research articles Journal of proteome research

Detection and Identification of Protein Isoforms Using Cluster Analysis of MALDI−MS Mass Spectra

Rikard Alm,[†] Peter Johansson,[‡] Karin Hjernø,[||] Cecilia Emanuelsson,[†] Markus Ringnér,[‡] and Jari Häkkinen*[,‡,§]

Department of Biochemistry, Lund University, Sweden, Complex Systems Division and Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University, Sweden, and Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark

Received October 20, 2005

We describe an approach to screen large sets of MALDI−MS mass spectra for protein isoforms separated on two-dimensional electrophoresis gels. Mass spectra are matched against each other by utilizing extracted peak mass lists and hierarchical clustering. The output is presented as dendrograms in which protein isoforms cluster together. Clustering could be applied to mass spectra from different sample sets, dates, and instruments, revealed similarities between mass spectra, and was a useful tool to highlight peptide peaks of interest for further investigation. Shared peak masses in a cluster could be identified and were used to create novel peak mass lists suitable for protein identification using peptide mass fingerprinting. Complex mass spectra consisting of more than one protein were deconvoluted using information from other mass spectra in the same cluster. The number of peptide peaks shared between mass spectra in a cluster was typically found to be larger than the number of peaks that matched to calculated peak masses in databases, thus modified peaks are probably among the shared peptides. Clustering increased the number of peaks associated with a given protein.

Keywords: hierarchical clustering • proteomics • mass spectra • protein identification • isoforms

Introduction

In proteomics, cellular function can be investigated on the protein level by observations of several hundreds or thousands of proteins simultaneously.[1,2] Mass spectrometry is a central tool in these experiments and is used to identify proteins and investigate their actual physical state, presence of covalent modifications, and their up- or down-regulation in response to various treatments or cellular states. A proteomic investigation usually involves sample preparation, protein separation, and mass spectrometric data acquisition and analysis. The last step, data analysis, is crucial for the interpretation of data. Novel ways to analyze acquired data are therefore important for conclusive results. With time, such analysis methods can be included into software for automated analysis, and become an important part of the actual capacity of a proteomics setup.

Protein isoforms can be detected as multiple spots in two-dimensional electrophoresis (2-DE), or by mass spectrometry (MS) as detection of modified peptide sequences. There are several explanations for protein isoforms: multiple gene copies (allelic variation), alternative splicing, truncation or degradation products, or the presence of various post-translational modifications (PTMs).[3−5] Experimental detection of PTMs and the assignment of correct isoforms of a protein are expected to be one of the major experimental challenges in proteomics.[6,7]

Clearly, there is a need for methods to rapidly screen for protein isoforms in any proteomics data set using mass spectrometry (MS) instrumentation. We describe an approach to facilitate mass spectrometric data analysis by matching the peptide mass fingerprints within a data set against each other to obtain clusters of mass spectra. The clusters represent similar proteins and isoforms that can be subjected to closer investigation.

Our approach is based on clustering lists of peak masses extracted from mass spectra and is available through a web interface named SPECLUST (http://bioinfo.thep.lu.se/speclust.html; Johansson P et al., work in progress). We compare peak lists by measuring a distance between each pair of peak lists. Many distance measures have been suggested (see e.g., ref 8), and most of them are histogram-based, i.e., binning data and counting how many bins contain peaks from both lists. Arbitrary bin boundaries may lead to sensitive to small measurement errors. To avoid this potential problem, we defined a measure where we calculate a match score between each pair of peaks based on their difference in masses. These scores are then used to align the peak lists and to calculate a

* To whom correspondence should be addressed. Jari Häkkinen, Department of Theoretical Physics, Lund University, Sölvegatan 14a, SE−223 62 Lund, Sweden. Phone: +46 (0)46-222 9347. Fax: +46 (0)46-222 9686. E-mail: jari@thep.lu.se.

[†] Department of Biochemistry, Lund University.

[‡] Complex Systems Division, Department of Theoretical Physics, Lund University.

[§] Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University.

[||] Biochemistry and Molecular Biology, University of Southern Denmark.

10.1021/pr050354v CCC: $33.50

Published on Web 03/16/2006

distance between the lists. Similar methods have been used in tandem mass spectrometry (MS/MS) database search algorithms.[9−11] Finally, we cluster the peak lists using these distances as a starting point. The result of the approach is presented as a dendrogram in which protein isoforms cluster together.

Clustering of mass spectra has been suggested for many other applications in proteomics. Schmidt et al. also clustered peak lists extracted from mass spectra of spots on 2-DE gels.[12] They used clustering to purify peak lists by removing peaks stemming from neighboring spots, thereby improving protein identification. Müller et al. used data from molecular scanners[13] to cluster peptide masses according to the similarity of the spatial distributions of their signal intensities.[14] This clustering improves identification of weakly expressed proteins. Tibshirani et al. used clustering of peaks across many mass spectra in a method to classify samples from patients according to disease status from protein MS data.[15] Beer et al. used clustering of LC−MS/MS spectra to reduce the large amounts of data generated in this process to a manageable size.[16] Monigatti and Berndt proposed a method to cluster MS spectra to generate consensus mass spectra from a large mass spectrum database, with the aim to achieve more unambiguous identification and decreased numbers of false positives in high throughput screening.[17]

We applied our clustering approach to two data sets. First, we used a data set consisting of 62 mass spectra derived from nine *Arabidopsis thaliana* proteins that appeared in multiple spots on two-dimensional electrophoresis (2-DE) gels. The peak lists, derived from mass spectra acquired for isoforms and replicate samples for each of the nine proteins, clustered together perfectly. Second, we applied the cluster analysis to another data set with unknown numbers of protein isoforms present. Several clusters suggested protein isoforms that were verified by protein identification based on MS/MS. We examined clusters further by identifying peaks being shared between mass spectra within a cluster. These shared peaks were submitted to a peptide mass fingerprint (PMF) search and yielded improved identification compared to using peaks from individual mass spectra. The clustering aided identification by increasing the number of peaks associated with a given protein, and recognized shared peaks not matched to calculated peak masses in databases. These peaks represent possible isoforms and post-translational modifications (PTMs), amenable for closer investigation.

