to the 50th iteration (identification).

Influence of Sequence Coverage. The impact of the sequence coverage on the quality of protein identification in complex mixtures was studied. The result displayed in Figure 5 elucidates that the sequence coverage is of key importance for the significance values when identifying proteins in complex mixtures.

Besides from the sequence coverage itself, we also investigated the influence of the minimum number of peptides from an individual protein present in the data. Figure 6 displays a comparison of identification results obtained from data with 30−60% sequence coverage, but filtered to guarantee a minimum of 2 or 3 peptides per protein, respectively. As shown in Figure 6, a minimum of 2 peptides is a considerably more difficult case than is that of the data having a minimum of 3 peptides per protein. We hypothesize that the practical use of data containing only 2 mass values from an individual protein is very limited unless additional means are employed to derive further information on, e.g., peptide p*I*, amino acid composition or sequence.[24]

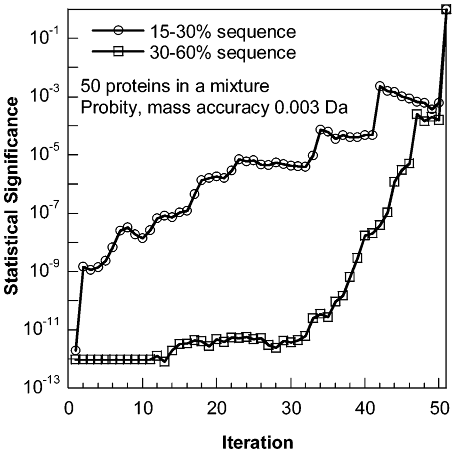

Figure 5. Statistical significance as a function of the number of iterative identifications for a mixture of peptides from 50 different proteins. The influence of the sequence coverage on the level of significance is apparent.

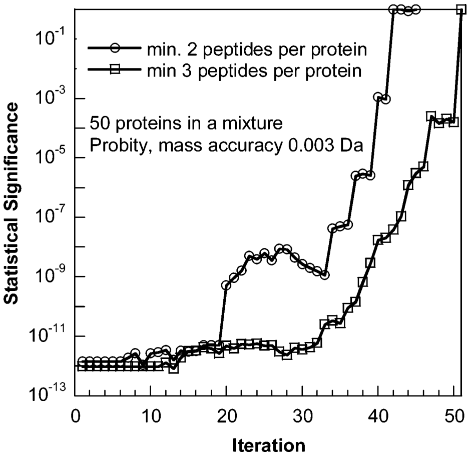

Figure 6. Statistical significance of proteins identified from a mixture of 50 proteins with 30−60% sequence coverage for data having a minimum of 2 and 3 peptides per protein, respectively.

Extreme Complexity. There is an astonishing potential for protein identification in very complex mixtures if the data are characterized by reasonable sequence coverage and good mass accuracy, and if the data are submitted to the Probity algorithm iteratively. Figure 7 displays the result from identifying proteins in a mixture composed of tryptic peptides from 300 proteins (with a total of 3100 monoisotopic mass values). In this example, 280 proteins were correctly identified at the significance level of 0.1%, 284 proteins at the 1% level and 293 at the 5% significance level.

Simulation B. Detection Probability in Mixtures. (i) No Peptide Separation. Although the real distribution of the abundances of various proteins in various proteomes is unknown, our hypothetical Gaussian distribution (eq 1) allowed us to monitor the principles of the information losses due to a limited dynamic range of mass spectrometers. It is seen in Figure 8a,b that the loss of low abundance proteins is more pronounced when (1) the more complex is the mixture, and (2) the lower is the dynamic range of the MS-analysis. It is seen in Figure 8a that for a dynamic range of 10^2 the number of proteins detectable in a mixture is poor already for a mixture of 10 proteins, whereas for a dynamic range of 10^5 most proteins in complex mixtures of 50 proteins are detectable.

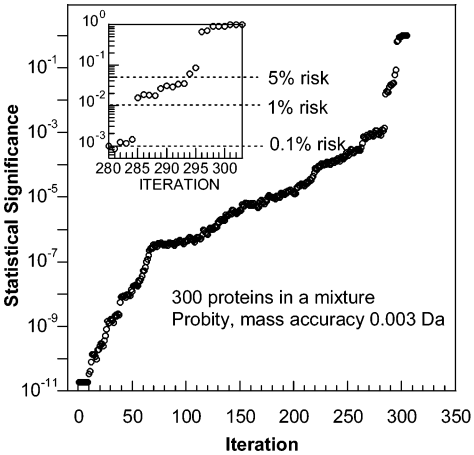

Figure 7. Statistical significance of proteins identified in a mixture of 300 proteins. The inset displays a magnified portion of the graph for the 280th to the 300th protein identified (iteration).

