How Much Peptide Sequence Information Is Contained in Ion Trap Tandem Mass Spectra?

Jürgen Cox, Nina C. Hubner, and Matthias Mann
Department for Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Martinsried, Germany

Matching peptide tandem mass spectra to their cognate amino acid sequences in databases is a key step in proteomics. It is usually performed by assigning a score to a spectrum-sequence combination. De novo sequencing or partial de novo sequencing is useful for organisms without sequenced genome or for peptides with unexpected modifications. Here we use a very large, high accuracy proteomic dataset to investigate how much peptide sequence information is present in tandem mass spectra generated in a linear ion trap (LTQ). More than 400,000 identified tandem mass spectra from a single human cancer cell line project were assigned to 26,896 distinct peptide sequences. The average absolute fragment mass accuracy is 0.102 Da. There are on average about four complementary b- and y-ions; both series are equally represented but y ions are 2- to 3-fold more intense up to mass 1000. Half of all spectra contain uninterrupted b- or y-ion series of at least six amino acids and combining b- and y-ion information yields on average seven amino acid sequences. These sequences are almost always unique in the human proteome, even without using any precursor or peptide sequence tag information. Thus, optimal de novo sequencing algorithms should be able to obtain substantial sequence information in at least half of all cases. (J Am Soc Mass Spectrom 2008, 19, 1813–1820) © 2008 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

In "bottom up" MS-based proteomics, proteins are digested to peptides that are then mass measured, isolated in the mass spectrometer, and fragmented, leading to characteristic ion series in the MS/MS spectra [1, 2]. Popular database search programs like Mascot [3], Sequest [4], and many others score these MS/MS spectra against in silico digested peptides whose calculated precursor masses fall into a suitable window around the measured mass, leading to statistically significant identification for a fraction of the mass spectrometric sequencing events [5]. In most cases, the proportion of identifiable peptides is quite low for samples of high protein complexity [6]. Despite recent improvements in identification rates [7, 8], many MS/MS spectra remain unassigned, even though they are of reasonable quality.

The peptide database search approach has the disadvantage that it is blind towards the unexpected: only peptides that result from the digestion of known protein sequences, possibly having a few missed cleavages and a very limited number of standard variable modifications, can be identified in this way. The sequence tag approach [9] is an alternative to the conventional peptide database search that does not suffer from these limitations. Instead of operating in the restricted space of in silico digestions of known protein sequences, one starts by looking for a series of peaks that correspond to consecutive members of a fragment series. Each of the mass differences between two neighboring peaks must be equal to one of the 20 amino acid masses. Much of the specificity of a sequence tag in database searches comes from the mass information encoded in the two flanking masses. In this way, even a tag of two or three amino acids is usually unique in the proteome, especially given the very high precursor mass accuracy possible with modern, high-resolution mass spectrometers. A tag sequence that is part of an in silico peptide but with a wrong parent mass points to a novel and potentially interesting modification or mutation, while a sequence tag that does not match any in silico peptide might be evidence for the expression of a novel and not-predicted protein.

The de novo sequencing problem consists of finding the correct amino acid sequence from the tandem mass spectrum without the help of a database. This problem has fascinated mass spectrometrists for at least three decades and is still not completely solved. Until recently, algorithms have been developed on the basis of restricted datasets. Even the latest efforts in de novo sequencing, i.e., the work of the Pevzner group [10, 11], have not yet taken advantage of recent improvements in performance and in the size of datasets. A fundamental question in the development of partial or complete de novo sequencing algorithms is how much information is present in tandem mass spectra as generated by state of the art proteomics projects. Determination of

Address reprint requests to Dr. M. Mann, Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz, 18, D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de

© 2008 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry.
1044-0305/08/$32.00
doi:10.1016/j.jasms.2008.07.024

Published online August 7, 2008
Received June 27, 2008
Revised July 24, 2008
Accepted July 24, 2008

contiguous peptide sequence generally requires the presence of a fragmentation product from each amino acid bond. Here we set out to determine how often this information is present in tandem mass spectra in very large proteomics projects. We use a large-scale dataset from our group [7], which was analyzed with the MaxQuant set of algorithms [12] and the Mascot search engine. MaxQuant uses the entire mass information present in all survey (precursor) mass spectra and employs sophisticated, peptide length dependent scoring statistics. As a result, the requirements for tandem mass spectra data quality are substantially reduced compared with standard database search, and more than 50% of the fragmentation events are generally assigned in any dataset. We use this very large and high quality dataset to determine the peptide sequence information in linear ion trap fragmentation data. We find that substantial sequence information is embedded in the majority of tandem mass spectra and we extrapolate these results to similar quality tandem mass spectra that are not identified by standard search engines.