Material and Methods

Mass Spectra Derived from Nine *Arabidopsis thaliana* Proteins. Published data from a study by Schubert et al. of the proteome of the chloroplast lumen of *Arabidopsis thaliana*[18] was used to compile a data set by selecting proteins represented by at least three spots on a single 2-DE gel. Mass spectra derived from in total five replicate gels run at different dates, and with MS data acquisition performed at different dates and on different instruments were used. The data set contained mass spectra from nine different, identified proteins: O22609; DEGP1_ARATH DegP-like protease, O82660; HC136_ARATH PSII stability factor HCF136, P82281; TL29_ARATH Ascorbate peroxidase, Q39249; Q39249_ARATH Violaxanthin deepoxidase, Q41932; PSBQ2_ARATH OEC 16 kDa subunit, Q42029; PSBP1_ARATH OEC 23 kDa subunit, Q9FYG5; Q9FYG5_ARATH Glyoxalase-like, Q9S841; PSBO2_ARATH OEC 33 kDa subunit, and Q9SW33; TL1Y_ARATH Lumenal 17.9 kDa protein, where the nine proteins were identified in 3 to 12 mass spectra each. In total, this set consisted of 62 mass spectra, each originating from a different spot.

Mass Spectra Derived from *Fragaria ananassa* Proteins. Data were generated by 2-DE of a protein extract from strawberry, *Fragaria ananassa*, in order to display differential expression of proteins,[19] especially the isoforms of the strawberry allergen.[20] Spots were selected for mass spectrometric analysis on the basis that they showed differential expression between two different types of strawberry.[19] This selection yielded a data set consisting of 88 mass spectra, each originating from a different spot. The MALDI−MS mass spectra were acquired in data-dependent mode on a Waters Micromass MALDI micro MX Mass Spectrometer (Waters, Manchester, UK) followed by automated protein identification by searching the PMFs against the NCBI nr database, limited to green plant (*Viridiplantae*), with either the search engine Mascot,[21] or the software PIUMS.[22] Although the MS spectra were of good quality, this PMF only yielded a 10% success rate in protein identification due to the lack of strawberry sequence information in NCBI nr.

After cluster analysis, a new sample set was prepared for a final round of mass spectrometric investigation to improve the protein identification rate. Manual MS and MS/MS data acquisition was performed using an Applied Biosystems 4700 Proteomics Analyzer with time-of-flight/time-of-flight (TOF/TOF) optics (Applied Biosystems, Darmstadt, Germany).

Peak Extraction and Preprocessing for Cluster Analysis. Peaks were extracted from the raw files with the software PIUMS.[22] This software allows automated recalibration of mass spectra based on recognized trypsin and keratin peaks from an automatically generated filter, and removal of trypsins, keratins, and other contaminant peaks.[23]

We used PIUMS with default parameter setting with the following exceptions: (i) Bin width 0.8, (ii) peaks with masses below 750 or above 4000 Da were removed, and (iii) the manually adjustable minimal number of hits parameter in PIUMS was set differently for the two data sets. Peaks found in many spectra are considered contaminants and the minimal number of hits parameter is used to remove peaks common to at least the number of spectra set by this parameter. For the validation set from *Arabidopsis thaliana*, the minimal number of hits was set to 19. For the strawberry data set with an unknown but presumably lower number of similar proteins, the minimal number of hits was set to 12.

Clustering. For clustering, we used the agglomerative hierarchical clustering method first suggested by Ward.[24] The method starts by assigning each peak list to its own cluster and calculating a distance between each pair of peak lists. The closest pair is found and merged to a new cluster. Distances between the new cluster and each of the old clusters are calculated. The search for closest pair, merging the pair, and calculation of new distances are repeated until there is one single cluster. We clustered using average linkage as implemented in the clustering package provided by de Hoon et al.[25] In average linkage, the distance between two clusters is calculated as the average of the distances from each peak list in one cluster to each peak list in the other cluster. The application of hierarchical clustering to high-dimensional biological data has been reviewed by Quackenbush.[26]

We calculated distances between peak lists by first calculating a similarity score for each pair. The similarity score in turn was assessed by comparing how well individual peaks in the first list matched peaks in the second list. Therefore, we also

defined a peak match score between two peaks taken from different peak lists.

Having two peaks, from different peak lists, with measured masses m and m' and measurement uncertainty σ, we wanted a peak match score that reflects the probability that the two peaks originate from the same peptide. We assumed measurement errors to be Gaussian and defined the peak match score to be the probability to get a mass difference equal to or larger than $|m\text{-}m'|$ given that the difference is only due to measurement errors. This assumption gives the peak match score $s = P(\Delta > |m\text{-}m'|) = 1 - \text{erf}(|m\text{-}m'|/2\sigma)$, which is zero for measurements infinitely apart and unity for measurements being identical. In contrast to a binary score, where peak matches are given a score 1 when the mass difference is within a predefined window and zero otherwise, this score allows for smoother inclusion of measurement errors since it gives a continuous score value between zero and unity. In all analysis presented in this paper, we used σ equal to 1 Da.