(ii) Protein Identification Based on MS of Separated Peptides. The randomization with respect to individual proteins that result from separation of proteolytic peptides by RP−HPLC influences the ability to detect peptides. The primary effect of a limited mass spectrometric dynamic range as MS analysis is performed subsequently to RP−HPLC is a loss of sequence coverage. Figure 8c displays an example of how the sequence coverage distribution for a protein mixture is altered by limitations in the dynamic range for the detection of ions of separated proteolytic peptides eluting in an order randomized with respect to an individual protein. For a proteolytic peptide mixture from 50 proteins with 30−60% sequence coverage, a dynamic range of 10^5 yielded no losses of sequence coverage, whereas a dynamic range of 10^2 yielded complete loss of 5 proteins and the loss of sequence coverage was often pronounced (Figure 8c). Figure 8d indicates the prospects for detection *and* proteolytic peptide mass fingerprint based identification of complex mixtures as peptides have been separated (randomized) prior to a dynamic range limited MS-analysis. For a mixture of 50 proteins with a proteolytic peptide mass accuracy of 0.003 Da and using the Probity algorithm in the iterative manner, all results were significant at the 0.1% level for a dynamic range of 10^5, whereas for a dynamic range of 10^2 36, 32, and 29 results were significant at the 5%, 1%, and 0.1% significance levels, respectively. The corresponding numbers for a simulation under the same conditions but including 300 proteins were 293 (5%), 284 (1%), and 279 (0.1%) for a dynamic range of 10^5 and 224 (5%), 213 (1%), and 188 (0.1%) for a dynamic range of 10^2 (data not shown).

Discussion

Dynamic Range and the Detection of Ions. The protein identification results of Simulation A displayed here for several examples of rather complex protein mixtures (Figures 4−7) elucidate that MS data are extremely rich in information, which allows successful protein identification provided that the algorithm employed has a good tolerance against random matching and that an iterative search procedure is used. A key question is whether data generated by experiments would result in the same level of success? There is no principle difference between the mass values generated in silico (with the use of a realistic mass error generator) and mass values derived from measurements. The central experimental problems are to *detect*

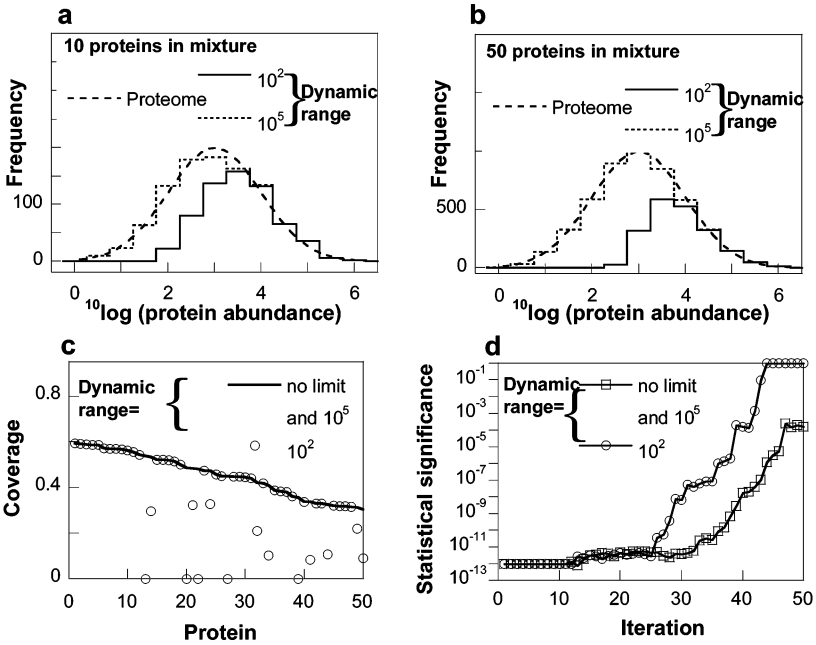

Figure 8. Simulation B. Top panel: Simulation results for the distribution of detectable proteins when assuming *no* peptide separation, a hypothetical Gaussian distribution of protein abundance in the proteome, and a mass spectrometric (MS) dynamic range of 10^2 or 10^5 for protein mixtures composed of **(a)** 10 proteins and **(b)** 50 proteins. **Bottom Panel.** Simulation results when utilizing *separation of peptides* originating from a protein mixture having a Gaussian distribution of the protein abundance. **(c)** An example of the distribution of the sequence coverage for each protein in a mixture of 50 proteins (no limit) together with the resulting sequence coverage when the peptides have been separated and subjected to MS with dynamic range limitation. A dynamic range of 10^5 yielded no losses of sequence coverage, whereas for a dynamic range of 10^2 several mass values were not detectable by MS resulting in pronounced losses of sequence coverage for some proteins. **(d)** The statistical significance of the proteins identified using iterative Probity for the same example of proteolytic peptides originating from a mixture of 50 proteins as displayed in (c). All mass values detectable within the MS dynamic range were submitted together to the first step of the sequence collection search ($\Delta m = \pm 0.003$ Da).