Experimental

Methods

Mass spectrometric data. We used the dataset from Cox and Mann [7], which was generated with SILAC labeled HeLa cells after EGF stimulation. Briefly, triplicates were separated into 24 isoelectric focusing fractions, which were analyzed with nanoLC-MS on an LTQ Orbitrap mass spectrometer. MS scans were acquired with high resolution (60,000 at m/z 400), and mass accuracy at the precursor ion level was extremely high, with an average absolute mass deviation of less than 300 ppb. Peptide identification additionally relied on the SILAC information present. Presence of a SILAC pair implies that the peak represents a peptide and not a contaminant molecule. Furthermore, the number of arginines and lysines is known before database search for SILAC pairs. MS/MS spectra were obtained at low resolution in linear ion trap mode and written out as centroid data. These spectra were filtered by retaining only the six most intense peaks in each 100 Th interval [13]. Fragment ions were matched with 0.5 Th mass tolerance. The international protein index (IPI) [14] human version 3.37 served as the sequence database. Processing of the 72 raw files with our MaxQuant software [12] leads to 461,336 identified MS/MS spectra at a 1% false discovery rate (FDR). For the sake of simplicity, we restrict our analysis to the 428,567 of these MS/MS spectra that correspond to completely unmodified peptides, accepting both light and heavy SILAC labeled forms. Together they identify 26,896 distinct peptide sequences with a length of at least six amino acids.

As indicated in the text, in some analyses unfiltered tandem mass spectra were used. The same data were processed but without the filtering of MS/MS spectra in 100 Th bins before submission to the database engine search. In this case, 428,567 MS/MS spectra corresponding to unmodified peptides were identified at 1% FDR, corresponding to 16,853 distinct peptide sequences.

Uniqueness of partial sequences in the human proteome and genome. The partial sequences from the approaches with and without MS/MS filtering were merged. All sub-sequences in identified partial sequences were also considered as partial sequences. For the determination of the multiplicity of partial sequences in the human proteome, we counted their occurrence in the ENSEMBL protein predictions, which attempts to provide a nonredundant set of sequences for the human genome [15]. To avoid underestimation of uniqueness due to the presence of protein isoforms we considered only one protein sequence for each ENSEMBL gene identifier, namely the longest one. The combinations of amino acids with the same molecular weight (leucine and isoleucine) and the same nominal weight (lysine and glutamine) were considered distinct for this calculation. For the statistics over the whole human genome, we downloaded all six frame translations of all human chromosomes from http://www.stateslab.org/data/6frameorfs/index.htm.

Results and Discussion

Fragment Mass Accuracy, Charge Distribution, and Fragment Mass Filtering

We first used our large dataset to determine average fragment mass accuracy. Figure 1a shows a histogram of the difference between measured and calculated fragment ion masses derived from several million matched fragments. The average absolute mass deviation in this histogram is 0.102 Da. The distribution is centered at zero indicating good calibration. All but 5% of fragments are measured within 0.3 Da of the true value and 99% within 0.42 Da. The graph indicates that the commonly used maximum mass deviation [16] of $\pm$0.5 Da for ion trap fragments encompasses close to 100% of measured fragment ions. On the basis of these results, would it be advantageous to set the fragment mass window more tightly? The answer is no, because fragment ion masses—particularly below 1000 Da—are confined to small bands of possible masses, given the restricted atomic composition of amino acids [17]. With the mass accuracy achieved in this dataset, there is virtually no chance that an observed fragment can be matched to a calculated fragment with a different nominal mass. This is of course only true for low-resolution ion trap data. High-resolution MS/MS data, i.e., by measuring the fragments in the Orbitrap, achieves low ppm mass accuracy. This high-resolution data additionally eliminate almost all incorrect fragment matches with the *same* nominal mass.

In Figure 1b the charge distribution of identified peptides is shown. About three-quarters of the tryptic

