Journal of proteome research

ARTICLE

pubs.acs.org/jpr

Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment

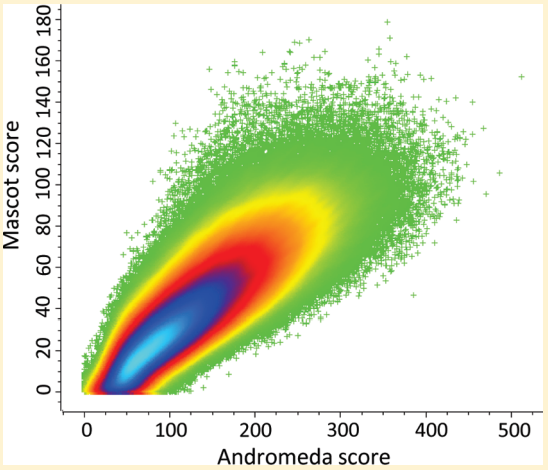
Jürgen Cox,[*,†] Nadin Neuhauser,[†] Annette Michalski,[†] Richard A. Scheltema,[†] Jesper V. Olsen,[‡] and Matthias Mann[*,†,‡]

[†]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[‡]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark

Supporting Information

ABSTRACT: A key step in mass spectrometry (MS)-based proteomics is the identification of peptides in sequence databases by their fragmentation spectra. Here we describe Andromeda, a novel peptide search engine using a probabilistic scoring model. On proteome data, Andromeda performs as well as Mascot, a widely used commercial search engine, as judged by sensitivity and specificity analysis based on target decoy searches. Furthermore, it can handle data with arbitrarily high fragment mass accuracy, is able to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides, and accommodates extremely large databases. The algorithms of Andromeda are provided. Andromeda can function independently or as an integrated search engine of the widely used MaxQuant computational proteomics platform and both are freely available at www.maxquant.org. The combination enables analysis of large data sets in a simple analysis workflow on a desktop computer. For searching individual spectra Andromeda is also accessible via a web server. We demonstrate the flexibility of the system by implementing the capability to identify cofragmented peptides, significantly improving the total number of identified peptides.



KEYWORDS: tandem MS, search engine, spectrum scoring, post-translational modifications, mass accuracy, collision induced dissociation, higher-energy collisional dissociation, Orbitrap

INTRODUCTION

Mass spectrometry (MS)-based proteomics is becoming a commonly used technology in a wide variety of biological disciplines.[1−6] In a "shotgun" format, very complex peptide mixtures are produced by enzymatic digestion of protein mixtures, which are analyzed by liquid chromatography followed by tandem mass spectrometry.[7,8] Per LC−MS/MS run, thousands of MS and MS/MS scans are acquired, often producing gigabytes of high resolution data per day and per mass spectrometer. Computational proteomics has become a key research area, dealing with the challenges of how to most efficiently extract protein identification and quantification results from the raw data. Both the proteomics community and the bioinformatics community have dealt with many areas of this novel field, and there is already a large literature outlining and reviewing the general tasks involved,[9−17] particular computational aspects of the field[18−22] and integrated data analysis pipelines.[23−30]

In this context, our group has developed the MaxQuant environment, a computational proteomics workflow that addresses the above tasks with a focus on high accuracy and quantitative data. It includes peak detection in the raw data, quantification, scoring of peptides and reporting of protein groups.[31] MaxQuant takes advantage of high resolution data such as those obtained by the linear ion trap−Orbitrap instruments and employs algorithms that determine the mass precision and accuracy of peptides individually. This leads to greatly enhanced peptide mass accuracy that can be used as a filter in database searching.[32] MaxQuant was also specifically designed to achieve the highest possible quantitative accuracy in conjunction with stable isotope labeling with amino acids in cell culture (SILAC).[33,34] Using high resolution data combined with individualized mass accuracies and robust peptide and protein scoring results in high peptide identification rates of typically 50% and even higher on SILAC peptide pairs.[31] This was an important foundation for the quantification of the first complete model proteome, that of budding yeast.[35]

Received: October 23, 2010
Published: January 21, 2011

The MaxQuant environment originally used the Mascot peptide search engine[36] to match tandem mass spectra to possible peptide sequences. Mascot together with SEQUEST[37] are commonly used search tools in proteomics today. However, there are many others including Protein prospector,[38] ProbID,[39] X!Tandem;[40] OMSSA,[41] ProSight[42] and Inspect[43] (see Nesvizhskii et al. for a review[14]). Mascot takes a probability based approach to match sequences from a database to tandem mass spectra.[36] Because it is a commercial program the exact algorithms it employs are neither known nor available for modification. Furthermore, Mascot is implemented in a client-server configuration, which imposes practical restrictions for some applications such as real-time searches. We therefore set out to develop a new search engine that would be free of these restrictions. We aimed at performance at least on par with Mascot, which has become a "gold standard" in proteomic analysis, and robustness for scaling up to extremely large and complex data sets. In combination with MaxQuant, the new search engine would then enable analysis of complex data sets on desktop machines by any proteomics researcher or biologist wishing to employ proteomics.

Database searching with fragment mass spectra typically follows one of three approaches:[44,45] (i) deriving a partial or full peptide sequence with associated mass information (first implemented by PeptideSearch[46] and graph theory based *de novo* methods[47]), (ii) autocorrelation between the experimental and a calculated spectrum (first used in SEQUEST) or (iii) calculating a probability that the observed number of matches between the calculated and measured fragment masses could have occurred by chance (pioneered in Mascot). We chose the probability based approach based on the binominal distribution probability and started from a score that we had originally developed for analyzing MS^3 data for which no search software was available at the time.[48] This score has already been used for ranking the peptides in MaxQuant searches from the beginning and it also determines the localization probability of modifications in peptides.[48]

In this paper, we describe the architecture of the Andromeda search engine and its scoring function. We perform a rigorous comparison against the Mascot search engine on several large-scale data sets. The ability of Andromeda to accurately handle many modifications of the same peptide is demonstrated. Due to the complexity of peptide mixtures in shotgun proteomics and the relatively low resolution of precursor isolation, two peptides are frequently 'cofragmented' and there are algorithms that try to identify them from mixture spectra.[49−52] We demonstrate the flexibility of the Andromeda search engine by implementing a novel second peptide identification algorithm.

MATERIALS AND METHODS

Benchmark Data Sets

Raw data from 84 LC−MS runs was taken from Luber et al,[53] a label-free proteome study of mouse dendritic cells to a depth of 5780 proteins. Cell subpopulations were obtained by FACS sorting, proteins were separated by 1D SDS-PAGE and digested with trypsin. Peptides from the gel pieces were analyzed on a nanoflow HPLC system connected to a hybrid LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific).

As a phosphoproteomics benchmark data set we took the raw data from 117 LC−MS runs produced in a phosphatase knockdown analysis.[54] Drosophila Schneider SL2 cells were differentially SILAC labeled as pairs with Lys-8/Arg-10 and Lys-0/Arg-0. Proteins were separated by 1D SDS-PAGE and digested with trypsin or in solution digested without gel separation. Peptides were subjected to TiO_2 chromatography and strong cation exchange chromatography and analyzed on a nanoflow HPLC system connected to a hybrid LTQ-Orbitrap (Thermo Fisher Scientific). For the analysis, we used only those MS/MS spectra that were acquired on a recognized SILAC pair. Modifications due to labeling with Lys-8 and Arg-10 can then be taken as fixed.

The benefits of second peptide analysis were investigated using data that was acquired on an LTQ-Orbitrap Velos. Briefly, HeLa cell lysate was in solution digested with trypsin, the peptide mixture was separated on a nanoflow HPLC system and analyzed using a data-dependent "top 10" method. Several runs were acquired with varying isolation windows. The precursor ions were isolated in selection windows of 1, 2, 4, 8, 16, and 32 Th followed by HCD fragmentation and high resolution data acquisition of the MS/MS spectra in the Orbitrap.