the peptide ions in the mass spectrometer and to accurately *derive* a mass from each *m*/*z* value measured. There are two fundamental obstacles to overcome in order to detect ions and accurately determine a mass based on a measured *m*/z-value. A highly complex sample can yield *unresolved* peaks in a mass spectrum. Once the probability for overlapping MS-signals from different ionic species is significant the accuracy is expected to become worse, and, if the overlap is observed and the peak is eliminated from the data, the sequence coverage is reduced for the proteins corresponding with the peptides with overlapping peaks. Given the theoretical distribution of mass values for tryptic peptides this problem is expected to be more pronounced for low mass values, where the abundance of peptides is high.[17,24] The problem of resolving power was not considered in our simulations. The other fundamental problem with mixtures in real samples is the dynamic range of mass spectrometers. As mentioned, the common region of dynamic range of mass spectrometers is in sharp contrast with the hypothesized ranges of levels of expressed genes. It is assumed that the range of expression levels of genes can be of the order of 10^{10}.[5] Mass spectrometers often display a dynamic range of as little as 10^2. The loss of information due to limitations in the dynamic range is difficult to estimate precisely without accurate knowledge about the actual gene expression level distribution of the biological system analyzed. The simulation results presented here concerning the influence of the dynamic range are therefore hypothetical in their nature. It is however obvious from the results displayed in Figure 8 that a dynamic range of 10^5 of a mass spectrometer is considerably more attractive for mixture analyses than is 10^2. The dynamic range of MS is limited either by the ion source or by the mass analyzer employed in the mass spectrometer. The two common means of ionization are matrix assisted laser desorption ionization (MALDI) and electrospray ionization (ESI). If we, for example, consider a MALDI ion source a *single* peptide species can be detected at concentration levels ranging at least from 0.001 μM to about 100 μM. Considering the mutual peptide influence on ionization sometimes observed in mixture analysis the practical dynamic range for mixtures could be lower. The limitations induced by mass analyzers are sometimes technical in their nature, e.g., the use of a low number of bits when converting an analogue detector signal to a digital number stored in the data acquisition computer of a time-of-flight mass analyzer, or the use of suboptimal ion filtering in ion inlets of ion trap mass analyzers. These limitations are possible to overcome by appropriate engineering efforts.[25,26] Hence, future experimental design for mixture analysis based on ESI or MALDI *ion sources* would presumably have to estimate a practical limit of the dynamic range of 10^5 or lower.

In proteome analysis experiments, separation methods are typically employed at the protein or the peptide level or at both levels. As indicated by the results presented here, the need for separation could be more critically due to the matter of handling the dynamic range rather than to the handling of a large amount of information acquired in parallel.

Improved Performance by MS/MS? The *identification* results presented in this paper are based entirely on the concept of identifying proteins using accurate measurements of proteolytic peptide masses. Many state-of-the-art proteome analysis experiments utilize partial sequence information obtained by isolation and fragmentation of each peptide ion in the mass spectrometer followed by *m*/*z*-measurements of the resulting fragment ions (MS/MS). These experiments are often performed on-line with RP−HPLC separation. Hence, in such experiments dynamic range considerations resemble the situation modeled in Simulation B, part (ii). It is expected that a lower number of detectable peptides can be sufficient for protein identification when utilizing fragment mass information.[24,27] However, the number of fragments as well as the accuracy of mass measurements can in some instances be too poor to confidently trace the gene expressed based on the detection of one peptide only. Comparisons of MS/MS versus MS only for the identification of protein mixtures need to be explored further. In general, MS/MS-analysis is clearly advantageous for the identification of post-translational modifications (PTM). Dynamic range issues also influence studies of PTM, which become meaningful only if a large fraction of a protein sequence is accessible for analysis or if PTM of the type studied can be enriched.[28,29]

Protein Mixtures from Higher Organisms. Sequence collections of higher organisms typically contain more sequences than the sequence collection of *S. cerevisiae* employed in the protein identifications simulated here. As the proteomes are more complex, experimental difficulties concerning separation and consequently the problems associated with the dynamic range increase. In addition, the difficulty of obtaining statistically significant results increases with the number of sequences in the sequence collection searched.[16,21] We have demonstrated recently that the dependence of the statistical significance on the size of the sequence collection can be computed accurately.[21] For example, the result displayed in Figure 7 can be converted to what is expected from *H. sapiens* by employing a simple formula. Instead of 284 proteins identified at the 1% significance level, the corresponding value for human is estimated to be about 280. Hence, we hypothesize that the scale-up of the complexity of the organism is likely to cause more severe problems due to the limitations of the dynamic range than it does from the informatic viewpoint.