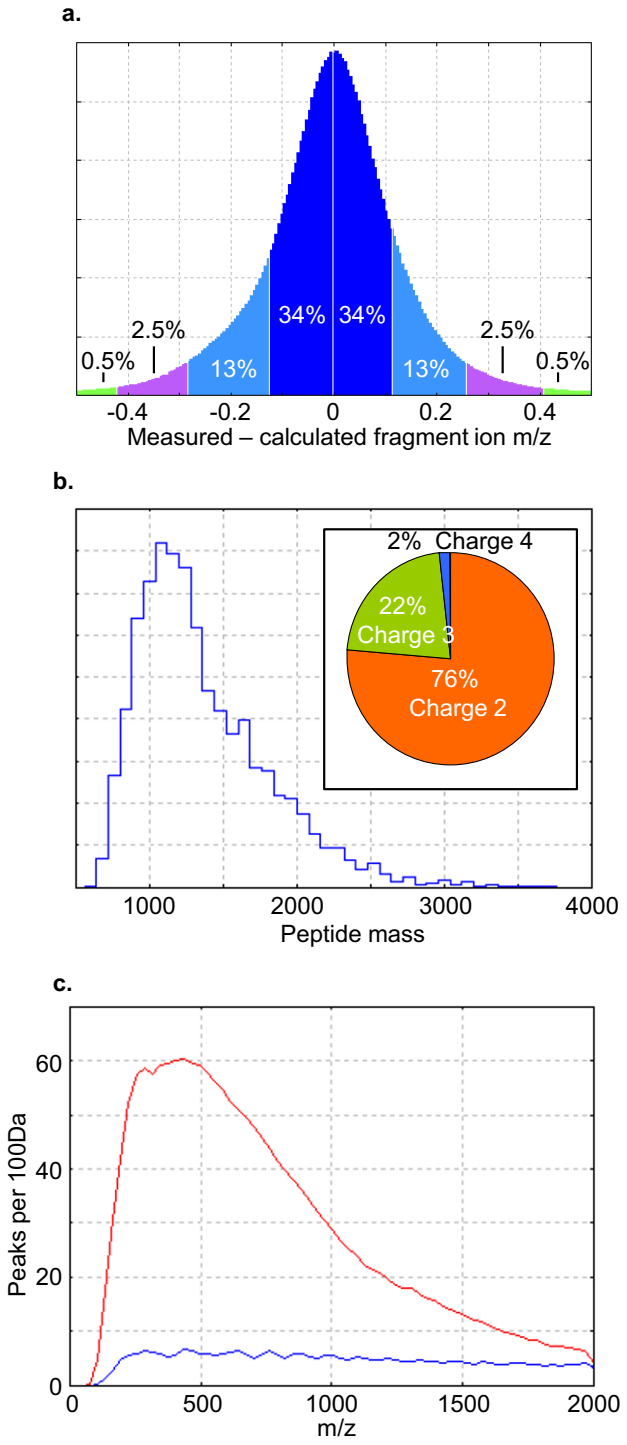

Figure 1. Properties of identified MS/MS spectra. (**a**) Histogram of measured minus calculated m/z of all matched fragment ion peaks. The average absolute mass deviation is 0.102, which fits well with the search tolerance of ±0.5 Da used in the database search. (**b**) Mass and charge distributions of the precursor peptides. (**c**) The average number of peaks per 100 Th mass interval in MS/MS spectra is plotted as a function of m/z for the unfiltered data (red) and for the data filtered to have at most six peaks per 100 Th interval (blue).

peptide precursors are doubly charged, and 22% are triply charged. Only 2% are quadruply or higher charged. (Singly charged ions were excluded from sequencing.)

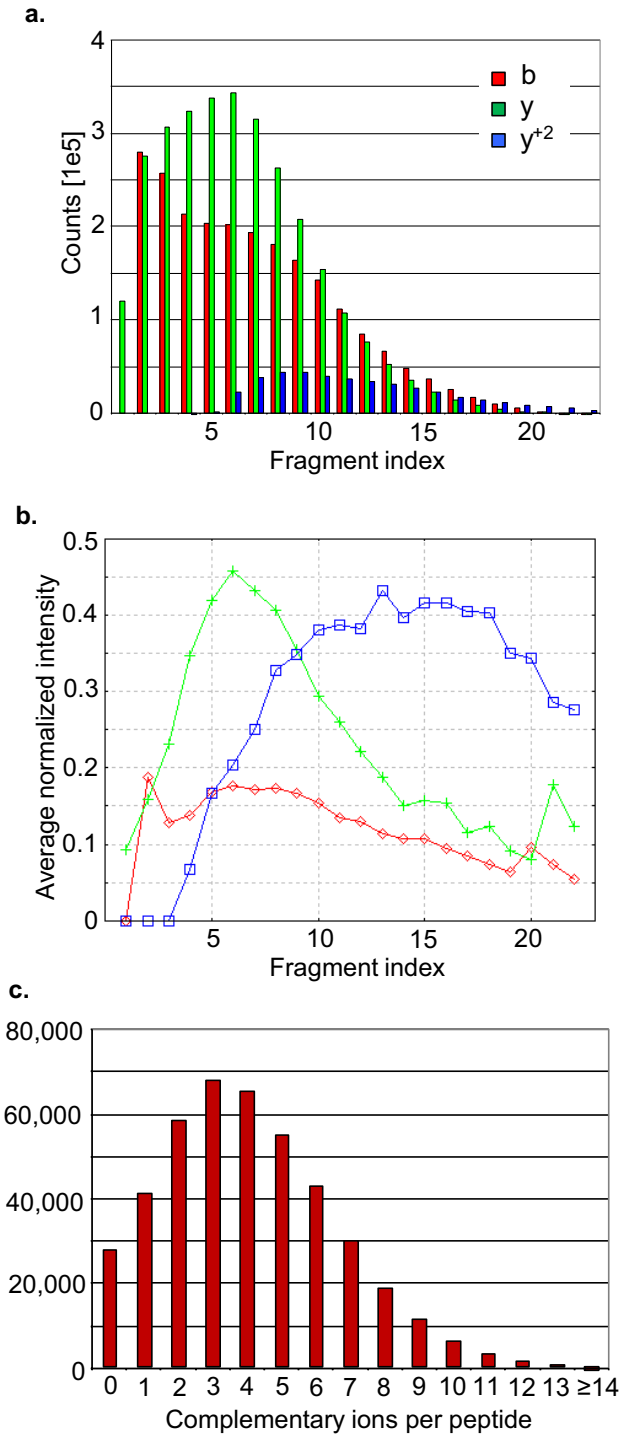

Figure 2. Statistics of fragment ions. (**a**) Total counts of fragment ion peaks matched in the identified MS/MS spectra corresponding to unmodified peptides separately for b-ions (red), y-ions (green), and y^{+2}-ions (blue). (**b**) Average normalized intensity of the fragment ion peaks. The intensities have been normalized in each MS/MS spectrum such that the highest matched peak has intensity one. (**c**) Distribution of the number of complementary ions per peptide. Two ions b_m and y_n are complementary if $m + n$ equals the length of the peptide sequence. On average about four complementary ion pairs are contained in each filtered MS/MS spectrum.