Data Preparation

MaxQuant, version 1.1.1.25, generated peak lists from the MS/MS spectra for the database searches. For the low-resolution MS/MS spectra recorded in "centroid" mode the 6 most abundant peaks per 100 Th mass intervals are kept for searching. High-resolution profile MS/MS data is deconvoluted (deisotoping and transfer of all fragment ions to single charge state) before extraction of the ten most abundant peaks per 100 Th. All statistical filters in MaxQuant like peptide and protein false discovery rates and mass deviation filters were disabled in order to score all submitted MS/MS spectra. Peptide masses were recalibrated by MaxQuant prior to both Andromeda and Mascot searches. For the Mascot search (using Mascot server version 2.2.04), peak lists written out by MaxQuant were converted to mgf format, the standard Matrix Science data format. Oxidation of methionine and N-terminal protein acetylation were used as variable modifications for all searches. A mass tolerance of 6 ppm was used for the peptide mass. To make Mascot and Andromeda searches comparable, we did not use the individual peptide mass tolerances in MaxQuant. A tolerance of 0.5 Th was used for matching fragment peaks produced by CID. The HCD fragment ion data used in the co-fragmentation study were searched with a 20 ppm window in Andromeda. A maximum of two missed cleavages were allowed in all searches. The "instrument" parameter was set to "ESI-TRAP" in the Mascot search. Mascot and Andromeda scores were matched to each other based on raw file name and scan number.

The search was performed against a concatenated target-decoy database with modified reversing of protein sequences as described previously.[31] Mouse and human data was searched against the respective IPI databases,[55] version 3.68, while the drosophila data was searched against protein sequences from flybase[56] version 5.24.

Search Engine Configuration

In Andromeda, the user specifies allowed peptide and protein modifications, enzymes used for protein cleavages and the protein sequence databases to be searched in the program AndromedaConfig.exe. Modifications are specified by their elemental composition. Neutral losses and diagnostic ions can be specified separately for each type of amino acid with the modification in question. Modifications that are interpreted as labels by MaxQuant can be defined here, such as SILAC labels. Searches with semispecific enzymes are supported as well, where

Table 1. Most Important Regular Expressions Defining How Protein Identifiers Are Extracted from the Headers of Fasta File Entries

regular expression	description
>(.*)	Everything after ">"
>([^])	Up to first space
>IPI:([^| .]*)	IPI accession
>(gi|[0−9]*)	NCBI accession
>([^\t]*)	Up to first tab character
>.*|(.*)|	Uniprot identifier

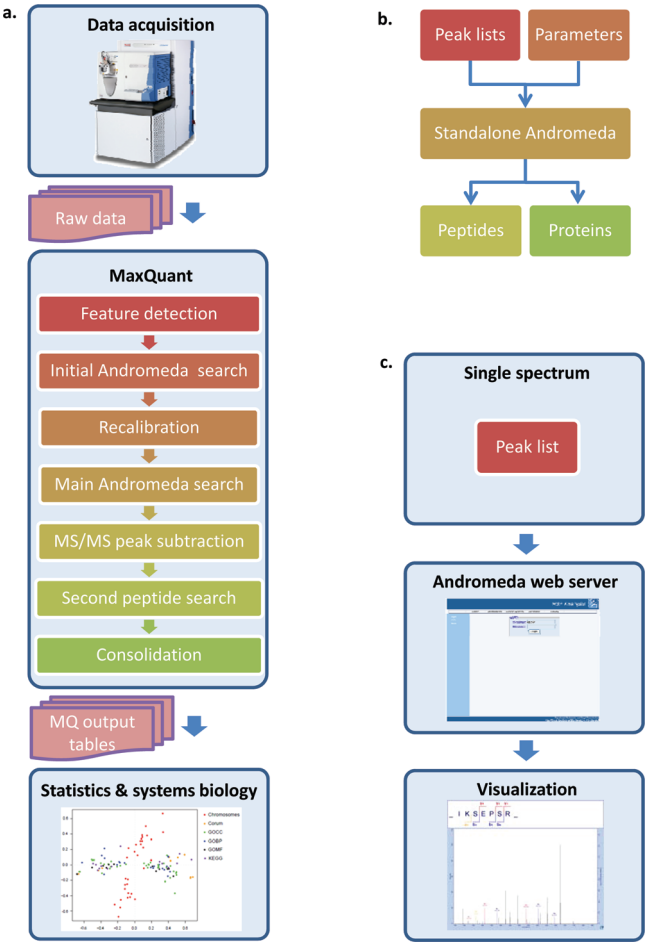


Figure 1. Three Andromeda configurations: (a) integrated in MaxQuant, (b) standalone search engine, and (c) web server.

only one peptide terminus needs to be a cleavage site according to the given protease digestion rule while the other terminus can be an arbitrary position in the protein. An unspecific search is also supported where both of the peptide termini can be arbitrary positions in a protein. Parse rules for regular expressions as defined in the Microsoft .NET framework (msdn.microsoft.com/en-us/library/az24scfc.aspx) are used to define how a protein identifier is extracted from the header line of a FASTA database file entry. Some of the most important regular expressions can be found in Table 1.

Input and Output Formats

Input files for peak lists and parameter values as well as output files for peptide identifications and a tentative protein list are all human-readable text files. Parameter files have the ending ".apar" and contain a list of key-value pairs where each pair is separated by a "=" sign. Expressions used for modifications, labels, enzymes and databases must have been defined previously in the AndromedaConfig.exe program. Peak list files have the extension ".apl" and can consist of arbitrarily many spectra, one following the other, each spectrum entry being enclosed by "peaklist start" and "peaklist end" lines. Some key-value pairs with peaklist-specific parameters are followed by two columns of numbers containing the m/z and intensity values. The peptide result files (".res") contain up to 15 candidate peptide matches for each peak list. For each candidate the peptide sequence, modification state, score, mass, mass deviation and all corresponding protein IDs are given.

Software Availability

MaxQuant with Andromeda as the integrated search engine can be downloaded from www.maxquant.org. A standalone version of Andromeda is available at www.andromeda-search.org. The source code is provided as Supporting Information 1. Both applications require Microsoft .NET 3.5, which is either already installed with Microsoft Windows or can be installed as a free Windows update. The Andromeda web server can be accessed at www.biochem.mpg.de/mann/tools/ for a limited number of submissions of MS/MS spectra. Andromeda has been written in the programming language C#, using the Microsoft .NET framework version 3.5.

RESULTS

Andromeda is a search engine based on a probability calculation for the scoring of peptide−spectrum matches. A version of it is fully integrated into the MaxQuant quantitative proteomics platform. Hence, all the data processing from the acquired raw data to the list of quantified peptides and proteins can be performed in a single end-to-end workflow (Figure 1a). In addition to the regular search Andromeda can be used in different contexts: for example in MaxQuant it is used for determining the mass-dependent recalibration function based on a preliminary database search, and for the identification of one or more cofragmented peptides (see below). We also provide a standalone version of Andromeda that produces scored peptide candidates, given a collection of MS/MS peak lists and a parameter file (Figure 1b). In this option, many of the statistical processing algorithms that are part of MaxQuant are not applied to the data and the reported list of identified proteins is only tentative without rigorous control of protein false discovery rate (FDR). The output consists of a raw list of scored peptide candidates per spectrum together with the protein list. Furthermore, there is a web server version of Andromeda for the submission of a limited set of spectra (Figure 1c), www.biochem.mpg.de/mann/tools/. In addition to the scoring results of the 15 best peptide candidates, the annotated spectrum can be inspected for the highest scoring and all other candidate peptide sequences. Despite these alternative uses, we anticipate that Andromeda will most commonly be employed as the search engine for MaxQuant.