To calculate a similarity score, S, between two peak lists, we added up all contributions from individual peak matches, Σs_{ij}, where s_{ij} is the peak match score between peak i in the first list and peak j in the second list. Recalling that we study mass spectra puts some restrictions on the summation. Each peak can only be matched to one other peak, and peak order (by mass) cannot be permuted (i.e., if peaks m and M from the first list are matched to peaks m' and M' from the second list, respectively, the only permissible relationships of their masses are $m < M$, $m' < M'$, or $m > M$, $m' > M'$). There are many possible combinations of peak matches (alignments) fulfilling these two conditions, and we chose the one that maximizes the sum Σs_{ij}. To find this maximum value, we used the Needleman-Wunsch algorithm,[27] commonly used in global sequence alignment.

The distance measure we used in clustering, $d = 1 - S/\min(N,N')$, is based on the similarity score, S, and the sizes of the two peak lists, N and N'. Intuitively, this distance measure corresponds to the fraction of peaks in the smaller peak list having no match to the larger list. Consequently, the distance is zero when each peak in the smaller list has a perfect match to a peak in the larger peak list. Because we use a distance measure that depends on a fraction of peaks, it is relatively insensitive to the number of peaks in spectra. This distance measure is the starting point in clustering the peak lists and building a dendrogram.

Extraction of Shared Peaks. To further investigate clusters, we examined pairs of peak lists from a cluster and identified shared peaks. A peak was considered shared between two spectra, if it was matched in the alignment of the spectra with a peak match score larger than 0.7 corresponding to a 0.5 Da mass difference.

Results and Discussion

Validation of the Clustering Method Using Mass Spectra from Nine *Arabidopsis* Proteins. To assess whether protein isoforms could be detected using hierarchical clustering of mass spectra, we performed clustering of peak lists from 62 *Arabidopsis thaliana* mass spectra, each originating from a different spot. This resulted in a dendrogram in which the nine proteins formed nine distinct clusters (Figure 1A). It is evident from Figure 1A, that every isoform and every replicate sample from all nine proteins cluster together perfectly. This result indicates that the clustering method is robust and is working although the mass spectra were obtained from different gels, different mass spectrometers, and at different dates. We tried different values for σ (0.1 to 10 Da), and found the clustering to be very robust. The sequence coverage in these mass spectra was typically around 25%, which is routinely obtained in automated MALDI−MS. This sequence coverage was obviously sufficient to yield clear clustering.

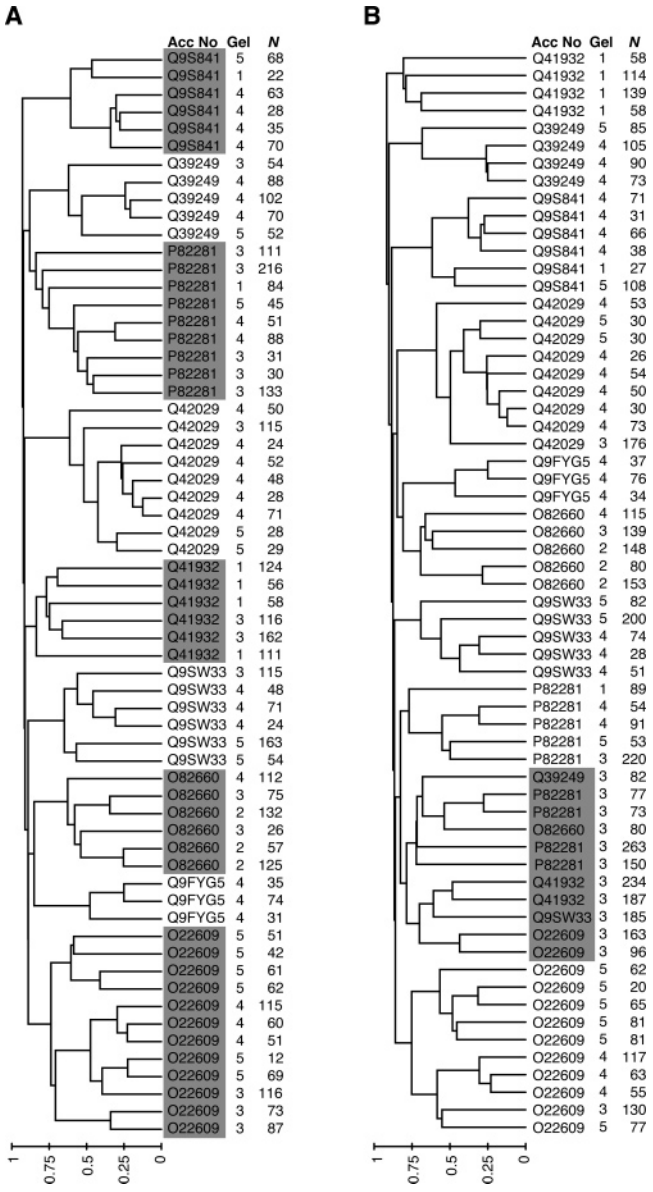

Figure 1. Hierarchical clustering of 62 peak lists from nine *Arabidopsis* proteins. Mass spectra were derived by MALDI−MS, and peak extraction and processing (calibration and filtering) were performed in PIUMS. For each mass spectrum the accession number, 2-DE gel identification number, and the size of peak list (N) used in the clustering are shown. (A) Filtered and calibrated peak lists. (B) Nonprocessed peak lists. The proteins listed are O22609; DEGP1_ARATH DegP-like protease, O82660; HC136_ARATH PSII stability factor HCF136, P82281; TL29_ARATH Ascorbate peroxidase, Q39249; Q39249_ARATH Violaxanthin deepoxidase, Q41932; PSBQ2_ARATH OEC 16 kDa subunit, Q42029; PSBP1_ARATH OEC 23 kDa subunit, Q9FYG5; Q9FYG5_ARATH Glyoxalase-like, Q9S841; PSBO2_ARATH OEC 33 kDa subunit, and Q9SW33; TL1Y_ARATH Lumenal 17.9 kDa protein. The scale below the dendrogram indicates the distance used in the clustering (see Materials and Methods).

