

# Expert System for Computer-assisted Annotation of MS/MS Spectra\*<sup>§</sup>

Nadin Neuhauser<sup>‡¶</sup>, Annette Michalski<sup>‡¶</sup>, Jürgen Cox<sup>‡</sup>, and Matthias Mann<sup>‡§</sup>

An important step in mass spectrometry (MS)-based proteomics is the identification of peptides by their fragment spectra. Regardless of the identification score achieved, almost all tandem-MS (MS/MS) spectra contain remaining peaks that are not assigned by the search engine. These peaks may be explainable by human experts but the scale of modern proteomics experiments makes this impractical. In computer science, Expert Systems are a mature technology to implement a list of rules generated by interviews with practitioners. We here develop such an Expert System, making use of literature knowledge as well as a large body of high mass accuracy and pure fragmentation spectra. Interestingly, we find that even with high mass accuracy data, rule sets can quickly become too complex, leading to over-annotation. Therefore we establish a rigorous false discovery rate, calculated by random insertion of peaks from a large collection of other MS/MS spectra, and use it to develop an optimized knowledge base. This rule set correctly annotates almost all peaks of medium or high abundance. For high resolution HCD data, median intensity coverage of fragment peaks in MS/MS spectra increases from 58% by search engine annotation alone to 86%. The resulting annotation performance surpasses a human expert, especially on complex spectra such as those of larger phosphorylated peptides. Our system is also applicable to high resolution collision-induced dissociation data. It is available both as a part of MaxQuant and via a webserver that only requires an MS/MS spectrum and the corresponding peptides sequence, and which outputs publication quality, annotated MS/MS spectra ([www.biochem.mpg.de/mann/tools/](http://www.biochem.mpg.de/mann/tools/)). It provides expert knowledge to beginners in the field of MS-based proteomics and helps advanced users to focus on unusual and possibly novel types of fragment ions. *Molecular & Cellular Proteomics* 11: 10.1074/mcp.M112.020271, 1500–1509, 2012.

In MS-based proteomics, peptides are matched to peptide sequences in databases using search engines (1–3). Statistical criteria are established for accepted *versus* rejected pep-

tide spectra matches based on the search engine score, and usually a 99% certainty is required for reported peptides. The search engines typically only take sequence specific backbone fragmentation into account (*i.e.* a, b, and y ions) and some of their neutral losses. However, tandem mass spectra—especially of larger peptides—can be quite complex and contain a number of medium or even high abundance peptide fragments that are not annotated by the search engine result. This can result in uncertainty for the user—especially if only relatively few peaks are annotated—because it may reflect an incorrect identification. However, the most common cause of unlabeled peaks is that another peptide was present in the precursor selection window and was cofragmented. This has variously been termed “chimeric spectra” (4–6), or the problem of low precursor ion fraction (PIF)<sup>1</sup> (7). Such spectra may still be identifiable with high confidence. The Andromeda search engine in MaxQuant, for instance, attempts to identify a second peptide in such cases (8, 9). However, even “pure” spectra (those with a high PIF) often still contain many unassigned peaks. These can be caused by different fragment types, such as internal ions, single or combined neutral losses as well as immonium and other ion types in the low mass region. A mass spectrometric expert can assign many or all of these peaks, based on expert knowledge of fragmentation and manual calculation of fragment masses, resulting in a higher degree of confidence for the identification. However, there are more and more practitioners of proteomics without in depth training or experience in annotating MS/MS spectra and such annotation would in any case be prohibitive for hundreds of thousands of spectra. Furthermore, even human experts may wrongly annotate a given peak—especially with low mass accuracy tandem mass spectra—or fail to consider every possibility that could have resulted in this fragment mass.

Given the desirability of annotating fragment peaks to the highest degree possible, we turned to “Expert Systems,” a well-established technology in computer science. Expert Systems achieved prominence in the 1970s and 1980s and were meant to solve complex problems by reasoning about knowl-

From the <sup>‡</sup>Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

Received May 5, 2012, and in revised form, July 19, 2012

<sup>§</sup> Author's Choice—Final version full access.

Published, MCP Papers in Press, August 10, 2012, DOI 10.1074/mcp.M112.020271

<sup>1</sup> The abbreviations used are: PIF, Precursor Intensity Fraction; FDR, False Discovery Rate; MS/MS, Tandem mass spectrometry; HCD, Higher Energy Collisional Dissociation; PEP, Posterior Error Probability; PDF, Portable Document Format; IM, immonium ion; SC, side chain fragment ion; Th, Thomson.

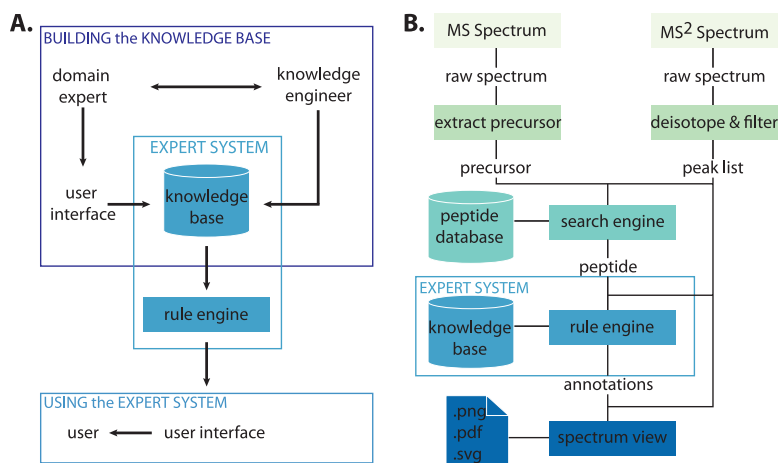


FIG. 1. **Basic concept of the Expert System.** A, An Expert System is constructed by interviewing an expert in the domain (here peptide fragmentation and the accumulated literature) and devising a set of rules with associated priority and dependence on each other. The knowledge base contains the rules whereas the rule engine is generic and applies the rules to the data. B, Data are automatically processed following the steps depicted.