Indexing Peptides and Proteins

To efficiently score an MS/MS spectrum it is important to be able to quickly retrieve all candidate peptides that have a suitable calculated precursor mass within a given tolerance. First we generate a list of all peptides obtained by the specified digestion

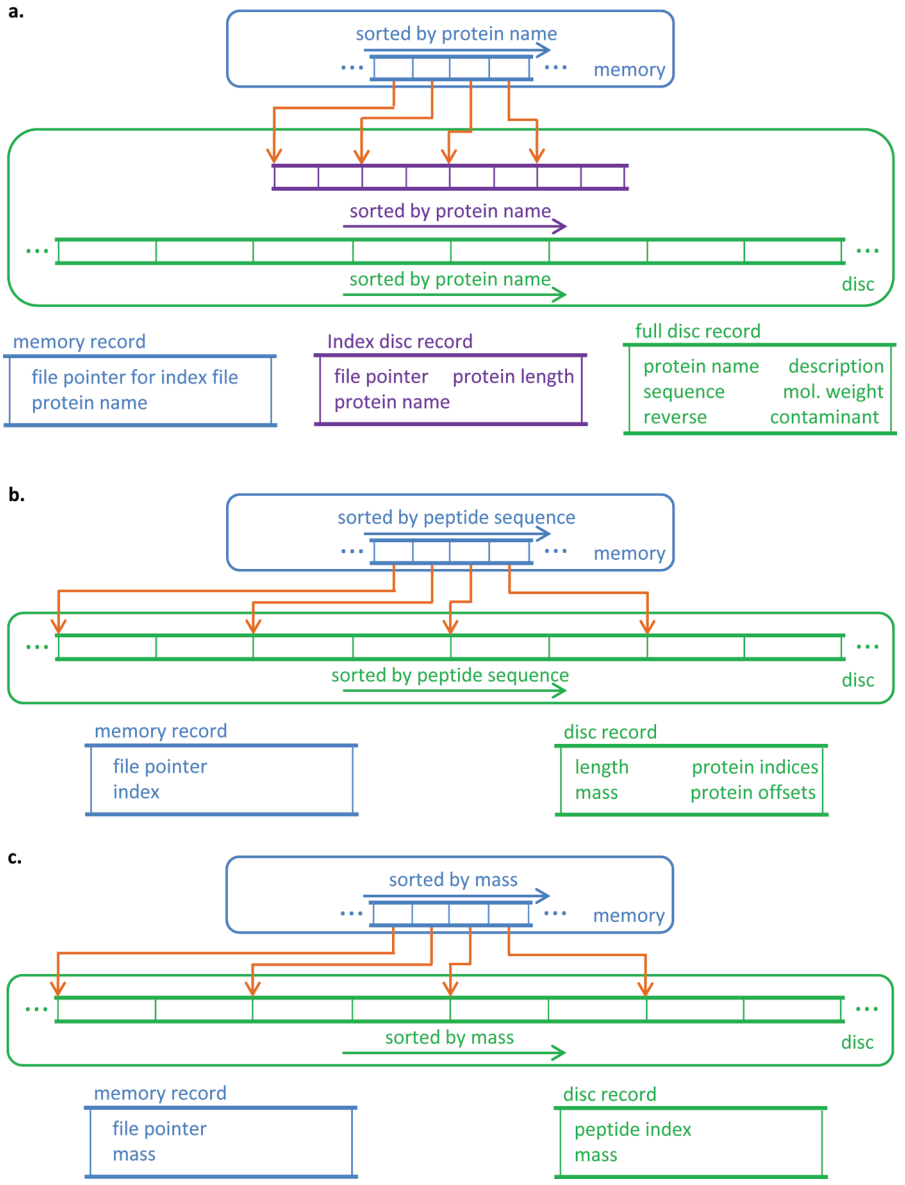


Figure 2. Memory and disk structure. (a) Protein list has a two-layer index structure. One small index is kept in memory whose entries point to blocks of multiple entries in the secondary index that is kept on disk. Each entry of the disk index points to the position of the protein entry in the file containing the complete information for each protein including the amino acid sequence. The protein lists are sorted alphabetically by the protein names. (b) Peptide index that resides in memory points to equally sized blocks of peptide entries, which are kept on disk. (c) Similar structure for the list of all combinations of peptide sequence and variable modifications. Index and disk entries are sorted by the peptide mass to allow for quick retrieval of all peptide candidates within a given mass interval.

rule from the protein sequences considering all possible combinations of preset variable modifications. At this stage we are only interested in the peptide masses, therefore only the number but not the positions of the modifications are important. The list of all of these peptides is sorted by mass for quick search access, which only grows slowly with increasing size (proportional to the log of the number of peptides for a binary search). The number of peptides with specific modifications can become very large, either when searching in an extended protein sequence database or by specifying many variable modifications. One common setting is to search the human IPI database including reverse sequences and common contaminants digested with trypsin and allowing for up to two missed cleavages. The number of modifications to consider can also grow rapidly. For example, in a phosphoproteomic experiment with triple SILAC labeling of lysine and arginine, one may simultaneously deal with phosphorylation of serine, threonine and tyrosine, Lys4, Lys8, Arg6, Arg10 and oxidation of methionine as variable modifications. (This is the case for those MS/MS spectra where the SILAC state could not be determined prior to the database search; otherwise the modification state of Arg and Lys are set by MaxQuant.) For the human IPI database and including the reversed sequences, this corresponds to a list of 174 618 protein sequences resulting in 7 837 653 peptide sequences and 76 937 183 modification-specific peptides (without taking modification positioning into account). These numbers can become even larger, for example in cases where one wants to search against a six-frame translation of the whole genome.

dx.doi.org/10.1021/pr101065j |J. Proteome Res. 2011, 10, 1794–1805

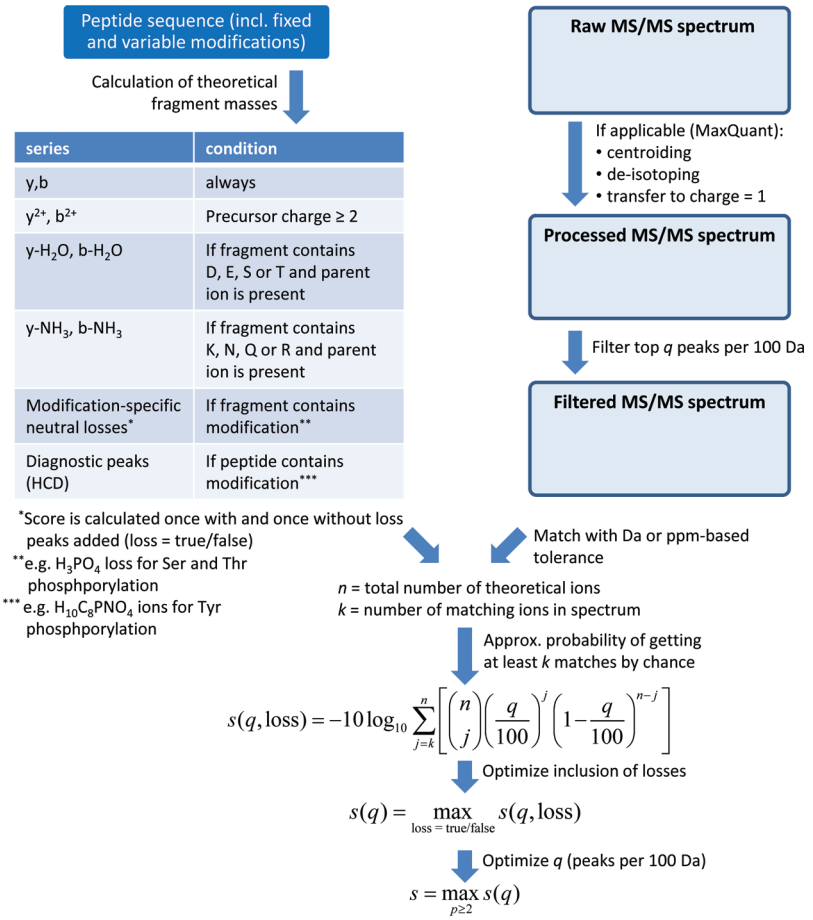


Figure 3. Schematic of the peptide scoring algorithm. The upper left branch shows the calculation of the theoretical fragment ion masses while the right branch indicates the processing of the experimental MS/MS spectra. In particular, all ion types that are used for the scoring can be found in the table on the left. The final score involves an optimization of the number of highest intensity peaks that are taken into account per 100 Da m/z interval and over the inclusion of modification-specific neutral losses.