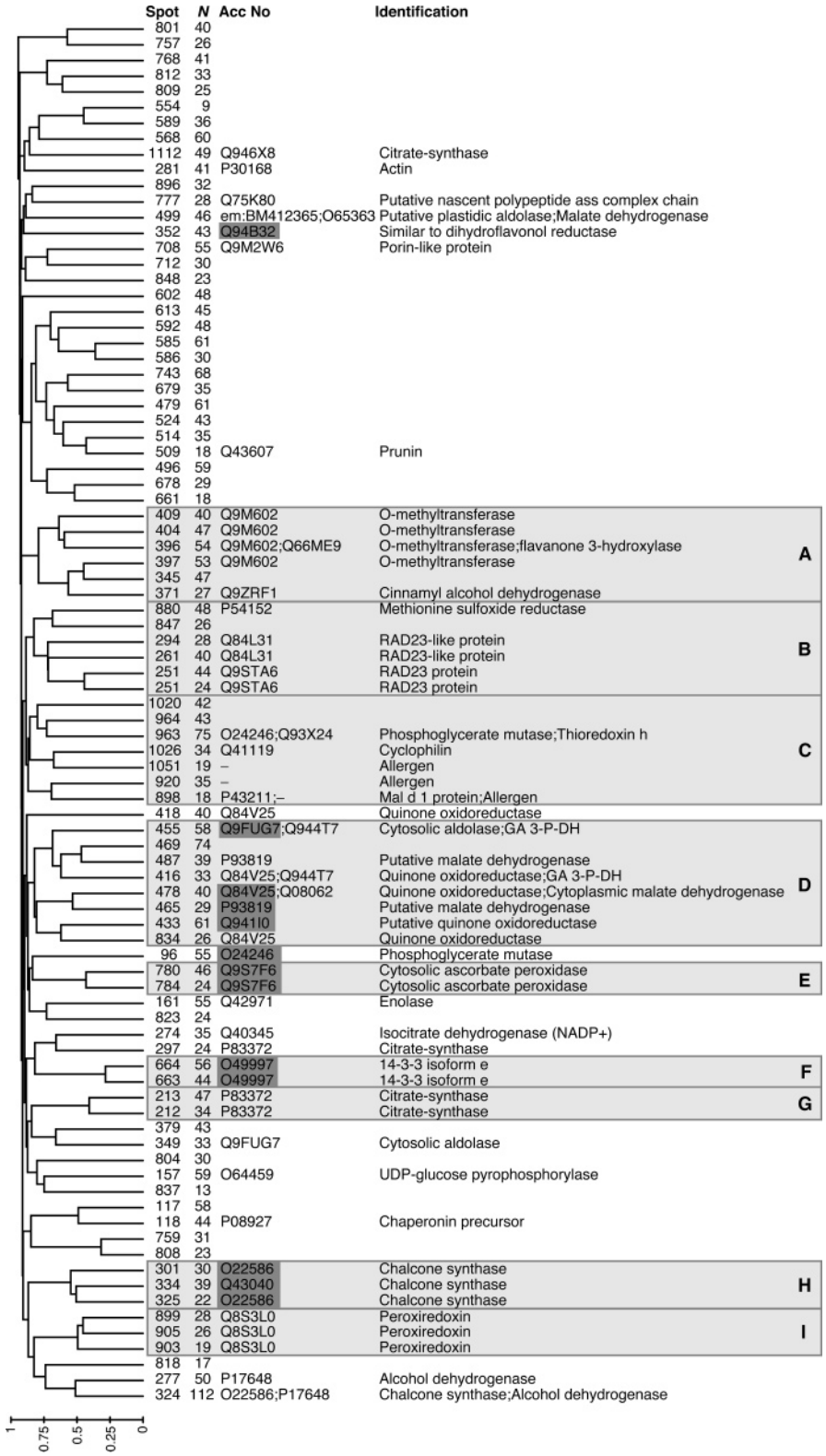

Figure 2. Hierarchical clustering of mass spectra from *Fragaria ananassa* proteins with an unknown number of isoforms. Mass spectra were derived by MALDI-MS and peak extraction and processing were performed using PIUMS. For each mass spectrum, the spot number and the number of peaks (*N*) are stated. In cases where proteins were successfully identified also an accession number and protein name are shown. Proteins having accession numbers on dark gray backgrounds could be identified initially by automated PMF and database searching with Mascot and PIUMS. The remaining identified mass spectra were identified in a second round of cluster affiliation, MS/MS and database searching, in combination with manual interpretation. Nine clusters that were selected for discussion are labeled by A to I. The scale below the dendrogram indicates the distance used in the clustering (see Materials and Methods).