Future Development and Direct Application. The predictive power of simulations of protein identification in protein mixtures could be improved significantly by an increased knowledge about the actual distribution of protein abundances in proteomes. Hence, experimental efforts to actually measure such distributions are of key importance for experimental design in the field of proteome analysis in general and for the challenging task of tracing low abundance biomarker proteins in particular. Once accurate distributions of protein abundances are established simulations of probabilities of detection can be combined with protein identification simulations in order to derive meaningful detailed information on the expected degree of successful identification under various experimental conditions. Although the present work is based on somewhat idealized model assumptions, we believe that the simulation results presented here can serve as a first step of a guideline for further optimization of comprehensive proteome analysis. The benefits of iteration and the performance of the probity algorithm demonstrated here should become useful for the identification of proteins unresolved by the electrophoresis typically employed in the state of the art proteome analysis methods.

Conclusions

We have demonstrated that the Probity algorithm employed in an iterative procedure can correctly identify >95% of proteins at a desired significance level in mixtures composed of hundreds of yeast proteins under realistic mass spectrometric experimental constraints. We have demonstrated that the great informatic potential for mass spectrometric identification of protein mixtures is held back by limitations of MS dynamic range. We have shown that simulations that account for a limited dynamic range could become a useful tool for predicting the degree of success in proteomics experiments, e.g., when developing experimental protocols that ensure the detection of low abundance proteins.

Acknowledgment. The Swedish Research Council provided funding to J.E.

References

(1) Jensen, O. N.; Podtelejnikov, A. V.; Mann, M. *Anal. Chem.* **1997**, *69*, 4741.
(2) Ramstrom, M.; Palmblad, M.; Markides, K. E.; Hakansson, P.; Bergquist J. *Proteomics* **2003**, *3*, 184.
(3) Ghaemmaghami, S.; Huh, W. K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* **2003**, *425*, 737.
(4) Tyers, M.; Mann, M. *Nature* **2003**, *422*, 193.
(5) Anderson, N. L.; Anderson, N. G. *Mol. Cell Proteomics* **2002**, *1*, 845.
(6) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 14440.
(7) Mortz, E.; Vorm, O.; Mann, M.; Roepstorff, P. *Biol. Mass Spectrom.* **1994**, *23*, 249.
(8) Yates, J. R. 3rd. *Trends Genet.* **2000**, *16*, 5.
(9) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys Res. Commun.* **1993**, *195*, 58.
(10) Beavis, R.; Fenyo, D. *Proteomics: A Trends Guide* **2000**, 22.
(11) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. *Science* **1996**, *274*, 546.
(12) Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M., and et al. *Science* **1995**, *269*, 496.
(13) Adams, M. D.; Celniker, S. E.; Holt, R. A., et al. *Science* **2000**, *287*, 2185.
(14) The *C. elegans* sequencing consortium. *Science* **1998**, *282*, 2012.
(15) The human genome project. *Nature* **2001**, 409.
(16) Eriksson, J.; Chait, B. T.; Fenyo, D. *Anal. Chem.* **2000**, *72*, 999.
(17) Eriksson, J.; Fenyo, D. *Proteomics* **2002**, *2*, 262.
(18) Zhang, W.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 2482.
(19) Lokhov, P. G.; Tikhonova, O. V.; Moshkovskii, S. A.; Goufman, E. I.; Serebriakova, M. V.; Maksimov, B. I.; Toropyguine, I. Y.; Zgoda, V. G.; Govorun, V. M.; Archakov, A. I. *Proteomics* **2004**, *4*, 633.
(20) Eriksson, J.; Fenyo, D. *J. Proteome Res.* **2004**, *3*, 32.
(21) Eriksson, J.; Fenyo, D. *J. Proteome Res.* **2004**, *3*, 979.
(22) Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768.
(23) Ossipova, E.; Nord, L.; Kenne, L.; Eriksson, J. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2053.
(24) Fenyo, D.; Qin, J.; Chait, B. T. *Electrophoresis* **1998**, *19*, 998.
(25) Beavis, R. C.; Chait, B. T. *Rapid Commun. Mass Spectrom.* **1989**, *3*, 233.
(26) Krutchinsky, A. N.; Kalkum, M.; Chait, B. T. *Anal. Chem.* **2001**, *73*, 5066.
(27) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390.
(28) Oda, Y.; Nagasu, T.; Chait, B. T. *Nat. Biotechnol.* **2001**, *19*, 379.
(29) Zhou, H.; Watts, J. D.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 375.

PR049816F