We therefore wished to be able to handle protein sequence information without limitation on the sizes of calculated protein and peptide lists. Our goal was to work within the memory limits of 32-bit operating systems, which is around 1.6 GB from within the Microsoft .NET framework. The data structures for the search engine have to have an even smaller memory footprint since other data might be required to be in memory at the same time. Obviously the full modification-specific peptide list is too large to keep in memory and it has to reside on the hard disk (or solid state disk for improved performance). This is also true for the peptide and protein lists because unlimited scalability is desired. Only an index for each of the files is kept in memory, which contains positions of the records relative to the beginning of the file. These memory indices can already exceed the memory limitations for very large numbers of peptides. Therefore the index points to beginnings of blocks of elements in the file with a suitably chosen block size such that the lengths of the indices in memory never exceed a fixed size. In Figure 2, the structure of these lists and the relationships between memory and files residing on the hard disk are shown for proteins, peptides and modification-specific peptides. The records always contain indices to the respective items in the hierarchy above, assuring easy navigation from a candidate peptide to all the proteins that it occurs in. The modification-specific peptide list is the one that is directly accessed in database searches. It is sorted by mass, which allows quick retrieval of peptides within the given mass window. Protein and peptide list are instead sorted alphabetically by protein name and peptide sequence, respectively.

Scoring Model

The probabilistic score employed in Andromeda is derived from the p-score that was introduced for the identification of MS^3 spectra.[48] Given a peptide sequence together with a configuration of fixed and variable modifications for that peptide, first the theoretical fragment ions are calculated (Figure 3). For CID and HCD the list of theoretical fragment ion masses always contains the singly charged b- and y-ions. If the precursor charge is greater than one, the doubly charged b- and y-ions are added. In case of low resolution ion trap MS/MS spectra the charge state of fragments usually cannot be determined. The calculated doubly charged m/z values are then added explicitly if it is desired to match more highly charged fragments. For high-resolution MS/MS the charge state can be assigned to a fragment if more than one isotopic peak is detected. For these cases we remove peaks of fragments with charge higher than 1 from the spectrum and reintroduce them into the spectrum as singly charged fragment ions. If there are several charge states for a fragment their intensities are added, taking account of the fact that signal is proportional to charge in the Orbitrap analyzer. We noticed that even for high-resolution MS/MS data, where charge state

detection is possible in general, it is beneficial to consider doubly charged b- and y-ions as well. This is because for lower mass fragments sometimes only the monoisotopic peak is detectable precluding charge state determination and hence also the transformation to charge state one. For example assuming that the elemental composition of fragments follows the averagine model[57] the ratio between the ^{13}C and monoisotopic peak intensities for a fragment of 400 Da is 4.6:1. For less abundant fragments this can obviously lead to nondetection of the ^{13}C peak while the monoisotopic peak is above the noise level.

Calculated peaks corresponding to water and ammonia losses are only offered for matching as singly charged ions in those cases where the main b- and y-ion fragment is present and contains the amine-, amide- or hydroxyl-containing amino acid side-chains that tend to lead to the respective side chain loss. Modification-specific losses are configurable in the program AndromedaConfig, which is included in the MaxQuant distribution. The above-mentioned modification-specific neutral losses, as well as ions that are diagnostic for the presence of a particular modification of an amino acid type can be freely configured there. For example, the loss of phosphate from a phosphorylated serine or threonine is much more likely than from a tyrosine, which instead produces a highly specific immonium ion at mass 216.0426 (see, e.g., Steen et al.[44]). If Andromeda is used within MaxQuant, the report for each modification site includes presence or absence of a diagnostic peak in the MS/MS spectrum. The score is calculated once including configurable neutral losses and once excluding them and the maximum of the two scores is chosen. (Note that all scoring procedures are carried out identically for sequences from the reverse database, so they do not introduce a bias.)

The first step in the actual calculation of the score is to count the number of matches k between the n theoretical fragment masses and the peaks in the spectrum. The higher k is compared to n, the lower the chance that this happened by chance.[48] Because there are many signals in MS/MS spectra, including many low intense noise signals, the number of peaks in a defined mass interval—here 100 Th, which is the typical distance between consecutive members of fragment series (average mass of amino acids)—are limited to a maximum number. The parameter q is defined as the number of allowed peaks in the mass interval and it is needed to calculate the probability of a single random match. If the difference between calculated and measured masses is less than a predefined value, a match is counted. This can be done with an absolute mass tolerance window specified in Th or a relative mass window specified in ppm. While the former is appropriate for ion trap spectra, the latter is more suitable for high-resolution FT-ICR or Orbitrap spectra.

The Andromeda score is calculated as -10 times the logarithm of the probability of matching at least k out of the n theoretical masses by chance as shown in Figure 3. This is slightly different from Olsen et al.,[48] where the probability of matching exactly k out of n theoretical masses is determined. The formula used here is more similar to a definition of a p value for the null hypothesis that there is no similarity between the theoretical mass list and list of the spectrum masses. In particular, the score has the desirable property to vanish for $k = 0$. The calculation of the probability is only approximate since the probability for a single random match is taken to be $q/100$, which is exact if there was only one possible match per nominal mass. For high resolution MS/MS data the true random match probability is considerably less than this and the true score would be higher but more complicated to calculate. However, this simplification is conservative as it decreases the calculated score and is justified by the excellent performance of the search algorithm on high-accuracy MS/MS data.

The intensities of the peaks in the MS/MS spectra are indirectly taken into account by calculating the score for all values for q (number of peaks per 100 Th) up to the specified maximum. The best of these scores for varying q is selected. Therefore two spectrum-sequence comparisons with the same values for n and k can result in different scores depending on the intensities of the matched peaks. Generally, the score is higher if the matches are among the more intense peaks because the optimal value of q will be lower (see formula in Figure 3). However, we have found it crucial that this intensity weighting is not done on the overall intensity scale over the whole spectrum, but that it is restricted to local mass regions (e.g., the 100 Th mass range intervals.). This compensates for underlying global peak density distributions which typically favor small fragment masses.

The inclusion of additional information like peptide length, number of modifications or of missed cleavages can aid the specificity of peptide assignments to spectra. Ideally this is done in a data-dependent manner in which different weights for different classes of peptides can be derived from the data by machine learning in a Bayesian framework. We wished to include such a weighting of peptide classes into the score while retaining a basic search engine score that is deterministic and only depends on the spectrum being scored rather than the ensemble of all other spectra. To capture the dependence of the score on peptide mass and on the number of modifications we introduced a fixed additive component to the Andromeda score, which depends on the number of modifications and is a linear function of the mass. The specific values are determined in a manner that adjusts the distributions of reverse hits from a target-decoy search so that they become equal. The net effect of this procedure is to minimize the FDR for a given cutoff value, because it does not depend on peptide mass and modification state any longer. We used a large data set of MS/MS spectra and incorporated the specific weights into the scoring function. A data-dependent Bayesian scoring can still be applied to the output of the Andromeda search engine. For instance, MaxQuant additionally performs a peptide length dependent Bayesian analysis in a data dependent manner.[31]

Comparison to the Mascot Search Engine

Mascot[36] is a widely used standard for database searching and most other search engines have been compared to Mascot. Therefore we investigated how Andromeda compares to Mascot in terms of scoring of peptide-spectrum matches. As the exact details of the Mascot scoring system are not known, we compared the performance of Andromeda vs Mascot empirically on very large sets of proteomic data.

