A Framework for Intelligent Data Acquisition and Real-Time Database Searching for Shotgun Proteomics*

Johannes Graumann‡‖, Richard A. Scheltema‡**, Yong Zhang‡¶, Jürgen Cox‡, and Matthias Mann‡§**

In the analysis of complex peptide mixtures by MS-based proteomics, many more peptides elute at any given time than can be identified and quantified by the mass spectrometer. This makes it desirable to optimally allocate peptide sequencing and narrow mass range quantification events. In computer science, intelligent agents are frequently used to make autonomous decisions in complex environments. Here we develop and describe a framework for intelligent data acquisition and real-time database searching and showcase selected examples. The intelligent agent is implemented in the MaxQuant computational proteomics environment, termed MaxQuant Real-Time. It analyzes data as it is acquired on the mass spectrometer, constructs isotope patterns and SILAC pair information as well as controls MS and tandem MS events based on real-time and prior MS data or external knowledge. Re-implementing a top10 method in the intelligent agent yields similar performance to the data dependent methods running on the mass spectrometer itself. We demonstrate the capabilities of MaxQuant Real-Time by creating a real-time search engine capable of identifying peptides "on-the-fly" within 30 ms, well within the time constraints of a shotgun fragmentation "topN" method. The agent can focus sequencing events onto peptides of specific interest, such as those originating from a specific gene ontology (GO) term, or peptides that are likely modified versions of already identified peptides. Finally, we demonstrate enhanced quantification of SILAC pairs whose ratios were poorly defined in survey spectra. MaxQuant Real-Time is flexible and can be applied to a large number of scenarios that would benefit from intelligent, directed data acquisition. Our framework should be especially useful for new instrument types, such as the quadrupole-Orbitrap, that are currently becoming available. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.013185, 1–11, 2012.***

Mass spectrometry-based proteomics is generally performed in a shotgun format, where the proteome of interest is digested by a sequence specific protease and resulting peptides are analyzed by on-line liquid chromatography tandem mass spectrometry (LC MS/MS)[1] (1–4). Complex protein mixtures can contain thousands of proteins and an even much larger number of peptides are generated by the enzymatic digestion. As a result many peptides elute at a given time during chromatographic separation and the mass spectrometer needs to schedule peptide fragmentation events based on the peptide mass and intensity information in the MS spectra ("data dependent acquisition"). A widely used acquisition scheme is a topN method in which the mass spectrometer continuously cycles through full MS scans that are each followed by up to N precursor isolation and fragmentation events (MS/MS scans). In complex mixtures MS scans contain many more precursors than can be fragmented in the available time before the next full scan (5). To select precursors for fragmentation, the manufacturer's software controlling the instrument sorts the precursor ions detected in each MS scan by intensity and also applies certain filters such as minimum signal intensity, charge state, and avoidance of already fragmented precursors. Inclusion and exclusion lists containing peptide masses of interest or those deemed uninteresting can also be used (6–8). For maximizing identification success in shotgun proteomics, more elaborate selection schemes have been developed. For example, the Coon group has implemented a decision tree algorithm that schedules precursors for fragmentation either by collision-induced dissociation (CID) or by electron transfer dissociation based on their mass and charge (9).

In computer science "intelligent software agents" are constructed that make decisions in complex environments (10). Examples of intelligent agents are "spiders" that scour the web for search engines or software controlling mobile robots.

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

⌘ *Author's Choice*—Final version full access.

Received August 1, 2011, and in revised form, December 11, 2011

Published, MCP Papers in Press, December 14, 2011, DOI 10.1074/mcp.M111.013185

[1] The abbreviations used are: LC-MS/MS, liquid chromatography-tandem mass spectrometry; CID, collision induced dissociation; HCD, higher energy dissociation; SILAC, stable isotope labeling with amino acids in cell culture; FDR, false discovery rate; LTQ, linear trap quadrupole; MS/MS, tandem mass spectrometry; OCX, Object Linking and Embedding Control eXtension; SIM, selected ion monitoring.