Tandem mass spectra produce peaks at many mass values due to fragmentation events leading to ions different from b- or y-ion types or chemical or electronic noise. These uninformative peaks would make identification difficult, and they are generally of lower abundance than sequence specific fragments. Therefore, spectra are frequently "filtered" so that only the most intense ions remain [18]. Depending on the application we have found a "top 4 filter" per 100 Th or a "top 6 filter" advantageous in separating signal (high intensity) from noise (low intensity). Here we chose a "top 6 filter". As can be seen in the blue curve in Figure 1c, there are usually enough signals in this centroided data that six masses can be obtained in each 100 Th interval, especially at masses below 1000. We also analyzed unfiltered data. Here we obtain signals at up to 60% of all nominal mass values—this number slowly declines to less than 30% at mass 1000 (red curve in Figure 1c).

Properties of Identifiable Tandem Mass Spectra

As mentioned above, our dataset contains low quality MS/MS spectra that are nevertheless unambiguously identified due to the SILAC information and the extremely high precursor mass accuracy. Figure 2a shows a histogram of the three major ion series, b-ions, y-ions, and y^{2+}-ions. The number of fragments for each index (i.e., index of y_6 ion is 6) is a smooth function that for y-ions increases to a maximum at y_8 after which it declines to a few percent of this maximum around y_{15}. Note that this distribution is a convolution of the actual number of fragments of a peptide of length n with the distribution of peptide lengths (Figure 1b). Surprisingly, there are almost as many b-ions as there are y-ions. As expected [19], the b_1 ion is not observed and the b_2 ion is the most frequent one. After this the distribution of b-ions decreases until b_4, where it stays about constant until it catches up to the number of y-ions at b_{11}/y_{11}. The doubly charged y-ion series starts at y_6 and continues relatively flat until y_{18}. However, it is a minor number compared with either b-ions or y-ions.

In Figure 2b, we have plotted the average intensity of each ion index, normalized to the largest peak in the tandem mass spectrum. Here, the difference between b- and y-ion series is much more pronounced. The y-ions are up to three times more intense compared with b-ions, particularly in the "tag region" of y_4 to y_8. This may partly account for the fact that it is often very easy to define a partial sequence of three to four y-ions in any spectrum. (This is even more true in "triple quadrupole type" spectra in which b-ions tend to fragment further.) Unexpectedly, the y^{2+} series, despite its infrequent presence, is as intense as the most abundant y-ions and much more intense than b-ions on average.

One of the major challenges in de novo sequencing is to avoid connecting fragments from different series. Complementary ions (N- and C-terminal ions from the same position in the peptide sequence) can help define the nature of each ion series or at least distinguish one from the other in de novo sequencing algorithms. Furthermore, complementary ion pairs are more likely to be genuine fragment peaks rather than noise or internal fragments and they therefore provide excellent "anchor sites". We counted the number of complementary ion pairs in all spectra and found that on average there are about four (Figure 2c). This indicates that preprocessing of tandem mass spectra for such pairs is a generally useful first step in spectral interpretation. Note, however, that the presence of a pair of complementary fragment masses is not an absolute indication that they are actually a pair: for each given b- or y-ion there is a 6% chance of finding a complementary ion by chance as 6% of all mass values have a signal after "top 6 filtering".

Occurrence of Partial Sequences

In the set of all unmodified peptides (461,336 spectra), we looked for consecutive stretches in the singly and doubly charged y- and in the singly charged b-ion series using the results of the prior database search. An ion fragment series consisting of $i + 1$ peaks determines a (partial) sequence of length i. If we speak of a sequence of length i we mean a series of i amino acids defined by $i + 1$ peaks. Note that the length of sequences present in the spectrum depends on the depth of filtering. All numbers given here are for the top 6 filter per 100 Th. Figure 3a provides an example of a tandem mass spectrum in which partial sequences have been assigned. It contains two sequences of length 3 and 7 in