In Figure 4a, we plot the Mascot score against the Andromeda score for a data set of 732 287 MS/MS spectra derived from a label-free mouse proteome measurement as described in Materials and Methods. For each MS/MS spectrum the highest scoring peptide is taken which is not necessarily the same for the Mascot and the Andromeda scoring. In Figure 4b, the fraction of cases for which the top-scoring Andromeda and Mascot peptide sequences coincide is displayed as a histogram depending on the Andromeda score. As can be seen, above an Andromeda score of 100 the top-scoring peptides coincide in almost all cases.

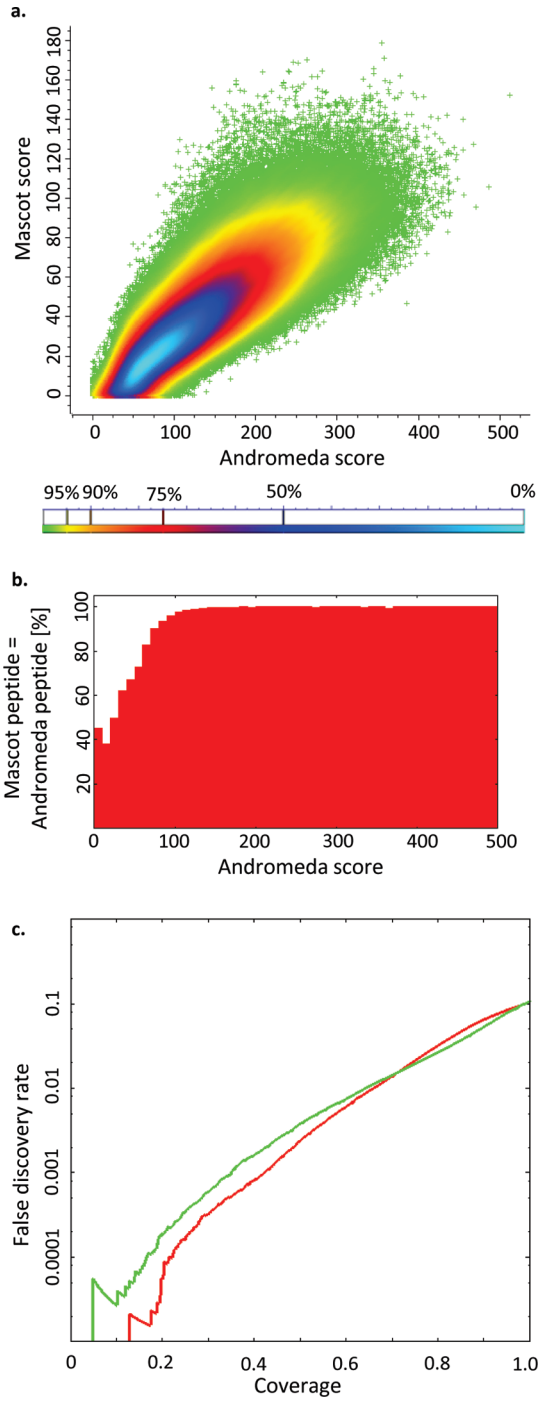


Figure 4. (a) Andromeda vs Mascot score for a data set of 732 287 MS/MS spectra derived from a label-free mouse proteome measurement.[53] The score for the top-scoring peptide for each MS/MS spectrum is shown which is not necessarily the same peptide sequence for the Mascot and the Andromeda identification. The color code indicates the percentage of points that are included a region of a specific color. (b) Histogram of the percentage of cases in which the top-scoring Andromeda and Mascot peptide sequences are equal as a function of Andromeda score. For the comparison leucine and isoleucine were treated as the same amino acid. (c) False discovery rate as a function of coverage for the same data set calculated based on the reverse hits from the target-decoy search.

Of the recorded MS/MS spectra, 89.1% correspond to unmodified peptides and most of the identified modified peptides have an oxidized methionine. The point density is indicated by the color code in Figure 4a which encodes the percentage of points that are included a region of a specific color. For example, the yellow line in Figure 4a encloses 95% of all data points. This visualization allows the visual detection of outliers (like a two-dimensional data plot), while at the same time retaining information about the density of points that would normally only be visible in a 3D data plot. It is immediately apparent from the figure that the scores correlate well overall. There are no distinct populations of peptides that are only identified by one of the search engines. A linear regression results in the equation $M = 0.311 * A - 32.231$, where M is the Mascot score and A the Andromeda score, with an R^2 value of 0.708. This indicates that Andromeda scores are generally about 3-fold larger than Mascot scores. However, this does not indicate a 3-fold larger confidence. The statistical power is better determined by calculating coverage and false discovery rates as a function of score threshold as is done below. A rough conversion between Andromeda and Mascot scores can be performed by a division by three or application of the regression line. Note that there are only very few and dispersed outliers on either side; of the order of tens of spectra out of the total of more than 700 000. Furthermore, there are virtually no high-scoring outliers near either axis, indicating an absence of spectra that were ranked highly with one method but scored close to zero with the other. This demonstrates that no populations of peptides would be lost entirely by employing one score or the other.

Next we compare the performance of the Andromeda and Mascot search engines as a function of False Discovery Rates estimated as the number of hits from the reverse database divided by the number of forward hits at any given minimum score. The sensitivity of the database search is defined as the number of accepted forward hits relative to the total number of forward hits at the same score. Mascot and Andromeda have very similar characteristics over the whole range of FDRs, in particular including the often used 1% FDR rate (Figure 4b). This shows that the two scores are very close in discriminatory power.

Scoring of Phosphopeptides

Figure 5a shows the same type of plot as in Figure 4a but for a data set that is enriched for phosphopeptides. Of the recorded 586 883 MS/MS spectra in Figure 5a, 27.4% have one or more phosphorylations. Outliers are visible in the region of high Andromeda and low Mascot score and most of them correspond to peptides with three to five phosphorylation events. Figure 5b displays the MS/MS spectrum of a peptide with five phosphorylation sites that has a Mascot score of 5.2 and an Andromeda score of 199.3. The y-series coverage is almost complete with most fragments occurring with a neutral loss of a phosphate molecule. An FDR coverage curve for the phosphopeptide data set is depicted in Figure 5c. The performances of Mascot and Andromeda are similar over the entire range with an advantage for Andromeda in the high specificity region. At the typical operation point of 1% FDR results are very close. We speculate that the better scoring in the region of higher specificity may be due to a better matching of spectra of phosphopeptides in Andromeda due to more comprehensive combinatorics of positioning of phospho-groups on the available serine, threonine and tyrosine sites in the peptide sequences, including a more complete offering of neutral losses. During the Andromeda search we offer up to 1000 positionings of variable modifications within any given peptide which is exhaustive for most situations.