Our preprocessing of mass spectra that only removes contaminant peaks and recalibrates the mass spectra was important to yield clear clustering of peak lists. Without this preprocessing, the clear clustering of the nine proteins (Figure

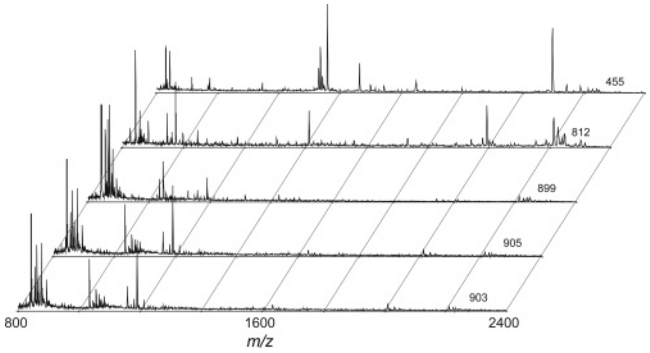

Figure 3. Example of five MALDI mass spectra from *Fragaria ananassa* 2-DE spots. The mass range shown is from 800 to 2400 and spot numbers are indicated on the right. Spectra from spots 903, 905, and 899 are similar and cluster together (cluster I in Figure 2), whereas spectra from spots 812 and 455 are different and did not end up in cluster I.

1A) cannot be observed (Figure 1B). In contrast, a cluster appears (Figure 1B, gray shaded) that is comprised of peak lists from mass spectra derived from six of the nine different proteins. All mass spectra in this cluster were derived from one gel, indicating that this gel was heavily contaminated. Removal of contaminants from the mass spectra is also well-known to be important prior to PMF searches not to obscure matching of peak lists to calculated masses in databases.[23]

Clustering of Mass Spectra from Strawberry Proteins with An Unknown Number of Isoforms. Hierarchical clustering was also applied to a data set that was not explicitly compiled to contain isoforms. This realistic data set contains 88 mass spectra from spots selected after separation on 2-DE because they showed differential expression between different types of strawberry.[19] This data set contained an unknown number of protein isoforms, and many spots for which the protein identity was not known.

Since only few of the strawberry proteins could be identified in the first round of automated PMF, only 13 of 88 spots were assigned identifications (Figure 2). This is typical for proteins from species with nonsequenced genomes, like strawberry. For such genomes, protein identification based solely on MS data is dependent on identification by sequence homology. However, the clustering analysis was used to decide how to proceed with manually performed protein identification by combined MS and MS/MS. Application of clustering to the 88 mass spectra, acquired from 88 spots, yielded the dendrogram presented in Figure 2. Several clusters of mass spectra suggesting possible isoforms could be discerned, and nine clusters that are marked by boxes and labeled by A to I were selected for further investigation. Identifications were finally obtained for 51 of 88 spots[19] with names and accession numbers as stated in Figure 2.

The mass spectra from the spots in cluster I, together with two other spectra, are shown in Figure 3 illustrating the similarity of spectra clustering together.

On the basis of the first round of PMF, some of the selected clusters were found to reinforce our conclusion from analyzing the Arabidopsis data that known isoforms cluster together. These clusters were E (Cytosolic ascorbate peroxidase), F (14-3-3 isoform e), and H (Chalcone synthase). The final identification confirmed that other clusters were also dominated by isoforms, including A (O-methyltransferase), B (RAD23 proteins), G (Citrate-synthase), and I (Peroxiredoxin). On the other hand, two spots (1112 and 297) outside of cluster G were identified as citrate-synthases, and one spot (324) outside of cluster H was found to contain chalcone synthase as well as another protein.

It is well known that isoforms due to phosphorylation can be seen as a string of pearls on a 2-DE gel. Other modifications can also be detected visually on a gel as long as the p*I*-shift or mass-change is small. This is true for the 14-3-3 isoform e (cluster F) for which the spots are physically very close to each other on the gel (Figure 4). On the other hand, cytosolic ascorbate peroxidases (cluster E) are physically far apart. Hence, clustering can reveal isoforms that are not so easily suspected to be isoforms by inspection of the gel.

In Figure 2, most mass spectra cluster together with other mass spectra. An exception is the unidentified protein derived from spot number 602 that is an outlier, suggesting that it has no isoforms in the analyzed data set.

Some mass spectra, which do contain protein isoforms according to the stated names and accession numbers, do not cluster together as nicely as those mentioned above. Many of these less perfect clusters contain spots from the upper region of the 2-DE gel (Figure 4), where the spot density was higher and many spots overlap with each other. Mass spectra derived from spots in this region are likely to contain more than one protein. One example is chalcone synthase, for which three mass spectra cluster together (cluster H), but a mass spectrum from a spot containing both chalcone synthase and alcohol dehydrogenase clusters with alcohol dehydrogenase. Another example is cluster D that contains four proteins intertwined in a complex manner. The fact that clustering is obscured when the mass spectra contain peaks from more than one protein resembles the situation for PMF searches, where peak mass lists from more than one protein usually give poor results in protein identification.

Improved Protein Identification by Clustering. To improve PMF-based protein identification, we utilized the cluster analysis in the following way. By first subjecting the peak lists to cluster analysis, peak masses shared within a cluster were identified. These shared peak masses are candidates to belong to the protein in question. Hence, a novel peak list comprised of these shared peak masses can then be used in a second PMF search. If successful, such a PMF-based identification can potentially be extrapolated to other protein members of the cluster. Moreover, the shared peak masses, hypothesized to belong to the protein in question, can also be used for a second round of data acquisition by MS/MS to improve and/or verify the protein identification. This approach, outlined in Figure 5, was utilized in two ways. First, the approach was used to improve protein identification, either within clusters with only one protein per mass spectrum in cases where protein identification initially was not successful (e.g., the allergen, O-methyltransferase, RAD23 protein, citrate-synthase, and peroxiredoxin clusters), or by deconvoluting clusters with more than one protein per mass spectrum (e.g., cluster D). Second, the approach was used to identify modified peaks.