edge (10, 11). Interestingly, one of the first examples was developed by Nobel Prize winner Joshua Lederberg more than 40 years ago, and dealt with the interpretation of mass spectrometric data. The program's name was Heuristic DENTRAL (12), and it was capable of interpreting the mass spectra of aliphatic ethers and their fragments. The hypotheses produced by the program described molecular structures that are plausible explanations of the data. To infer these explanations from the data, the program incorporated a theory of chemical stability that provided limiting constraints as well as heuristic rules.

In general, the aim of an Expert System is to encode knowledge extracted from professionals in the field in question. This then powers a rule-based system that can be applied broadly and in an automated manner. A rule-based Expert System represents the information obtained from human specialists in the form of IF-THEN rules. These are used to perform operations on input data to reach appropriate conclusion. A generic Expert System is essentially a computer program that provides a framework for performing a large number of inferences in a predictable way, using forward or backward chains, backtracking, and other mechanisms (13). Therefore, in contrast to statistics based learning, the "expert program" does not know what it knows through the raw volume of facts in the computer's memory. Instead, like a human expert, it relies on a reasoning-like process of applying an empirically derived set of rules to the data.

Here we implemented an Expert System for the interpretation for high mass accuracy tandem mass spectrometry data of peptides. It was developed in an iterative manner together with human experts on peptide fragmentation, using the published literature on fragmentation pathways as well as large data sets of higher-energy collisional dissociation (HCD) (14) and collision-induced dissociation (CID) based peptide identifications. Our goal was to achieve an annotation perform-

ance similar or better than experienced mass spectrometrists (15), thus making comprehensively annotated peptide spectra available in large scale proteomics.

#### EXPERIMENTAL PROCEDURES

The benchmark data set is from Michalski *et al.*<sup>2</sup> Briefly, *E. coli*, yeast and HeLa proteomes were separated on 1D gel electrophoresis and in gel digested (16). Resulting peptides were analyzed by liquid chromatography (LC) MS/MS on a linear ion trap - Orbitrap instrument (LTQ Velos (17) or ELITE (18), Thermo Fisher Scientific). Peptides were fragmented by HCD (14) or by CID, but in either case fragments were transferred to the Orbitrap analyzer to obtain high resolution tandem mass spectra (7500 at  $m/z$  400). We scanned tandem mass spectra already from  $m/z$  80 to capture immonium ions as completely as possible. Data analysis was performed by MaxQuant using the Andromeda search engine (8, 9). Maximum initial mass deviation for precursor peaks was 6 ppm and maximum deviation for fragment ions for both the search engine and for the Expert System was 20 ppm. MaxQuant preprocessed the spectra to be annotated by the Expert System in the same way as it does for the Andromeda search engine: Peaks were filtered to the 10 most abundant ones in a sliding 100  $m/z$  window, de-isotoped and shifted to charge one where possible. From this data, sequence-spectra pairs were selected that had a certainty of identification of 99.99% PIF values (7) larger than 95% and that were sequence unique (more than 16,000 peptides).

The Expert System was written in the programming language C#, using the Microsoft .NET framework version 3.5 and the Workflow Activities library, which contains a rule engine to implement an Expert System (Microsoft Corporation, Redmond, WA).

MaxQuant contains the Expert System as an integrated option in its Viewer—the component that allows visualization of raw and annotated MS data. MaxQuant can freely be downloaded from [www.maxquant.org](http://www.maxquant.org). It requires Microsoft .NET 3.5, which is either already installed with Microsoft Windows or can be installed as a free Windows update. In our group we have implemented the Expert System both on a Windows cluster and in a desktop version. Additionally, we provide an Expert System web server, which can be accessed at

<sup>2</sup> Michalski, A., Neuhauser, N., Cox, J., and Mann, M., unpublished data.

www.biochem.mpg.de/mann/tools/. Although MaxQuant allows the Expert System annotation of arbitrary numbers of MS/MS spectra, the webserver is currently limited to the submission of one MS/MS spectrum at a time. After upload of a list of peaks with  $m/z$  value and their intensities—together with the corresponding peptide sequence—the spectrum with all annotations is displayed. This can then be exported in different graphical formats.