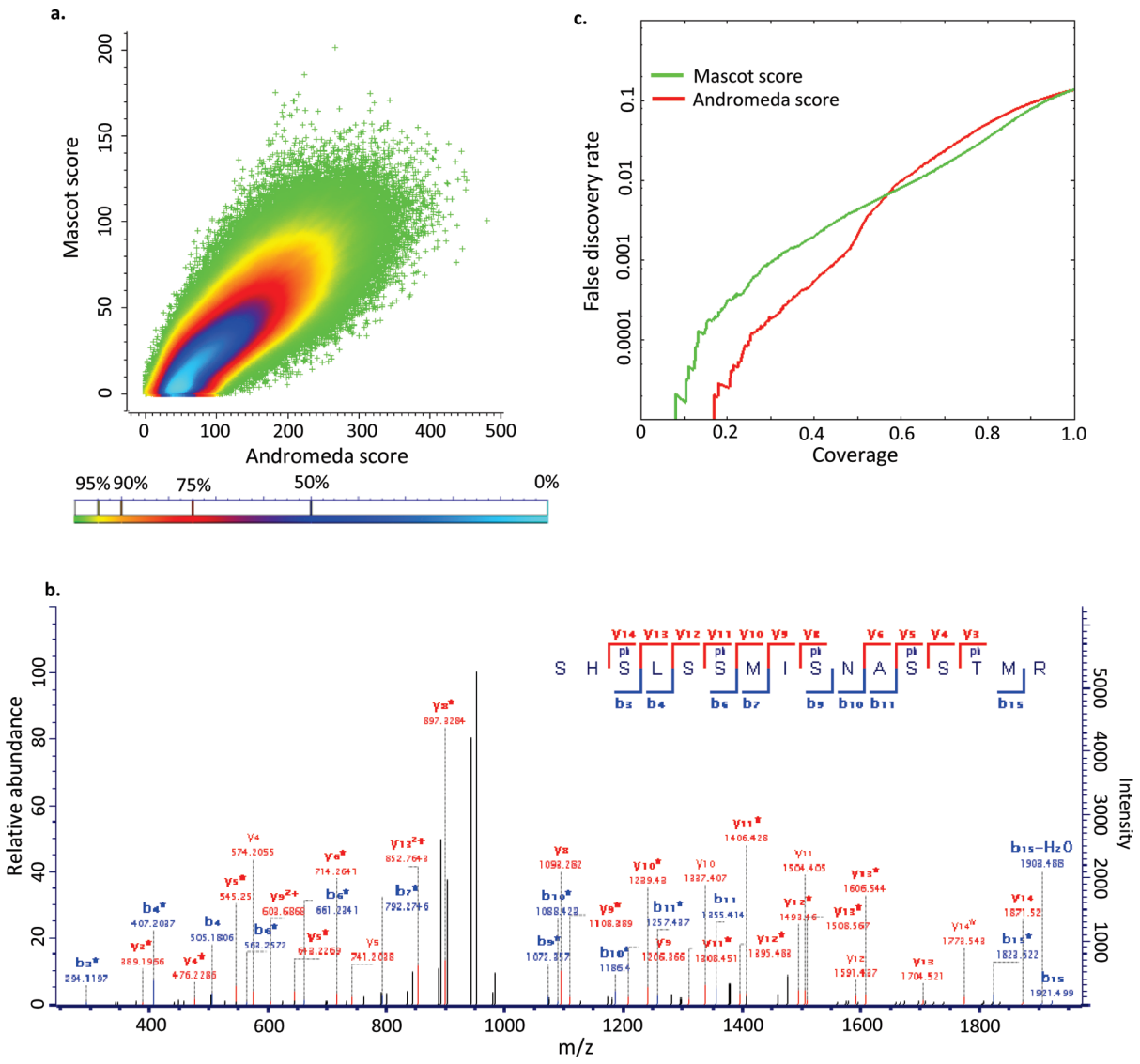


Figure 5. (a) Andromeda vs Mascot score for 586,883 MS/MS spectra from the phospho-proteome data by Hilger et al.[54] (b) Annotated MS/MS spectrum of the peptide SHpSLSpSMIpSNApSSpTMR. Mascot and Andromeda produce the same top-scoring peptide sequence with a Mascot score of 5.2 and an Andromeda score of 199.3. (c) False discovery rate as a function of coverage for the same data set calculated in the same way as in Figure 4c.

In MaxQuant, the top-scoring peptide is furthermore rescored with essentially exhaustive positioning of modifications. We merely restrict the combinatorics to 100 000 possibilities to exclude the rare instances where single peptides cause long calculation times due to "combinatorial explosion". In Supplementary Figure 1 (Supporting Information), the same data as in Figure 5a is shown six times—each time highlighting another population of top-scoring peptides with a fixed number of phosphorylations. Peptides with higher phosphorylations tend to have many data points in the high Andromeda score but low-to-moderate Mascot score region further indicating that Andromeda performs better on highly phosphorylated peptides.

Identification of Second Peptides

Even in high-resolution MS, the selection of the precursor ion for fragmentation is always performed with low resolution (typically a few Th) to ensure adequate sensitivity for MS/MS. In complex mixtures, this results in frequent cofragmentation of coeluting peptides with similar masses. These 'chimerical' MS/MS spectra[52] can be detrimental for identification of the peptide of interest, especially if the cofragmented peptide is of comparable intensity. Co-fragmentation generally reduces the number of peptides identified in database searches and poses special problems for reporter fragment based quantification methods because both peptides contribute to the measured ratios.

However, this situation can be turned to an advantage if both peptides can be identified. In particular, this presents the opportunity to identify peptides that have not been targeted for MS/MS and to obtain two or more peptide identifications from a single MS/MS spectrum. Although this problem has been addressed before,[49−52] to our knowledge it has not been adopted in mainstream search engines yet. Here we describe a second peptide identification algorithm that we have integrated into the Andromeda/MaxQuant workflow.

To illustrate the principles of our algorithm, Figure 6a shows an LC−MS map, where 3D peaks are indicated as lines marking the peak boundaries. The blue isotope pattern has been selected for fragmentation at the position of the cross on the