Improved Protein Identification within Clusters. As one example of how clustering can assist in protein identification, we describe how our approach was applied to the allergen protein. After the first round of automated MS data acquisition and PMF protein identification, one spot (898) matched with an insignificant score to a homologous allergen from apple (Mal d 1 protein). Subsequently, clustering was used to search for mass spectra similar to the mass spectrum from spot 898. The

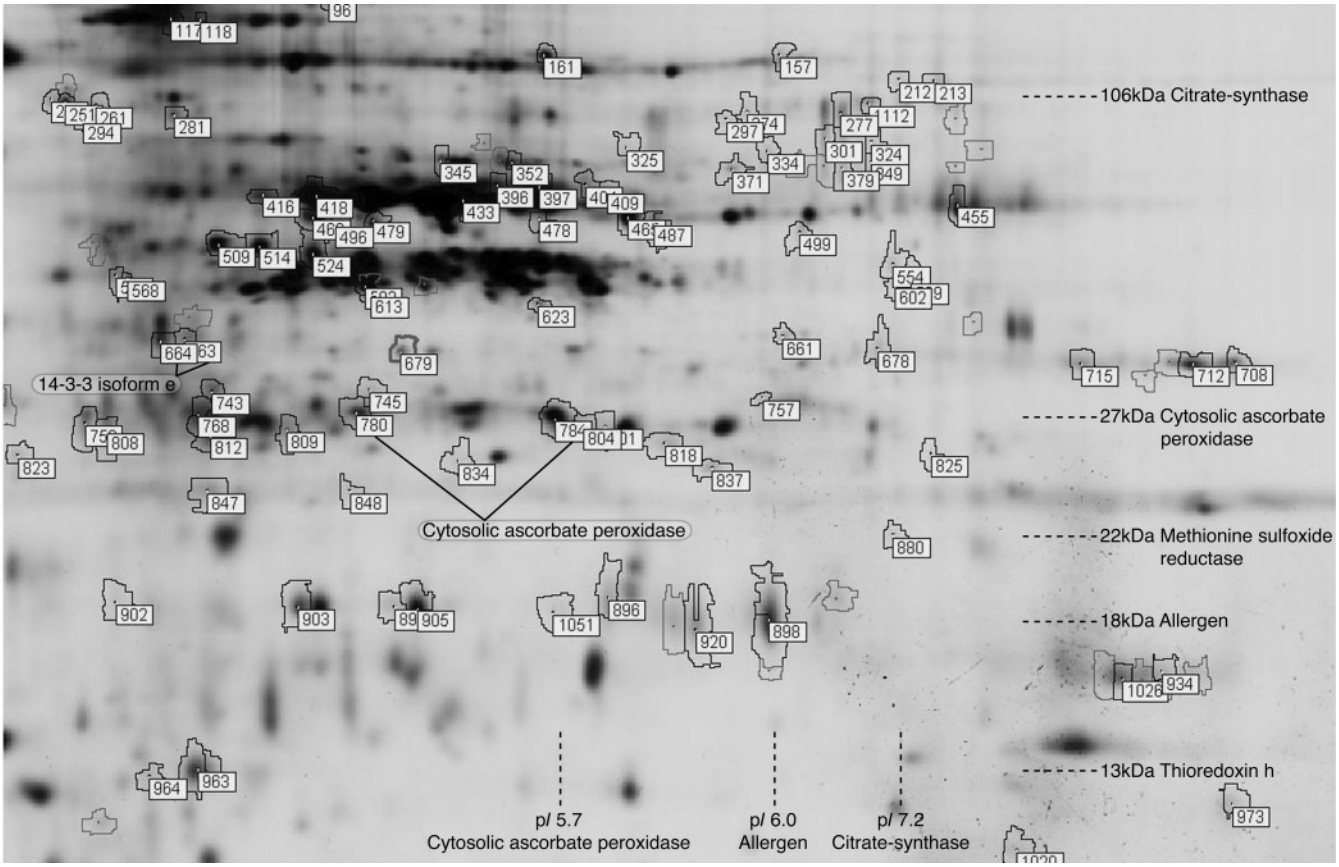

Figure 4. Gel image of *Fragaria ananassa* proteins showing spots selected for mass spectrometric analysis. Spots encircled and annotated with a spot number were selected for mass spectroscopic analysis. Theoretical mass and isoelectric point (p*I*) values for some of the identified proteins are indicated with dotted lines. Note that clustering analysis can identify spots physically close to (e.g., 14-3-3 isoform e), as well as spots far apart from (e.g., ascorbate peroxidase), each other.