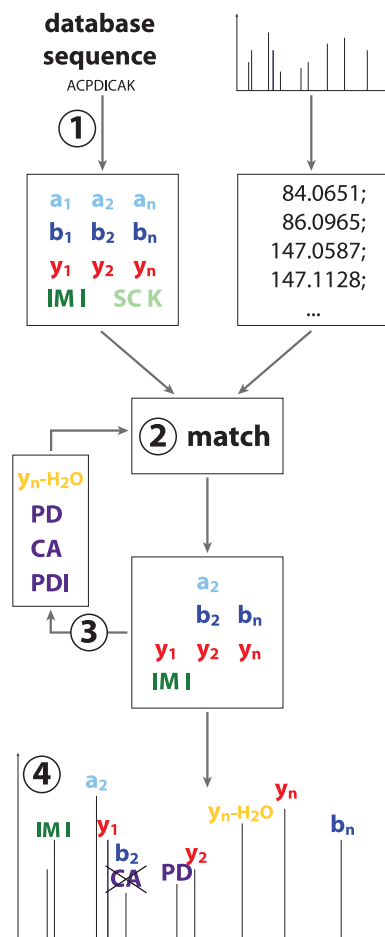
#### RESULTS AND DISCUSSION

**Construction of the Expert System**—Human experts perform a generic set of tasks when solving problems such as the interpretation of an MS/MS spectrum. These rules have to be codified in the Expert System, mainly in the form of a series of IF-THEN rules. Fig. 1 shows the major steps involved in building and using the Expert System. It is important to acquire all relevant rules to interpret MS/MS spectra as comprehensively as possible. However, to avoid over-annotation leading to false positives (see below), the number of rules and their interactions should not become too large. This balance was struck by evaluating the performance of different set of rules on large data sets in conjunction with human experts.

Rules were encoded in a table-like structure, where they could be activated, deactivated or modified. To create the knowledge base, the extent of interactions of the rules also had to be determined—for instance, which combination of neutral losses to allow. After iterative construction of the knowledge base, the rule engine then applied the encoded knowledge to MS/MS spectra and displayed the result to the user (Fig. 1A). The processing steps that are performed on the raw MS and MS/MS spectra are shown in Fig. 1B (see also EXPERIMENTAL PROCEDURES). Note that the workflow is entirely automated and that user interaction is possible but not required. Arbitrary numbers of annotated spectra of peptides of interest can be produced as interactive screen images or high resolution, printable PDF files. The Expert System is very fast, and 16,000 spectra can be annotated in less than four hours on a desktop system.

The IF-THEN constraints of our Expert System can be divided into four major parts (Fig. 2). At first the Expert System calculates any specific backbone fragments ( $a$ ,  $b$ , and  $y$ -ion series), the charged precursor ion, the immonium ions as well as side chain fragments in the low-mass region and places them into a queue. In the second part of the workflow every element in this queue is filtered with respect to the actual MS/MS spectrum. Even if there is a peak corresponding to a calculated item in the queue, it may still be filtered out (symbolized by missing annotations after the filter in Fig. 2). For instance, a  $b_1$  ion is only allowed in very restricted circumstances.

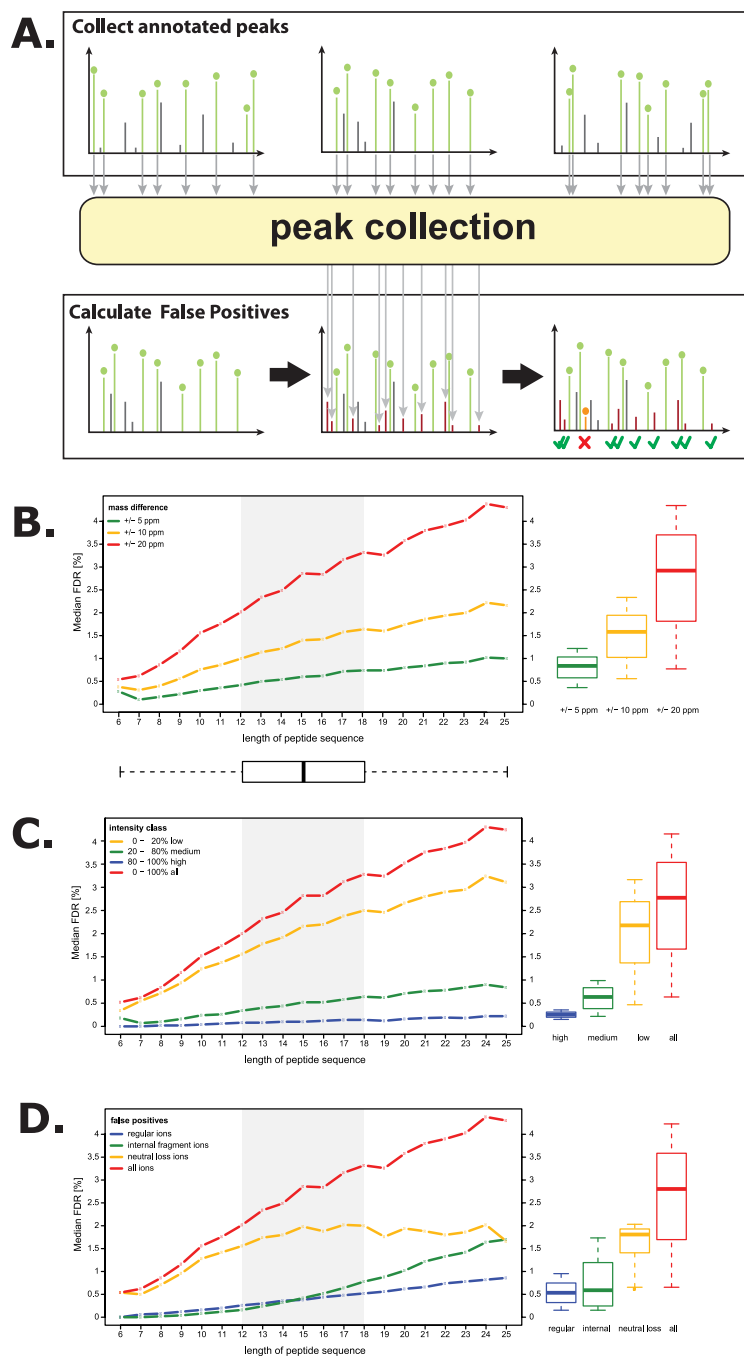
In the third step, neutral losses and internal fragments for the filtered values are calculated and added to the queue. They are then subjected to the same filtering rules as in step 2. Step 3 is iterative, as several subsequent neutral losses may be allowed.