dx.doi.org/10.1021/pr101065j |J. Proteome Res. 2011, 10, 1794–1805

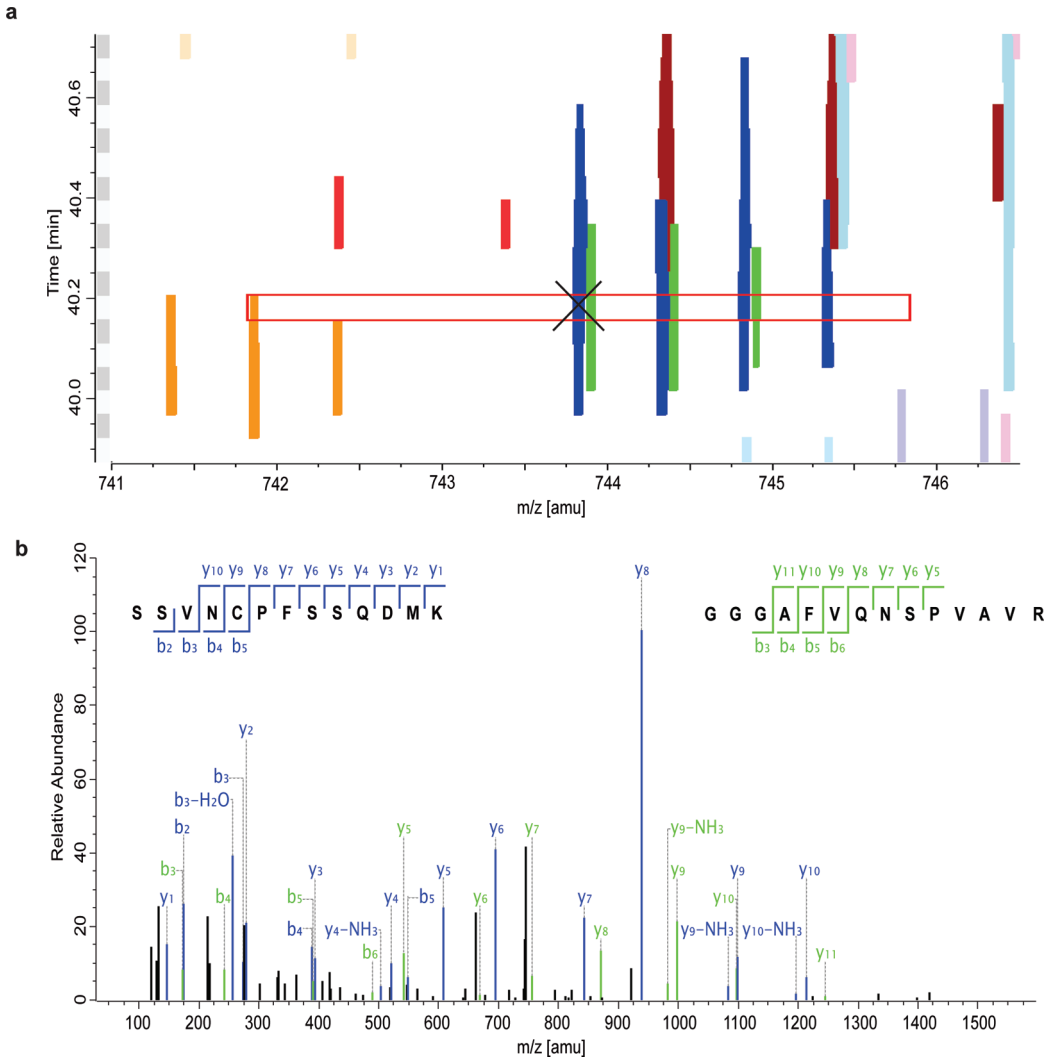


Figure 6. Second peptide identification. (a) LC−MS map of the sequenced (blue) and cofragmented (green) peptide described in the main text. The blue peptide has been selected for fragmentation at the position of the cross. The red rectangle indicates the isolation window. (b) MS/MS spectrum leading to the identification of both peptides. Fragments of the two peptides are indicated in blue and green, respectively. The blue peptide is identified in the conventional database search while the green peptide has been identified as "second peptide".

monoisotopic peak of that peptide. The red rectangle indicates the region from which ions have been isolated for fragmentation. Clearly the peptide corresponding to the green isotope pattern that has not been selected for sequencing intersects with the isolation rectangle. Therefore its fragments should be present in the MS/MS spectrum as well. The actual fragment spectrum is shown in Figure 6b where the fragments originating from the targeted and identified peptide (blue isotope pattern) are indicated in blue. This process is repeated for the entire LC−MS/MS run. For every 3D MS isotope pattern that has not been selected for sequencing the algorithm checks whether it intersects with the isolation window of any MS/MS spectrum. If this is the case then the fragments in this MS/MS spectrum that have already been assigned to a peptide sequence during the main Andromeda search are subtracted. The remaining fragments are submitted to a new database search with the precursor mass from the peptide that was not targeted for MS/MS. The collection of these "subtracted" peak lists is submitted to Andromeda in the same way as in a conventional search. However, the results of this second peptide search are further processed with their own peptide length based posterior error probability and precursor mass filtering. Since these spectra are on average of lower quality than the original MS/MS spectra we have found it to be crucial that they have their own data-dependent statistical model for peptide identification. The resulting peptides are then accepted up to a 1% FDR and integrated into the usual protein identification and quantification workflow.

The HCD data set used for testing (see Materials and Methods) was acquired with a total isolation width of 4 Th for every MS/MS spectrum. The identification rate for the set of second peptide spectra is much lower compared to the normal MS/MS identification rate of 50%. Nevertheless, since the number of the second peptide spectra is quite high compared to normal MS/MS spectra considering cofragmentation still leads to a considerable increase in peptide identifications. In our example, the number of identified peptide features increased by 10.7% by the inclusion of second peptide identifications. The gain in the number of identified peptides depends on the isolation width for the acquisition of MS/MS spectra. For instance, at an isolation width of 2 Th we observe that the increase

dx.doi.org/10.1021/pr101065j |J. Proteome Res. 2011, 10, 1794–1805

in identified peptides through second peptide identifications is only 5.7%. The relative gain is larger at increased isolation width because the average number of additional peptides within the window increases. However, the chance to identify the main peptide decreases due to the mixing of the spectrum with fragments from other peptides. The dependence of the number of peptide identifications for conventional and second peptides is shown in Supplementary Figure 2 (Supporting Information).

DISCUSSION AND OUTLOOK

Here we have described Andromeda, a novel search engine for matching MS/MS spectra to peptide sequences in a database. Andromeda can either be used in a stand-alone mode or—more typically—as part of the MaxQuant environment. Apart from an optimal scoring model our intention was to develop a very robust architecture with unlimited scalability. We have demonstrated this on large scale data sets with hundreds of thousands of spectra. Andromeda has been "stress tested" in ongoing studies and has been the default search engine in our laboratory for some time. A practical advantage of the MaxQuant/Andromeda combination is that it runs locally on the user's computer. This eliminates client-server set up and communication issues. The computational proteomics pipeline starting from raw data files to reported protein groups and their quantitative ratios now appears unified to the user. Despite the local search architecture, processing speeds are generally not different from the previous MaxQuant/Mascot environment in which Mascot was run on an external server. Furthermore, we have added a separate module called Perseus (www.maxquant.org), which performs bioinformatic analysis of the output of the MaxQuant/Andromeda workflow. Perseus is already available and in use[58] and completes the pipeline for computational proteomics analysis but will be described in a future publication (Cox et al., in preparation).

The scoring function at the heart of Andromeda is built on a simple binominal distribution probability formula (Figure 3), which we have previously used in scoring MS^3 spectra and localizing PTMs.[59] Andromeda divides the MS/MS spectrum into mass ranges of 100 Th. In each of these ranges the number of experimental peaks offered for matching is dynamically tested in an intensity prioritized manner.

False discovery rates for the same initial probability score can still depend on the number of modifications and on the mass of the peptide. This is accounted for in Andromeda by an additive component to the score. Comparison to Mascot on very large data sets reveals very few outliers—in particular almost no peptides are exclusively identified by one of the two search engines. Furthermore, the coverage of identified peptides at any given FDR is likewise similar, including at the generally used operating point of 1% expected false positives. We did notice improved identification of heavily modified peptides in Andromeda compared to Mascot, which we attribute to the more exhaustive combinatorial analysis of placing PTMs on all possible amino acids. As the Mascot search engine has become one of the standards in proteomics, equivalent performance fulfills the goal that we had set for the development of Andromeda and likely implies favorable comparison to other search engines as well. Apart from describing the score we have also made the actual code used in Andromeda available for inspection with this publication (Supporting Information 1).

A key advantage of Andromeda is its extensibility. For example, proteomics with high accuracy MS and MS/MS data (high−high mode[60]), is becoming increasingly common. Andromeda, in contrast to Mascot, allows arbitrarily accurate MS/MS requirements specified in ppm. Similarly, Mascot precludes identification of SILAC pairs if the same amino acid can bear a fixed and a variable modification. This causes a substantial loss of quantification information, for example in the analysis of lysine acetylated peptides[61] because all MS/MS spectra of lysine-acetylated peptides that were sequenced on the heavy SILAC partner will not be identified by Mascot. All these quantitative ratios are retrieved in the MaxQuant/Andromeda workflow.