Here we set out to conceptualize and construct an intelligent agent framework for proteomics and evaluate its applications to selected examples. The agent was implemented in the MaxQuant computational proteomics environment, which performs feature detection in raw MS data files, extracts high accuracy mass and quantification values, searches MS/MS data, and includes downstream bioinformatic analysis tools (11–13). Unlike MaxQuant, this agent—termed "MaxQuant Real-Time"—reconstructs isotope patterns and stable isotope labeling with amino acid in cell culture (SILAC) patterns based on incomplete information from incoming MS scans and makes decisions about data acquisition within the chromatographic time scale. Communication with the Orbitrap instruments was provided by a special software module made available by Thermo Fisher Scientific ("instrument Object Linking and Embedding Control eXtension (OCX)" for Microsoft Object Linking and Embedding Control eXtension) (14). We demonstrate the capabilities of MaxQuant Real-Time in proof-of-principle applications to several generic and long-standing challenges in peptide based proteomics. These include implementation of the first real-time peptide search engine as well as targeted quantification of SILAC pairs whose ratios could not be unambiguously determined from preceding scans or runs.

EXPERIMENTAL PROCEDURES

Construction of the Intelligent Agent—The intelligent agent makes use of the instrument OCX library (Thermo), which allows programmatic access to the mass spectrometer. The agent provides much of Xcalibur's functionality, while offering flexibility for novel acquisition methodology and advanced real-time analysis capabilities. It triggers scans, retrieves acquired spectral information, feeds this information to analysis software, and proceeds with acquisition in a data-dependent manner. The analysis of spectral data uses the implemented algorithms from the MaxQuant computational proteomics environment for centroiding the individual spectra, construction of three-dimensional features, as well as the construction of isotope clusters and SILAC partners. However, these algorithms have been adapted to deal with the partial chromatographic peak data while the full scan data is gradually becoming available.

For the correct operation of the machine a large number of parameters are required, which are stored in an XML-file read at the start of the measurement. This makes the software applicable to different measurement modes and machine types. At the initialization stage of the measurement the tune file is loaded, and the following parameters are passed to the instrument control software: acquisition file, acquisition time, polarity, automatic gain control settings (ion target values, injection waveforms and machine based automatic gain control), and communication with the high-performance liquid chromatography (contact closure). For each scan definition pre- and post-settings are applied. The presettings consist of the mass-range, spectrum averaging, and the mass analyzer type with the resolution if applicable. The postsettings consist of the number of micro-scans (always 1 in our experiments) and the data format (centroid or profile). For both the Full and selected ion monitoring (SIM) scan definition the machine is instructed to do single MS and the maximum inject time is defined. For the MS2 scan definition the machine is instructed to do MS/MS together with parameters specifying the isolation window, the reaction type (CID, HCD, or ETD), collision energy and the maximum injection time. For CID, the minimum m/z is calculated based on the one-third cutoff rule as described in the Thermo LTQ Orbitrap Velos manual and the maximum to the single charged m/z of the precursor ion. For higher energy collisional dissociation, the minimum m/z is set fixed to 80 Th and the maximum calculated as for CID.

With the intelligent agent framework, we now also have the possibility to write out large amounts of meta-information. For example, the uncorrected precursor mass information is much more accurate than that provided by Xcalibur (only specified to two decimal places by the Xcalibur software), which makes the link back to the correct precursor in post-processing robust. This is especially useful when retrieving the isotope patterns of specific peptides of interest (*e.g.* with targeting).

Sample Preparation—HeLa cells were SILAC-labeled with arginine 10 ($^{13}C_6H_{14}{}^{15}N_4O_2$; Sigma) and lysine 8 ($^{3}C_6H_{14}{}^{15}N_2O_2$; Sigma) or labeled with normal arginine/lysine for five duplications in high glucose Dulbecco's modified Eagle's medium devoid of those amino acids (Invitrogen, Carlsbad, CA) and supplemented with dialyzed fetal bovine serum (10 kDa size cutoff, Invitrogen). Cytolplasmic extracts were prepared by douncing and subsequent centrifugal removal of nuclei and membranous components (15, 16). The resulting extracts were mixed in the required heavy/light ratio by protein concentration and subjected to in-solution digest as described earlier (17). In brief, extracts were adjusted to 8 M urea (Sigma), 50 mM ammonium bicarbonate pH 7.8, reduced using 1 mM dithiothreitol (Sigma) for 30 min at room temperature and cysteines were alkylated by 5.5 mM iodoacetamide (Sigma) for 20 min at room temperature in the dark. Endoproteinase Lys-C (Roche, Basel, Switzerland) was added at 1 μg/50 μg protein and digestion proceeded for 4 h at room temperature. The urea concentration was adjusted to 2 M by addition of 50 mM ammonium bicarbonate and sequencing grade trypsin (Promega, Charbonnières, France) was added to 1 μg/50 μg protein. After overnight incubation at room temperature, proteolysis was stopped by addition of trifluoroacetic acid to 2% final and a volume corresponding to 5 μg peptides was loaded onto StageTips (18) for clean-up and storage. Prior to LC-MS/MS analysis, peptides were eluted using 80% ACN, 0.1% acetic acid and the solvent was subsequently evaporated.