**Fig. 2. Work flow of the Expert System.** ① From the database sequence of the peptide identified by the search engine, a list of possible fragment ions is created. ② Peaks from the measured spectrum are compared with the possible fragments and preliminarily annotated if they pass the rules of the Expert System. ③ Neutral losses and internal fragments are generated from the candidate, annotated peaks and exposed to the Expert System rules. ④ Potential conflicts are resolved via the priority of the annotations and peaks are labeled. Note that possible internal fragment 'CA' is crossed out because the  $b_2$  ion has the higher priority.

In the fourth and last step each potential annotation is given a priority. If there is more than one possible annotation, the one with the highest priority is chosen (*i.e.* the one that triggered the rules with higher priority). However, in this case the Expert System provides a pop-up (or "tool-tip") containing the other possibility when hovering the mouse over the peak. (This can still happen if the FDR is properly controlled and is then typically caused by two different chemical designations for the same ion; or by different ions with the same chemical composition, such as small internal fragments with different sequence but the same amino acids).

**Determining a False Discovery Rate for Peak Annotation**—Use of a very high threshold for peptide identification (99.99%) ensured that virtually none of the peptides in our collection should be misidentified. However, when building



**FIG. 3. Calculation of false discovery rate for peak annotations.** *A*, The upper panels represent a large number of identified MS/MS spectra from which annotated peaks are drawn to form a large peak collection of possible fragment masses. From each identified spectrum in the data set, 10 random fragments are inserted and the number of annotations by the Expert System is counted. This process is repeated 500 times for each peptide. *B*, Median FDR as determined in *A* as a function of peptide length distinguished by the mass difference of fragment ion and theoretical mass. The FDR for peak annotation rises with peptide length and is strongly dependent on the mass difference. Box plot at the bottom shows that 50% of the peptides were between 12 and 18 amino acids long. The box plots on the right summarize the range of FDR values regardless of peptide length. *C*, Graph of the median FDR as a function of peptide length but separated by intensity classes of the false annotated fragment peaks. Most false positives come from the low abundant peaks (*blue*) rather than the medium (*green*) or high abundance fragment peaks (*yellow*). *D*, Same plot as above but differentiated by the fragment ion type of the false positives. Getting lower number of false positives from regular fragment annotations (*blue*), compared with internal fragment (*green*) and neutral loss annotations (*yellow*).