More generally, additional scoring modes can be added to Andromeda. We demonstrated this by implementing a second peptide identification algorithm into the MaxQuant/Andromeda workflow. For each isotope cluster that is detected in the LC−MS data but that was not targeted for fragmentation the algorithms checks if the precursor isotope pattern intersects the selection window of any MS/MS event. If so, fragment ions belonging to the identified peptide are subtracted and the search is repeated with the cofragmented peptide in a statistically rigorous way. As demonstrated here, this leads to an appreciable increase in peptide and protein identifications in complex mixtures. As another example, special algorithms are necessary for peptide identification in data independent MS/MS where the whole mass range is fragmented.[62,63] Using the MaxQuant/Andromeda infrastructure our group recently developed an implementation of this principle on the Exactive instrument, which consists only of an Orbitrap analyzer with HCD capability.[64]

In conclusion, we have developed, described and tested a robust and scalable search engine that in combination with MaxQuant represents a powerful and unified analysis pipeline for quantitative proteomics, which is freely available to the community.

ASSOCIATED CONTENT

Supporting Information

Supplemental figures and materials. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*J.C. e-mail cox@biochem.mpg.de, fax +49 (89) 8578 3209, phone +49 (89) 8578 2088 or M.M. e-mail mmann@biochem.mpg.de, fax +49 (89) 8578 3209, phone +49 (89) 8578 2557.

ACKNOWLEDGMENT

We thank other members of our groups in Martinsried and Copenhagen for fruitful discussion and help. We also thank early adopters of Andromeda in other laboratories for helpful feedback. This work was partially supported by the European Union seventh Framework Program (HEALTH-F4-2008-201648/PROSPECTS).

REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.

(2) Yates, J. R., 3rd; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6* (9), 702–14.

(3) Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science* **2006**, *312* (5771), 212–7.

(4) Cox, J.; Mann, M. Is proteomics the new genomics? *Cell* **2007** *130* (3), 395–8.

(5) Vermeulen, M.; Selbach, M. Quantitative proteomics: a tool to assess cell differentiation. *Curr. Opin. Cell Biol.* **2009**, *21* (6), 761–6.

(6) Choudhary, C.; Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell. Biol.* **2010**, *11* (6), 427–39.

(7) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–82.

(8) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–7.

(9) Nesvizhskii, A. I.; Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today* **2004**, *9* (4), 173–81.

(10) Listgarten, J.; Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (4), 419–34.

(11) Chalkley, R. J.; Hansen, K. C.; Baldwin, M. A. Bioinformatic methods to exploit mass spectrometric data for proteomic applications. *Methods Enzymol.* **2005**, *402*, 289–312.

(12) Colinge, J.; Bennett, K. L. Introduction to computational proteomics. *PLoS Comput. Biol.* **2007**, *3* (7), e114.

(13) Matthiesen, R. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics* **2007**, *7* (16), 2815–32.

(14) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4* (10), 787–97.

(15) Deutsch, E. W.; Lam, H.; Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **2008**, *33* (1), 18–25.

(16) Mueller, L. N.; Brusniak, M. Y.; Mani, D. R.; Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* **2008**, *7* (1), 51–61.

(17) Matthiesen, R.; Jensen, O. N. Analysis of mass spectrometry data in proteomics. *Methods Mol. Biol.* **2008**, *453*, 105–22.

(18) Bandeira, N.; Clauser, K. R.; Pevzner, P. A. Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell. Proteomics* **2007**, *6* (7), 1123–34.

(19) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. Clustering millions of tandem mass spectra. *J. Proteome Res.* **2008**, *7* (1), 113–22.

(20) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **2008**, *7* (1), 286–92.

(21) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 47–50.

(22) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7* (1), 245–53.

(23) Rauch, A.; Bellew, M.; Eng, J.; Fitzgibbon, M.; Holzman, T.; Hussey, P.; Igra, M.; Maclean, B.; Lin, C. W.; Detter, A.; Fang, R.; Faca, V.; Gafken, P.; Zhang, H.; Whiteaker, J.; States, D.; Hanash, S.; Paulovich, A.; McIntosh, M. W. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **2006**, *5* (1), 112–21.

(24) Rinner, O.; Mueller, L. N.; Hubalek, M.; Muller, M.; Gstaiger, M.; Aebersold, R. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat. Biotechnol.* **2007**, *25* (3), 345–52.

(25) Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R., 3rd A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **2008**, *5* (4), 319–22.

(26) Brusniak, M. Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L. N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. D. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinform.* **2008**, *9*, 542.

(27) May, D.; Law, W.; Fitzgibbon, M.; Fang, Q.; McIntosh, M. Software platform for rapidly creating computational tools for mass spectrometry-based proteomics. *J. Proteome Res.* **2009**, *8* (6), 3212–7.

(28) Deutsch, E. W.; Shteynberg, D.; Lam, H.; Sun, Z.; Eng, J. K.; Carapito, C.; von Haller, P. D.; Tasman, N.; Mendoza, L.; Farrah, T.; Aebersold, R. Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation data sets. *Proteomics* **2010**, *10* (6), 1190–5.

(29) Mortensen, P.; Gouw, J. W.; Olsen, J. V.; Ong, S. E.; Rigbolt, K. T.; Bunkenborg, J.; Cox, J.; Foster, L. J.; Heck, A. J.; Blagoev, B.; Andersen, J. S.; Mann, M. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.* **2010**, *9* (1), 393–403.

(30) Kumar, C.; Mann, M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* **2009**, *583* (11), 1703–12.

(31) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–72.

(32) Cox, J.; Mann, M. Computational Principles of Determining and Improving Mass Precision and Accuracy for Proteome Measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1477–85.

(33) Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7* (12), 952–8.

(34) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1* (5), 376–86.

(35) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455* (7217), 1251–4.

(36) 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), 3551–67.

(37) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–89.

(38) Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **1999**, *71* (14), 2871–82.

(39) Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2* (10), 1406–12.

(40) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–7.

(41) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958–64.

(42) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W701–6.

(43) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally

modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–39.

(44) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.* **2004**, *5* (9), 699–711.

(45) Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195–202.

(46) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390–9.

(47) 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* (3−4), 327–42.

(48) Olsen, J. V.; Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (37), 13417–22.

(49) Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005**, *5* (16), 4096–106.

(50) Bern, M.; Finney, G.; Hoopmann, M. R.; Merrihew, G.; Toth, M. J.; MacCoss, M. J. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **2010**, *82* (3), 833–41.

(51) Wang, J.; Perez-Santiago, J.; Katz, J. E.; Mallick, P.; Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* **2010**, *9* (7), 1476–85.

(52) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152–60.

(53) Luber, C. A.; Cox, J.; Lauterbach, H.; Fancke, B.; Selbach, M.; Tschopp, J.; Akira, S.; Wiegand, M.; Hochrein, H.; O'Keeffe, M.; Mann, M. Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **2010**, *32* (2), 279–89.

(54) Hilger, M.; Bonaldi, T.; Gnad, F.; Mann, M. Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Mol. Cell. Proteomics* **2009**, *8* (8), 1908–20.

(55) 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* (7), 1985–8.

(56) Tweedie, S.; Ashburner, M.; Falls, K.; Leyland, P.; McQuilton, P.; Marygold, S.; Millburn, G.; Osumi-Sutherland, D.; Schroeder, A.; Seal, R.; Zhang, H. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.* **2009**, *37* (Database issue), D555–9.

(57) Senko, M. W.; Beru, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229–233.

(58) Geiger, T.; Cox, J.; Mann, M. Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet.* **2010**, *6* (9), e1001090.

(59) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, In Vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635–48.

(60) Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (47), 18132–8.

(61) Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **2009**, *325* (5942), 834–40.

(62) Geromanos, S. J.; Vissers, J. P.; Silva, J. C.; Dorschel, C. A.; Li, G. Z.; Gorenstein, M. V.; Bateman, R. H.; Langridge, J. I. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **2009**, *9* (6), 1683–95.

(63) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77* (7), 2187–200.

(64) Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **2010**, *9* (10), 2252–61.