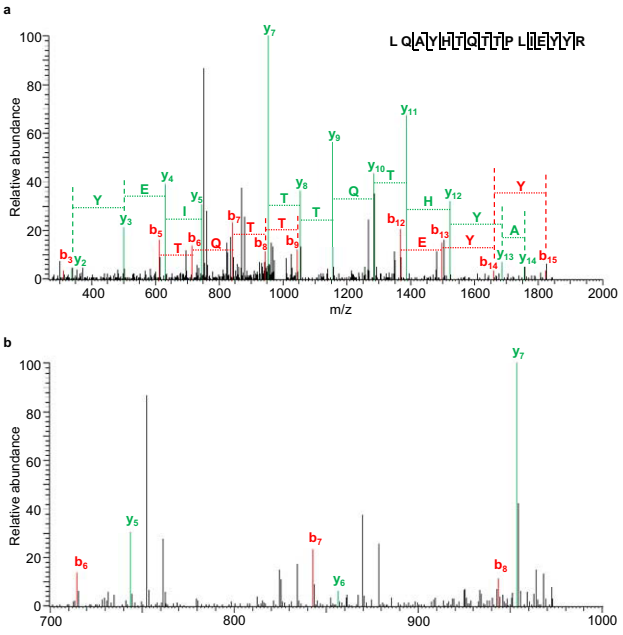

Figure 3. MS/MS spectrum with partial sequences. (**a**) MS/MS spectrum identifying the peptide LQAYHTQTTPLIEYYR with a Mascot score of 86.4. It contains four partial sequences. (**b**) Zoom into the range 700–1000 Th, showing the y_6, which is clearly present but was excluded during "top 6 filtering".

the y-ion series and another two series (lengths 3 and 4) in the b-ion series. The longest sequence is AYHTQTT. Figure 3b indicates the presence of the missing fragment ion between the two y-ion sequences within the low abundance peaks (outside of the top 6 per 100 Th). With this fragment the partial sequence becomes AYHTQTTPLIEY, indicating that the low abundance peaks can be important (see also below).

Starting with $i = 1$, we find 263,142 partial sequences in total, of which 116,254 belong to the b-ion series, 122,708 belong to the y-ion series, and 24,180 belong to the y^{+2} series. To investigate the effects of removing low intensity peaks, we analyzed the unfiltered dataset for partial sequences. This yielded 198,682 sequences for b-ions, 70,336 sequences for y-ions, and 40,271 sequences for y^{2+} ions. These numbers are smaller than the numbers for the top 6 filter because the total number of identified peptides is smaller (by 38%) due to decreased statistics in database matching.

Figure 4a shows the distribution of partial sequences found in filtered MS/MS spectra from the whole dataset consisting of 72 LC-MS runs from HeLa cells. Many sequences are short and the distribution decays nearly exponentially towards longer sequences. Y-ion se-