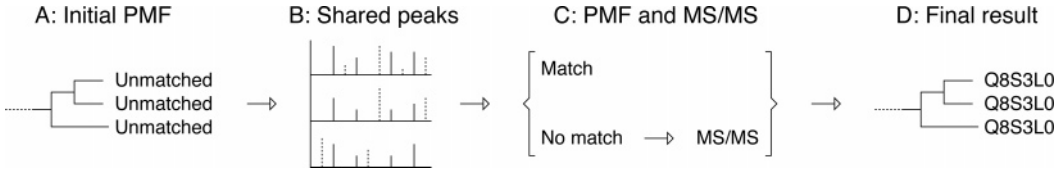

Figure 5. Strategy for improved protein identification by cluster analysis and identification of shared peak masses. (A) After initial MS data acquisition, data is used for initial PMF and clustering. (B) Peak masses shared by mass spectra in a cluster are identified. (C) The shared peaks are used for a second round of PMF identification, and still unidentified mass spectra are run through MS/MS for further investigation, (D) eventually leading to successful identification for mass spectra.

mass spectrum from spot 898 clustered together with spectra from six other spots in cluster C (Figure 2) and the spots in this cluster were subjected to a closer investigation as outlined in Figure 5. By manual protein identification, we found that spots 920 and 1051 also contained the allergen.[19] Thus clustering can be used to suggest which spots in a large data set should be investigated more closely for the presence of a particular protein.

Spectra containing more than one protein may cluster together with spectra containing any of these proteins, depending on how clusters are merged. An example of this behavior is cluster D (Figure 2). We prefer the clustering to perform like this because it helps to disentangle spots that contain multiple proteins. Our choice of merging clusters (average linkage) is sensitive to such multiple protein spots without introducing poor clustering of independent single protein spots.

A spectrum with more than one protein is in general a problem in PMF searches, but with our approach it was possible to deconvolute such a spectrum using other spectra from the same cluster and improve PMF searches. The automated identification in Figure 2 was a combination of PIUMS and Mascot. We reexamined the spectra with Mascot to measure how much the PMF search could be improved by using shared peaks. As an example, we used mass spectra from the following: (i) spot 478 that contained both quinone oxidoreductase and malate dehydrogenase, (ii) spot 465 that contained malate dehydrogenase, and (iii) spot 433 that contained quinone oxidoreductase (see Figure 2).

Mascot gave the following results: spot 478, malate dehydrogenase (score 82) and no hit for quinone oxidoreductase, spot 465, malate dehydrogenase (no significant score), and spot 433, quinone oxidoreductase (score 87). Thereafter, peak masses shared between mass spectra were identified and novel

Table 1. Clustering Can Identify Shared Peaks Which Do Not Match the Theoretical Sequence

allergen spot	total[a]	matched[b]	shared and matched[c]	shared but not matched[d]
898	32	5	5	8
919	8	2	2	1
920	36	6	5	7
1051	21	4	3	3

[a] The number of peaks in the mass spectrum after filtering (keratins, trypsins, and contaminants removed). [b] The number of peaks in the mass spectrum that matched the theoretical sequence. [c] The number of peaks in the spectrum that matched the theoretical sequence and found to be shared between at least two of the four allergen spots. [d] The number of peaks in the spectrum that did not match the theoretical sequence but found to be shared between at least two of the four allergen spots.

peak lists were created and subjected to Mascot as follows. The novel peak list shared between spots 478 and 465, corresponding to malate dehydrogenase, gave a higher score (88) than the two original peak lists from spots 478 and 465. The novel peak list shared between spots 433 and 478, corresponding to quinone oxidoreductase, gave a score of 91. Quinone oxidoreductase was not at all detected with the original peak list from spot 478. Hence, in total two new identifications would have been found using shared peak lists, based solely on Mascot. Thus, clustering can assist in identification of spectra with more than one protein per mass spectrum, and improve PMF searches. The protein identifications stated in Figures 2 and 4 were confirmed by manually performed MS/MS.[19] When peptide masses are selected for MS/MS from a spectrum containing several proteins, it is advantageous if selected peptides belong to the same protein. For such a spectrum, our approach to find shared peaks can assist in the selection of peptides from one protein. Each protein in the spectrum can thereby selectively be subjected to MS/MS.

Using Clustering for Identification of Modified Peaks. To investigate if clustering can be used not only to detect but also to benefit the characterization of isoforms, the allergen spectra were investigated with an additional round of MS data acquisition. New mass spectra were obtained for the three allergen spots as well as for a fourth spot (spot 919, not shown in Figure 4) also containing the allergen.[19] These four mass spectra were investigated for shared peaks. Most of the peak masses that could be assigned to calculated peak masses in databases were found to be shared by at least two mass spectra (Table 1). However, only approximately half of the shared peak masses could be assigned to calculated peak masses. This finding suggests that several of the peaks shared between the mass spectra were modified peptide peaks because contaminant peaks were removed in preprocessing. Thus, clustering can assist in the selection of tentatively modified peptides for further characterization by MS/MS analysis. For example, the peptide with mass 1516.7 Da, shared by spots 898, 919, and 920, was confirmed to be a modified peak. This peptide is a modified variant of the peptide CAEILEGDGGPGTIK.[19]

Clustering revealed one modified peptide and focused the investigation to the four spots containing the allergen. To further characterize isoforms a protocol was developed in ref 19 with a double-derivatization to obtain a complete y-ion series in MS/MS, which yielded sequence information and confirmed that for example the peptide LVSAPHGGTLLK (1192.7 Da) is present in two more isoforms. These isoforms were contained within the same spot, 920, and within the same MS spectrum.