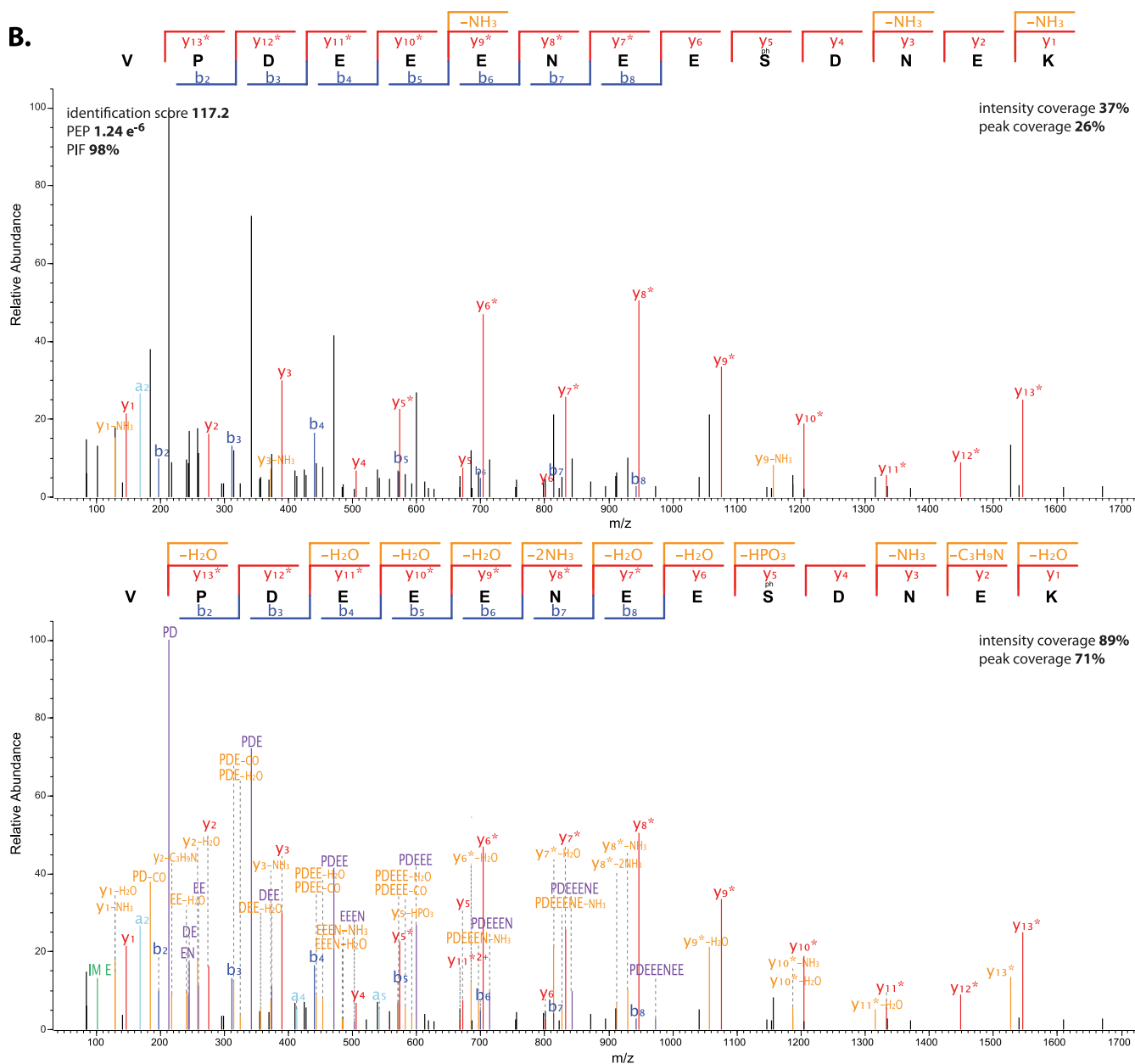
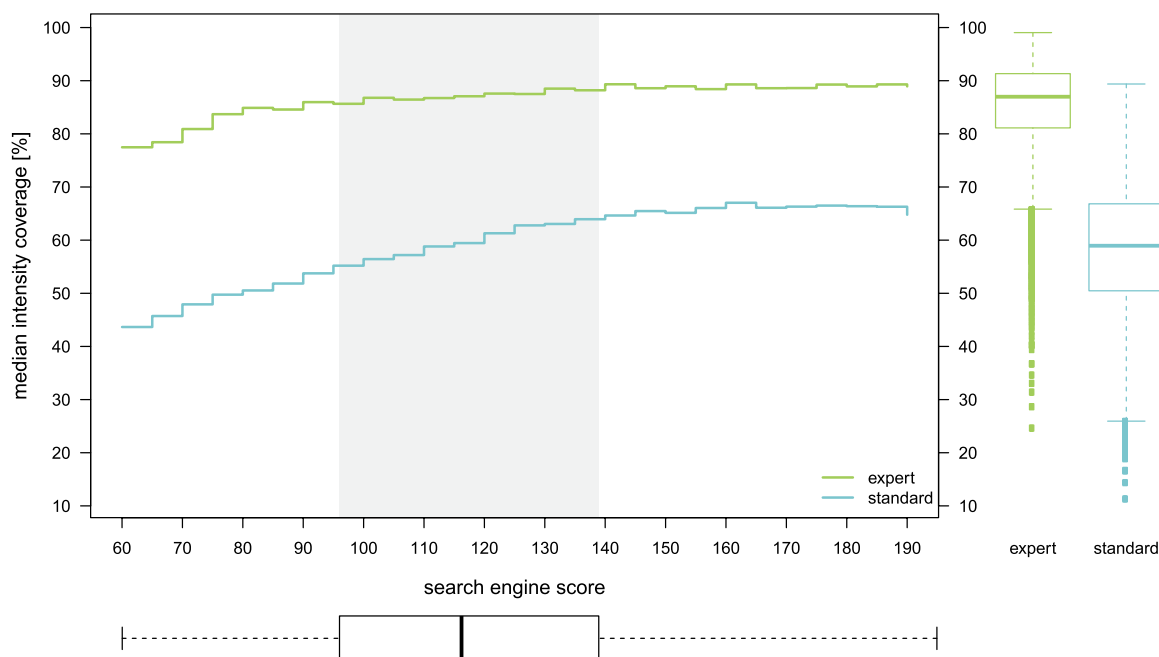


FIG. 4—continued

chance because its mass fits one of the Expert System rules for the peptide sequence. To calculate a proper FDR, we therefore needed to provide a set of background peaks that would represent false positives when they are labeled by the Expert System. Producing realistic background peaks turned out to be far from trivial because they need to have possible masses that can in principle be generated from peptide sequences and they need to be independent of the sequence of the peptide in question. The principle of our solution to this problem is shown in Fig. 3A. From the large data set underlying this study, we collect the  $m/z$  values of all annotated peaks, except those coming from immonium or side chain

ions. They were stored in a large peak collection of several million entries, together with the respective peptide sequences and the relative intensity of the peak. For each spectrum in which we wanted to determine the FDR, we then inserted a random set of 10 peaks from the collection, where after we checked if the sequence of the selected peaks was independent from the sequence of the current spectrum. If one of the inserted peaks overlapped with an existing peak, it was discarded. By definition these 10 peaks represent possible peptide fragments and, because they are chosen randomly from millions of other peaks, they collectively represent a good approximation to a true background set. This would