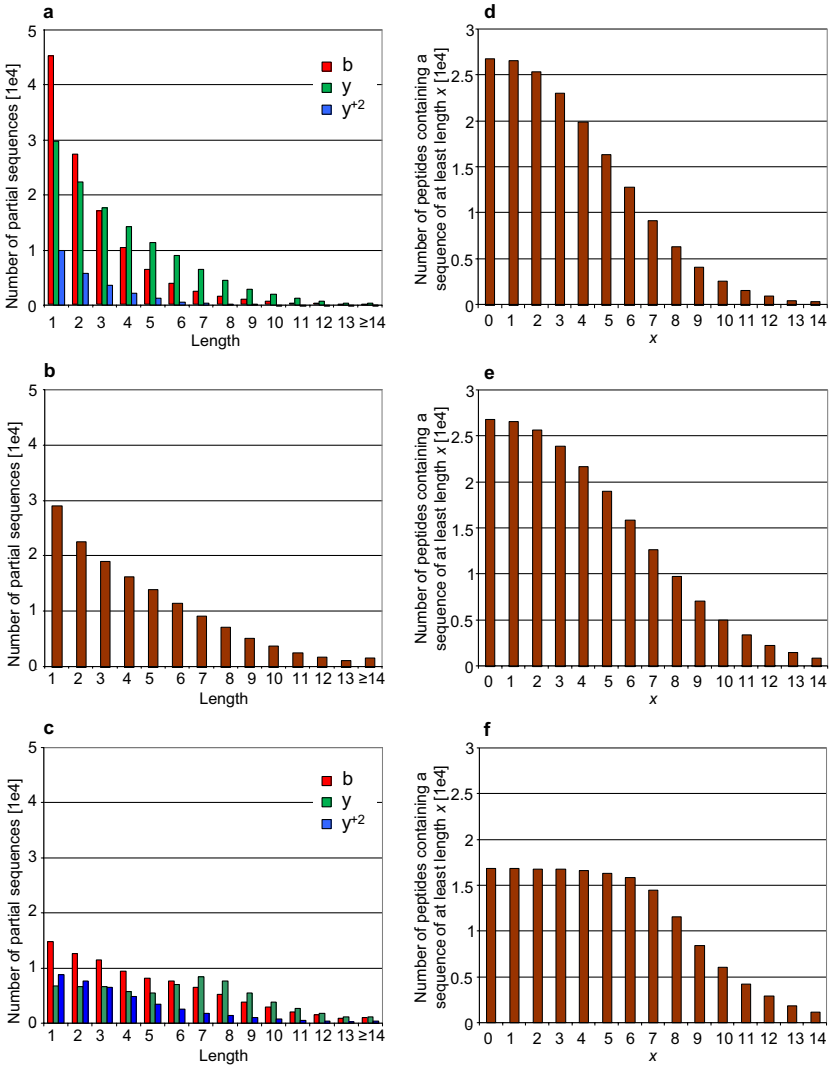

Figure 4. Statistics of sequence partial sequences. (**a**) Length dependent histogram of the 263,142 sequence partial sequences found when filtering top six fragment ion peaks per 100 Th intervals. Sequences from the b-series are in the usually short and are outnumbered by y-series sequences from length three on. Doubly charged y fragment series represent a small fraction compared to singly charged tags. (**b**) Same as (**a**) but using combined b- and y-ion series. (**c**) Same histogram as in (**a**) but for unfiltered fragment ion spectra. The vertical axes have the same scale. While many short tags are missing due to the lower identification rate, considerably more long tags were found. The absolute number of sequences from doubly charged series has increased 2-fold. In contrast to (**a**), there are now more singly charged b-ions than y-ions in total. (**d**) Number of nonredundant peptides that have a sequence of at least x; 12,769 peptides (47.5% of all identified peptides) have a tag of at least length six. (**e**) Same as (**d**) but with combined b-, y- and y^{2+}-ion series. (**f**) Same as (**d**) without filtering; 15,833 peptides (93.9% of all identified peptides) have a tag of at least length 6.

quences are longer than b-ion sequences on average. Sequences in doubly charged series constitute only a small fraction of the total. We reasoned that some partial sequences may be extendable when connecting between the three ion series. The result of this analysis is plotted in Figure 4b. Indeed, sequences are on average longer by one amino acid when considering all three ion series together. Furthermore, the decay to long sequences is shallower.

When using unfiltered MS/MS spectra, there is a sharp increase in the long sequences, in particular for the ones belonging to the b-series (Figure 3c). Sequences from doubly charged ions increase in absolute numbers as well. This indicates that for some partial sequences found in the filtering approach there are peaks present at low abundance that could either extend or join the sequences consisting of high abundance fragments. Many long b and charge two sequences appear to be present in the data but get shortened or interrupted by the filtering. However, this is difficult to state with certainty because of the high density of peaks in the unfiltered data (Figure 1c), which presents many opportunities to randomly connect fragment ion series.

In Figure 4d, the number of peptides containing a partial sequence of at least length x in any of the three ion series is shown as a function of x. More than 85% of all identified peptides have a tandem spectrum that contains a partial sequence of at least three amino acids, and half of the peptides have spectra that contain a six amino acid fragment sequence. Combining all three ion series and performing the same analysis (Figure 4e) yields sequences that are on average one amino acid longer (just as when counting total sequence occurrence in Figure 4b). Finally, we investigated how many peptides have spectra with partial fragment sequences of length x for unfiltered data. As can be seen in Figure 4f, almost all of these peptides contain sequences of at least 6 amino acids and half of them contain sequences of 9 amino acids.

Uniqueness of Short Peptide Sequences in the Human Proteome and Genome

We next investigated the usefulness of the partial peptide sequences contained in most tandem mass spectra in locating the corresponding site encoding the peptide in the human proteome. For this purpose, we prepared a database containing a single transcript per entry in the ENSEMBL database (see the Experimental section). One can see in Figure 5 that partial sequences of length 4 or shorter are virtually never unique. Partial sequences of four amino acids occur on average in about 100 candidate positions in the proteome. Going to length 5 reduces the number of candidates to 10 on average and 5% of tags are unique in the proteome. A sharp increase in uniqueness follows, and partial sequences of length 7 are already unique in most cases. Here we ignore the non-uniqueness due to proteins that result from alternative splicing of the same gene by selecting for each gene only the isoform with the longest sequence. One would expect that long partial sequences would become completely unique in the proteome. This is, however, not the case; instead, a plateau is reached at about 86%. This is due to the presence of proteins encoded at different gene locations with a high pair-wise sequence similarity, or also due to highly conserved protein domains. For instance, the sequence TGIVMDSGDGVTHTVPIYEGYAL that is found in our dataset should be highly unique. However, we find that it is contained in two protein sequences encoded by two different genes, β-actin (ACTB) and γ-actin (ACTG1), which are located on different chromosomes. Both proteins have a length of 375 amino acids and their sequences differ only at four positions.

Figure 5c shows the histogram of occurrence in the proteome for partial sequences of length 5, 6, and 7. For sequences of length 5, there is still a small fraction that match 10 or more times in the proteome, while for sequences of length 6, half are already unique. For length 7, three-quarters are unique and almost all others only occur twice. Thus, there is little need to de novo sequence more than seven amino acids to uniquely "lock down" the peptide in the human proteome. However, for organisms without sequenced genome, longer amino acid sequences may be desired for homology searching or cloning.