Conclusions

Cluster analysis after MS data acquisition can be used to screen for possible protein isoforms in large proteomic studies. Clustering is not dependent on database content and can be applied to mass spectra from different sample sets, dates, and instruments provided that mass spectra are calibrated and filtered. Peaks that are shared within a cluster, likely to represent the protein in question, can be further characterized with MS/MS. Also, shared peaks that do not match theoretical masses may represent modified peaks that can be identified. This approach is well suited for MALDI-TOF/TOF, where it is possible to first scan in MS mode and, following the cluster analysis, to perform MS/MS on shared peaks. To fully investigate differences between protein isoforms high sequence coverage is needed. Nevertheless, we have presented a clustering approach that benefits the characterization of isoforms even for the sequence coverage routinely obtained in MALDI−MS data acquisition.

Acknowledgment. We thank Wolfgang Schröder and Thomas Kieselbach at Umeå University, Sweden for the *Arabidopsis thaliana* mass spectrometric data and Björn Samuelsson for valuable discussions. This work was in part supported by the Swedish Research Council, the Knut and Alice Wallenberg Foundation through the Swegene consortium, and the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS, 2002-0042).

References

(1) Pandey, A.; Mann, M. Proteomics to study genes and genomes. *Nature* **2000**, *405* (6788), 837−846.
(2) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.
(3) Larsen, M. R.; Roepstorff, P. Mass spectrometric identification of proteins and characterization of their post-translational modifications in proteome analysis. *Fresenius J. Anal. Chem.* **2000**, *366* (6−7), 677−690.
(4) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255−261.
(5) Jensen, O. N. Modification-specific proteomics: characterization of posttranslational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **2004**, *8* (1), 33−41.
(6) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **2002**, *27* (2), 74−78.
(7) Appel, R. D.; Bairoch, A. Posttranslational modifications: a challenge for proteomics and bioinformatics. *Proteomics* **2004**, *4* (6), 1525−1526.
(8) Hansen, M. E.; Smedsgaard, J. A new matching algorithm for high-resolution mass spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (8), 1173−1180.
(9) Pevzner, P. A.; Dancik, V.; Tang, C. L. Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.* **2000**, *7* (6), 777−787.
(10) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* **2001**, *11* (2), 290−299.
(11) Potthast, F.; Ocenasek, J.; Rutishauser, D.; Pelikan, M.; Schlapbach, R. Database independent detection of isotopically labeled MS/MS spectrum peptide pairs. *J. Chromatogr. B Analyt Technol. Biomed Life Sci.* **2005**, *817* (2), 225−230.
(12) Schmidt, F.; Schmid, M.; Jungblut, P. R.; Mattow, J.; Facius, A.; Pleissner, K. P. Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J. Am. Soc. Mass Spectrom.* **2003**, *14* (9), 943−956.
(13) Binz, P. A.; Muller, M.; Walther, D.; Bienvenut, W. V.; Gras, R.; Hoogland, C.; Bouchet, G.; Gasteiger, E.; Fabbretti, R.; Gay, S.; Palagi, P.; Wilkins, M. R.; Rouge, V.; Tonella, L.; Paesano, S.; Rossellat, G.; Karmime, A.; Bairoch, A.; Sanchez, J. C.; Appel, R. D.; Hochstrasser, D. F. A molecular scanner to automate proteomic research and to display proteome images. *Anal. Chem.* **1999**, *71* (21), 4981−4988.

(14) Muller, M.; Gras, R.; Appel, R. D.; Bienvenut, W. V.; Hochstrasser, D. F. Visualization and analysis of molecular scanner peptide mass spectra. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (3), 221−231.
(15) Tibshirani, R.; Hastie, T.; Narasimhan, B.; Soltys, S.; Shi, G.; Koong, A.; Le, Q. T. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* **2004**, *20* (17), 3034−3044.
(16) Beer, I.; Barnea, E.; Ziv, T.; Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **2004**, *4* (4), 950−960.
(17) Monigatti, F.; Berndt, P. Algorithm for accurate similarity measurements of peptide mass fingerprints and its application. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (1), 13−21.
(18) Schubert, M.; Petersson, U. A.; Haas, B. J.; Funk, C.; Schroder, W. P.; Kieselbach, T. Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. *J. Biol. Chem.* **2002**, *277* (10), 8354−8365.
(19) Hjerno, K.; Alm, R.; Canback, B.; Matthiesen, R.; Trajkovski, K.; Bjork, L.; Roepstorff, P.; Emanuelsson, C. S. Down-regulation of the strawberry Bet v 1-homologous allergen in concert with the flavonoid biosynthesis pathway in colourless strawberry mutant. *Proteomics* **2005**, *6* (5), 1574−1587.
(20) Karlsson, A. L.; Alm, R.; Ekstrand, B.; Fjelkner-Modig, S.; Schiott, A.; Bengtsson, U.; Bjork, L.; Hjerno, K.; Roepstorff, P.; Emanuelsson, C. S. Bet v 1 homologues in strawberry identified as IgE-binding proteins and presumptive allergens. *Allergy* **2004**, *59* (12), 1277−1284.
(21) 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* **1999**, *20* (18), 3551−3567.
(22) Samuelsson, J.; Dalevi, D.; Levander, F.; Rognvaldsson, T. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics* **2004**, *20* (18), 3628−3635.
(23) Levander, F.; Rognvaldsson, T.; Samuelsson, J.; James, P. Automated methods for improved protein identification by peptide mass fingerprinting. *Proteomics* **2004**, *4* (9), 2594−2601.
(24) Ward, J. H. Hierarchical Grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58* (301), 236−244.
(25) de Hoon, M. J.; Imoto, S.; Nolan, J.; Miyano, S. Open source clustering software. *Bioinformatics* **2004**, *20* (9), 1453−1454.
(26) Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2001**, *2* (6), 418−427.
(27) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48* (3), 443−453.

PR050354V