**FIG. 5. Expert System performance on a large data set.** Median sequence coverage by summed fragment ion intensity is plotted as a function of identification score. Statistics is based on more than 16,000 spectra. For every identification score, the Expert System adds a large proportion of explainable peaks. Box plot below the graph indicates that 50% of peptides in the set have an Andromeda score between 98 and 140. Box plots on the right indicate the range of values for the intensity coverage for standard and Expert System annotation.

not be the case for permutation of the sequence of the precursor in question, for instance, because many of the fragment peaks in permuted sequences are identical. Whenever the Expert System annotated one of these peaks, it was counted as a false positive. To find the number of repeats necessary to obtain a stable FDR for this procedure, we chose a set of spectra and simulated a thousand times on each one. We found that the FDR was constant after 500 iterations. For the final FDR calculation, for each spectrum we added a different set of 10 random peaks from the collection and repeated this 500 times. This was then applied to each of the more than 16,000 pure (high PIF) spectra in the large scale data set.

Beyond providing a solid FDR estimate for each rule set, this procedure also allowed us to identify the rules or rule combinations that were responsible for miss-annotation, *i.e.* the rules that falsely annotated the inserted peaks. These mostly turned out to be chains of subsequent neutral losses. In conjunction with detailed evaluation of the frequency of ion types, we iteratively designed an optimal rule set (supplemental Table S1). For instance, neutral losses from a particular amino acid were allowed if they occurred in more than five percent of the fragment sequences that contained that amino acid. Likewise, of a set of about 42 possible neutral side chain losses, only six were sufficiently important to retain them in the Expert System. The Figs. 3B–3D show the results of the median FDR as a function of the peptide length based on this final rule set. The overall FDR—indicated in red—is the same in all plots and shows a clear growing trend in the number of

false positives with the length of the peptides. For small peptides of 12 amino acids or less, the FDR was less than 2.1% and all peptides in the range investigated had a peak annotation FDR of less than 5%. With these settings, the annotations are correct in more than 97% of the cases for the vast majority of MS/MS spectra. The Expert System could of course be pruned to provide a lower FDR by narrowing the mass tolerance window; however, this would come at the expense of discarding correct annotations. To explore the influence of mass accuracy on potential false positive annotations, we repeated these calculations with required mass deviations no larger than 5 ppm or no larger than 10 ppm. As can be seen in Fig. 3B, this further reduced possible errors to less than 1%, or less than 0.3%, respectively. This highlights the value of high mass accuracy in unambiguously identifying fragment mass identity.

Furthermore, peaks with a low signal to noise are more likely to be miss-annotated than more intense peaks. In Fig. 3C we sorted the peak intensity of the false positives into three intensity classes (Fig. 3C). The median FDR of peaks with high or medium abundance are only 0.1 or 0.5%. For low abundance peaks it is higher but still with a median of no more than 2.1%.

Next we separately investigated the FDR as a function of peptide length for the different fragment ion types. As can be seen in Fig. 3C, regular ions and internal fragments contribute very little to overall false annotation (0.4 and 0.5%), whereas neutral loss ions are wrongly annotated in 1.8% of the case or even more.

**A.** copy & paste list of  $m/z$ -value \t intensity**Peak List**

Update

Number of peaks:53

101.07084	242861.7
115.08653	6455070
143.08148	5373670
158.09233	284282.1
159.07645	251449.4
169.09703	296378.4
175.11895	1177980
187.07145	366193.3
214.11883	394028.6
225.12225	555222.2

**B.** insert peptide sequence and modifications**Identification**

Update

Peptide sequence:

AAADVQLR

Peptide charge:

2+

Modifications:

**position site modification****delta mass**

N-Term

None

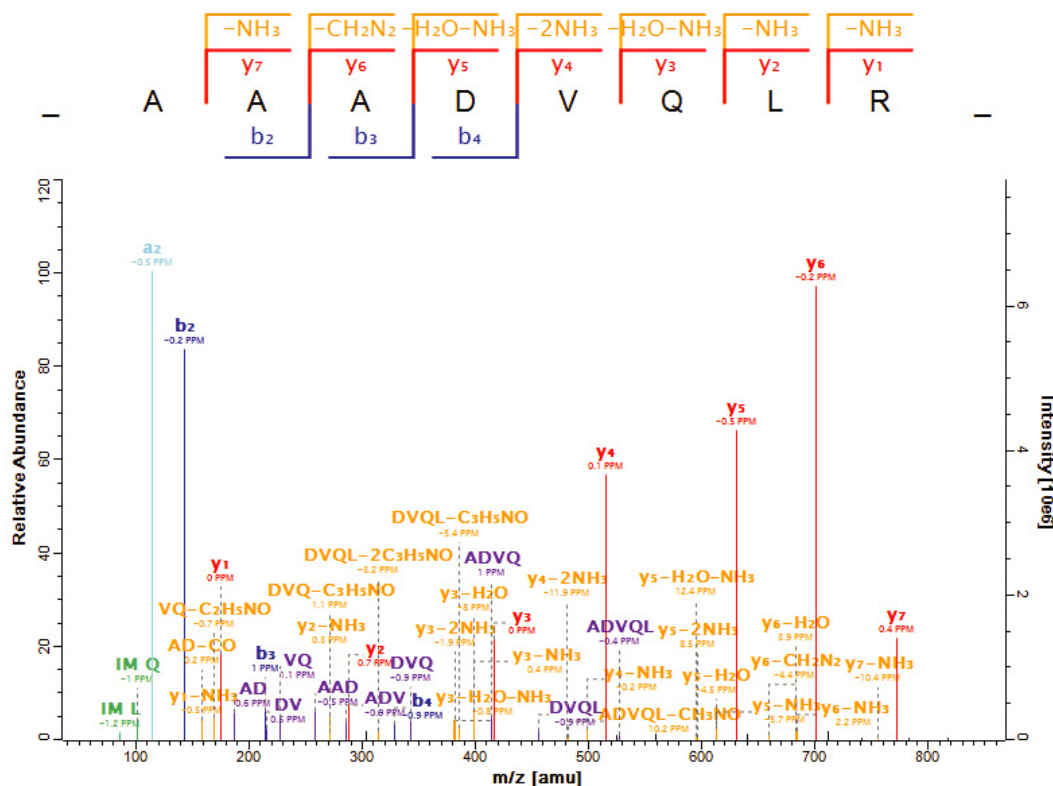
None

None

None

C-Term

None

**C.** inspect/export annotated spectrum

**FIG. 6. Web interface for the Expert System.** A, Text field to paste the spectrum in text format ( $m/z$  value; intensity in arbitrary units). B, Form to enter the peptide sequence, modifications and their positions. C, Detected backbone fragments and their neutral losses are indicated in the peptide logo. Scalable spectrum annotated by the Expert System. Note that neutral loss peaks are very small compared with the major backbone fragments. The spectrum can be downloaded with the desired resolution and in the desired graphical format.

**Performance of the Expert System**—Fig. 4 shows an illustrative example of an HCD fragmented peptide before and after Expert System evaluation. The peptide was identified with an Andromeda score of 136 and posterior error probability (PEP) of  $1.1E-21$  (the corresponding Mascot score was 83). The spectrum features an uninterrupted b-ion series from  $b_2$  to  $b_9$  and an uninterrupted y-ion series from  $y_1$  to  $y_{12}$ , together covering the entire peptide sequence. Despite this unambiguous identification, the peaks used by the search engine to identify the peptide only accounted for 35% of the

summed intensity of the peaks in the fragmentation spectrum. Coverage by number of explained peaks was even lower at 24% (allowing up to 10 peaks per 100 Th in the measured spectrum see EXPERIMENTAL PROCEDURES). There is a series of high abundance, high  $m/z$  fragments as well as a large number of low abundance peaks in the low and medium  $m/z$  range that are unexplained by the search engine. After annotation by the Expert System, this situation changes entirely. The high  $m/z$  series is revealed to be a prominent loss of  $CH_4SO$  from oxidized methionine. The low



mass ions are neutral losses, internal fragments and combinations between them and they were unambiguously and correctly assigned. Altogether, the Expert System accounted for almost all prominent ions and explained a total of 88% of the ion current. Manual annotation of this spectrum would have been possible but would have been very time consuming.

Interpretation of phosphorylated peptides, especially large ones, is more difficult than that of unmodified peptides. Furthermore, accurate placement of the phosphorylation site can be challenging. We used literature knowledge (19, 20) and the results of a large-scale investigation into the fragmentation of phosphorylated peptides to derive suitable fragmentation rules for the Expert System. This led to an additional six rules, which were easily integrated, illustrating the extensibility of the Expert System. Fig. 4B depicts an example annotation of the relatively complex fragmentation spectra typical of phosphorylated peptides. The large ion series from the low mass range to about mass 1000 is caused by an extensive and uninterrupted internal ion series starting from the proline in the second position of the peptide sequence. As these internal fragments contain several glutamines, they lead to additional water and ammonia losses. However, there are also newly annotated fragments resulting from neutral losses in addition to loss of the phosphorylation site. Moreover, the neutral loss of  $\text{HPO}_3$  is annotated.

*Large-scale Evaluation of the Performance of the Expert System*—We used the population of 16,000 spectra with high PIF—identified with a false discovery rate of 0.01% by the search engine—and annotated them automatically using the Expert System. For each spectrum we calculated the intensity coverage obtained by the fragments used by the search engine and the fragments explained by the Expert System. Higher scoring fragmentation spectra would be expected to have a larger fraction of their ion current annotatable than lower scoring peptides. Fig. 5A shows a plot of the median of these values for all search engine scores. A total of 95% of these Andromeda scores are within a range of 96 to 138. Here the median intensity coverage by standard annotation varies from 55% at 96 to 64% at 138. The Expert System, in contrast, annotated between 86 and 89% of the total ion current in the fragment spectra of the same peptides. This represents an average increase of 28%. There was only a small percentage of peptides that were lower scoring than 96 and for these the increased annotation percentage of the Expert System was even larger (34%). Interestingly, even in very high scoring HCD fragment spectra there are still many peaks not directly annotated by the search engine. For these, the average increase of annotated ion current because of the Expert System was still 23%.