Partial sequences of length 3 or 4 usually occur in 100 to 1000 locations in the human proteome. While this may appear to be a very large number, it is actually very manageable for computer algorithms. Just like in the peptide sequence tag approach, these loci can be expanded in N- and C-terminal direction to obtain a mass match. With only 1000 "seed points" and very high precursor mass accuracy, a very large number of possibilities can be tried to obtain a fit to the measured precursor mass and to the maximum number of measured fragments. Thus, far from being useless, even very short partial sequences should be able to allow unique reconstruction and matching of both the modified and unmodified peptides.

Searching the partial sequences in a complete six frame translation of the human genome resulted in similar patterns as for the proteome. However, due to the larger search space, partial sequences on average had to be longer by two amino acids for the same degree of uniqueness (Figure 5).

Conclusions and Perspectives

Here we have shown that tandem mass spectra from large proteomics projects are surprisingly rich in sequence information. A majority of spectra contains the fragment ions necessary to yield useful sequences. On-going advances in algorithm design, combined with progress in the theoretical understanding [20] and empirical modeling [21] of peptide fragmentation, should

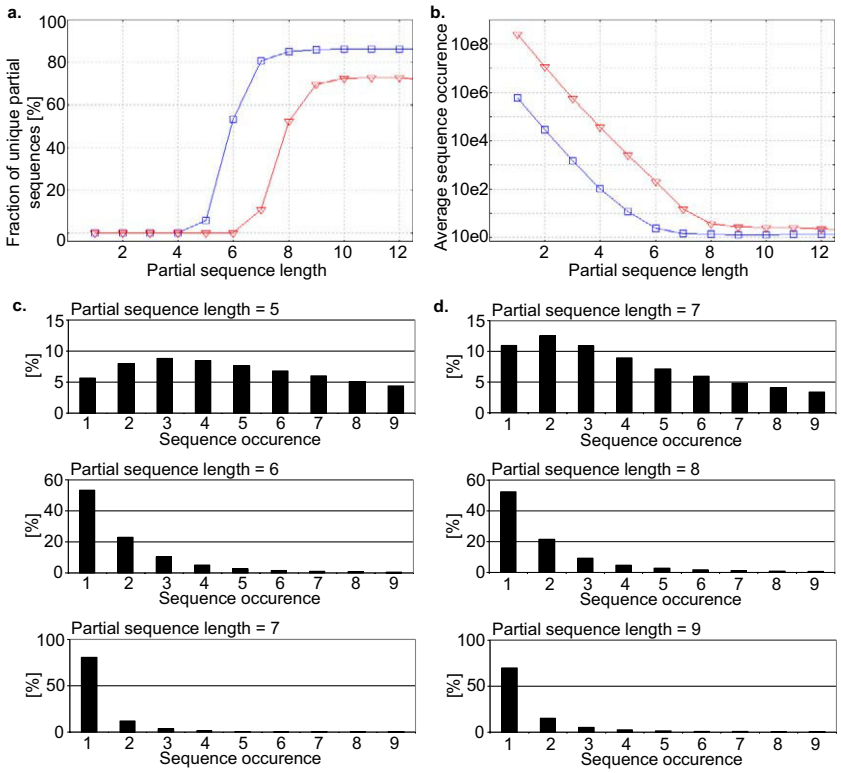

Figure 5. Partial sequence distributions in the human proteome and genome. (**a**) The fraction of unique tags in percent is plotted against sequence length for all identified sequences in the proteome (squares) and genome (triangles). In the proteome, up to length 4, all tags are non-unique. There is a steep crossover at length 6 after which the curve flattens in a plateau at around 86%. The curve for the genome shows similar behavior but it is shifted to sequences that are longer by two amino acids. (**b**) The average number of occurrences of a tag sequence as a function of sequence length in the proteome (squares) and genome (triangles). A tag of length 6 occurs on average about twice in the proteome. (**c**) Distributions of tag occurrences in the proteome separately for tags of length 5, 6, and 7. While for length 5 the distribution extends to larger counts, for length 6 it is beginning to be centered at 1. For length 7 the bins other than the first are sparsely occupied. (**d**) Same as (**c**), but for the genome sequences of length 7, 8, and 9 are plotted.

make it possible to reliably "read out" these sequences from most of the spectra.

There are several obvious directions for future improvements of the data to assists de novo or partial de novo algorithms. One is the use of high-resolution and high mass accuracy in tandem mass spectra. As the required ions are usually present even in large datasets (as shown here), they could be unambiguously identified if sensitivity and dynamic range of MS/MS measurement in a high accuracy setting was improved to approach that of the ion trap. On the LTQ-Orbitrap instrument, a particularly attractive option would be the use of higher energy dissociation (HCD), which does not have a low mass cut-off and which produces "triple quadrupole"-like fragmentation with long y-ion series [22].

Another direction is the more discriminating assignment of peaks in the current low-resolution tandem mass spectra. If the raw data, rather than the centroided data, could be saved, one could employ much more sophisticated algorithms for peak detection than are currently used "on the fly". This is not possible at the moment on the LTQ-Orbitrap because resulting files are larger than 2 Gbytes and cannot be opened by the acquisition software. Once this bottleneck is removed, most of the noise peaks can likely be eliminated, isotope patterns can be modeled, charge states determined and common side-chain losses accounted for, so that signals for the same fragment are collapsed into single, high confidence peaks. Among this smaller number of peaks, the same 'top 6 filtering' would include more sequence relevant ions. Finally, we suggest a 'two-step' strategy, where partial sequences are first found in the usual way (using graph theory as pioneered by Pevzner [23]) among the more intense fragments. Connections between sub-graphs can then be made through low abundance peaks employing empirical, modeling and theoretical knowledge about peptide fragmentation pathways. The first step would guarantee a low rate of false positives, since a tag of a certain length has to be found in the 'high quality' part of the data, while the sequence extension in the low abundant peaks would allow for a higher uniqueness of the tag in the proteome or genome.

Acknowledgments

The authors thank the members of the Department for Proteomics and Signal Transduction for fruitful discussion. The authors acknowledge partial support for this work by "Interaction Proteome", a 6th Framework EU grant.

References

1. Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003,** *422,* 198–207.
2. Steen, H.; Mann, M. The abc's (and xyz's) of Peptide Sequencing. *Nat. Rev. Mol. Cell. Biol.* **2004,** *5,* 699–711.
3. 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,* 3551–3567.
4. Eng, J. K.; McCormack, A. L.; J. R. Yates, I. An Approach to Correlate MS/MS Data to Amino Acid Sequences in a Protein Database. *J Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
5. Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-Scale Database Searching Using Tandem Mass Spectra: Looking Up the Answer in the Back of the Book. *Nat. Methods* **2004,** *1,* 195–202.
6. Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. Scoring Proteomes with Proteotypic Peptide Probes. *Nat. Rev. Mol. Cell. Biol.* **2005,** *6,* 577–583.
7. Cox, J.; Mann, M. Is Proteomics the New Genomics? *Cell* **2007,** *130,* 395–398.
8. Graumann, J.; Hubner, N. C.; Kim, J. B.; Ko, K.; Moser, M.; Kumar, C.; Cox, J.; Schoeler, H.; Mann, M. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol Cell Proteomics.* **2008**, 7, 672-683.
9. Mann, M.; Wilm, M. S. Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994,** *66,* 4390–4399.
10. Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A. Protein Identification by Spectral Networks Analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104,* 6140–6145.
11. Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. De novo Peptide Sequencing and Identification with Precision Mass Spectrometry. *J. Proteome Res.* **2007,** *6,* 114–123.
12. Cox, J.; Mann, M. High Peptide Identification Rates and Proteome-Wide Quantitation Via Novel Computational Strategies. In revision, **2008**.
13. Krutchinsky, A. N.; Kalkum, M.; Chait, B. T. Automatic Identification of Proteins with a MALDI-Quadrupole Ion Trap Mass Spectrometer. *Anal. Chem.* **2001,** *73,* 5066–5077.
14. Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: An Integrated Database for Proteomics Experiments. *Proteomics* **2004,** *4,* 1985–1988.
15. Birney, E.; Andrews, T. D.; Bevan, P.; Caccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cuff, J.; Curwen, V.; Cutts, T.; Down, T.; Eyras, E.; Fernandez-Suarez, X. M.; Gane, P.; Gibbins, B.; Gilbert, J.; Hammond, M.; Hotz, H. R.; Iyer, V.; Jekosch, K.; Kahari, A.; Kasprzyk, A.; Keefe, D.; Keenan, S.; Lehvaslaiho, H.; McVicker, G.; Melsopp, C.; Meidl, P.; Mongin, E.; Pettett, R.; Potter, S.; Proctor, G.; Rae, M.; Searle, S.; Slater, G.; Smedley, D.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Storey, R.; Ureta-Vidal, A.; Woodwark, K. C.; Cameron, G.; Durbin, R.; Cox, A.; Hubbard, T.; Clamp, M. An Overview of ENSEMBL. *Genome Res.* **2004,** *14,* 925–928.
16. Zubarev, R.; Mann, M. On the Proper Use of Mass Accuracy in Proteomics. *Mol. Cell. Proteom.* **2007,** *6,* 377–381.
17. Mann, M. Useful Tables of Possible and Probable Peptide masses. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics,* Atlanta, GA, 1995; p. 639.
18. Zhang, W.; Krutchinsky, A. N.; Chait, B. T. "De Novo" Peptide Sequencing by MALDI-Quadrupole-Ion Trap Mass Spectrometry: A Preliminary Study. *J. Am. Soc. Mass Spectrom.* **2003,** *14,* 1012–1021.
19. Hung, C. W.; Schlosser, A.; Wei, J.; Lehmann, W. D. Collision-Induced Reporter Fragmentations for Identification of Covalently Modified Peptides. *Anal. Bioanal. Chem.* **2007,** *389,* 1003–1016.
20. Paizs, B.; Suhai, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrom. Rev.* **2005,** *24,* 508–548.
21. Zhang, Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem.* **2004,** *76,* 3908–3922.
22. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nat. Methods* **2007,** *4,* 709–712.
23. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J Comput. Biol.* **1999,** *6,* 327–342.