The rule set of the Expert System was derived from HCD data. However, HCD and CID appear to produce similar ion types, although with different abundances. We therefore tested if the derived rule set was also applicable to high

resolution CID data. This was indeed the case, and a total of 85% of the ion current in high resolution CID spectra explained by the Expert System, although in CID spectra a higher percentage (79%) of the peaks are already accounted for by standard ion types. Therefore we conclude that the Expert System can be used equally well for high resolution HCD and CID data although the benefits for CID are not as large as they are for HCD.

*Webserver for Expert System Annotation of Spectra*—The Expert System is now part of the Viewer component of MaxQuant, which is freely available at [www.maxquant.org](http://www.maxquant.org). In this environment, the Expert System can annotate arbitrarily large data sets of identified peptides and visualize and export them in different graphical formats such as PDF. Additionally, we established a webserver to make the Expert System available to any proteomics scientist, regardless of the computational workflow that he or she is using. The webserver is located at <http://www.biochem.mpg.de/mann/tools/> and its graphical interface is shown in Fig. 6. The user needs to supply a mass spectrum in the form of an  $m/z$  and peak intensity list as well as the sequence of the identified peptide (Figs. 6A, 6B). Common modifications and their position in the sequence can also be specified. The webserver then provides an annotation of the spectrum within the stated mass tolerance as shown in Fig. 6C. The graph is scalable to enable detailed study of complex fragmentation spectra. Mass deviations in ppm (calculated mass – measured mass) can also be depicted. This annotated spectrum can be downloaded in a number of graphical formats for use in publications.

### CONCLUSION AND OUTLOOK

Here we have made use of Expert Systems—a well-known technology in computer science—to automatically but accurately interpret the fragmentation spectra of identified peptides. We have shown that the Expert System performs very well on high mass accuracy data, annotating the large majority of medium to high abundance peaks. For HCD spectra it explains on average 28% more of the peak intensities than the search engine results alone. We derived a rigorous false positive rate, ensuing that less than 5% of peaks can be miss-annotated—this rate is even lower for spectra with at least median scores and fragment ion intensities of at least moderate abundance. The rule set was derived by iterative interpretation of large HCD data set but we show that the Expert System is equally applicable to high resolution CID spectra.

We envision different uses for the Expert System: For beginners in MS-based proteomics, it enables efficient training in the interpretation of MS/MS spectra without requiring much input from a specialist. For advanced users, it allows focusing on unusual and potentially novel types of fragments. One caveat is that the Expert System currently cannot explain fragment peaks that belong to cofragmented precursors; a very common occurrence that we deliberately avoided here by selecting only pure MS/MS spectra. This limitation can be

addressed if both precursors are identified and communicated to the Expert System. Such a feature might be particularly useful for instruments that allow deliberate multiplexing of precursors, which leads to complex MS/MS spectra (21).

The Expert System has been in routine use in our laboratory for a number of months. During this time we have found that it provides helpful confirmation of the identification of the peptide and the identity of the previously unlabeled fragment ions. This is particularly welcome in the case of complicated spectra of important peptides, such as the ones regulated in the biological function in question. Compared with a human expert, the principal advantages of the Expert System are its speed, its ability to check for all supplied rules in a consistent manner as well as its rigorously controlled false positive rate. Obviously, the Expert System is limited to the knowledge supplied whereas an experienced mass spectrometrists can go beyond these rules and discover the origin of novel fragmentation mechanisms.

As we have shown here, Expert Systems can readily be applied to problems in computational proteomics. Given their relative ease of implementation, they may become useful in other areas in MS-based proteomics, too.

*Acknowledgments*—We thank Forest White for critical comments on this manuscript.

\* This work was supported by funding from the European Union 7<sup>th</sup> Framework project PROSPECTS (Proteomics Specification in Time and Space, grant HEALTH-F4-2008-201645).

§ This article contains [supplemental Table S1](#).

¶ These authors contributed equally.

§ To whom correspondence should be addressed: Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de.

#### REFERENCES

1. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711
2. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797
3. Granholm, V., and Käll, L. (2011) Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* **11**, 1086–1093
4. Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N. G., and Old, W. M. (2010) Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **9**, 4152–4160
5. Zhang, N., Li, X. J., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. (2005) ProbiDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **5**, 4096–4106
6. Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G., Toth, M. J., and MacCoss, M. J. (2010) Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82**, 833–841
7. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10**, 1785–1793
8. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372
9. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805
10. Giarratano, J. C., and Riley, G. (2005) *Expert systems: principles and programming*. PWS Pub. Co., Boston
11. Liao, S. H. (2005) Expert system methodologies and applications - a decade review from 1995 to 2004. *Expert Syst. Appl.* **28**, 93–103
12. Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J. (1969) Applications of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data. *J. Am. Chem. Soc.* **91**, 7440–7445
13. Russell, S. J., Norvig, P., and Davis, E. (2010) *Artificial intelligence: a modern approach*. Prentice Hall, Upper Saddle River, NJ
14. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712
15. Bin, M., and Johnson, R. (2012) De novo sequencing and homology searching. *Mol. Cell. Proteomics* **11**, O111.014902
16. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860
17. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8**, 2759–2769
18. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J. J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11**, 10.1074/mcp.O111.013698
19. Boersema, P. J., Mohammed, S., and Heck, A. J. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* **44**, 861–878
20. Kelstrup, C. D., Hekmat, O., Francavilla, C., and Olsen, J. V. (2011) Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. *J. Proteome Res.* **10**, 2937–2948
21. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, 10.1074/mcp.M111.011015