LC and LTQ Orbitrap Velos Setup—Using an EASY-nLC nano-flow high performance liquid chromatography pump (Thermo Fisher Scientific), ~2.5 μg of peptides were loaded onto in-house produced integrated fritless capillary chromatography columns/electrospray emitters. The 15 cm long fused silica columns (75 μm ID, spray nozzle about 5 μm ID) were packed with C_{18} material (Reprosil-Pur C18-AQ, 3 μm particle size, Dr. Maisch GmbH). Online tandem mass spectrometric analysis was performed on a Velos Orbitrap mass spectrometer (Thermo Fisher Scientific; Tune V. 2.6 build 1061 SP2) at 2.1 kV spray voltage. The following gradient was run at 250 nl/min: 5% to 30% solvent B in 85 min, to 60% solvent B in 12 min, to 80% solvent B in 7 min, to 100% solvent B in 2 min at 500 nl/min, 5 min of 100% B at 500 nl/min, to 5% solvent B in 3 min B at 500 nl/min and 8 min at 5% solvent B and 500 nl/min (with solvent A: 0.5% acetic acid; and solvent B: 80% acetonitrile/0.5% acetic acid).

The Xcalibur and MaxQuant Real-Time (MQRT) controlled runs were both set up with the following parameters: For survey (MS) scans resolution was set to 30,000 and target value to 1,000,000 with a maximum ion inject time of 100 ms and a scan range of 300 to 1700 Th. CID MS/MS scans were performed in the LTQ with a target value of 5000 and a maximum ion inject time of 25 ms. SIM scans were collected at a resolution of 15,000 with a target value of 150,000 and a maximum ion inject time of 150 ms. Ion inject time needed to reach the target values was determined with Automatic Gain Control. Linear ion trap activation time, collision energy, activation Q and wideband activation were set to 10 ms, 35%, 0.25 and "true", respectively and

the isolation window for precursor selection set to 2 Th. The mass range for CID scans was calculated based on the one-third cutoff rule (19). For the Xcalibur runs we used dynamic exclusion of 90 s with early expiration turned off.

Analysis of Proteomic Data—The resulting raw data were analyzed with the MaxQuant proteomics computational framework (version 1.1.1.32) (11). Time-dependent mass calibration (20) was performed and the resulting accurate precursor masses matched to the IPI human database (version 3.68, 87,061 entries), complemented with the standard MaxQuant contaminant database, with a maximum allowed deviation of 6 ppm. Andromeda (13) was used to search the acquired CID spectra against the in-silico fragmentation spectra of the matching database entries with a maximum fragment mass deviation of 0.5 Da. Enzyme specificity was set as C-terminal to Arg and Lys, also allowing cleavage adjacent to proline residues and a maximum of two missed cleavages. Carbamidomethylation of cysteine was set as fixed modification and N-terminal protein acetylation and methionine oxidation as variable modifications. The false discovery rate (FDR) was set to 0.01 for peptides and proteins and the minimum peptide length to 6 amino acids. Further analysis of the data provided by MaxQuant was performed in the R scripting and statistical environment (21). The data sets used for analysis are deposited at Tranche (www.proteomecommons.org).

RESULTS AND DISCUSSION

Here we set out to create a conceptual and practical framework for intelligent data acquisition and real-time database searching for proteomics applications. The intelligent agent was developed in C#.NET as part of the MaxQuant computational proteomics environment (EXPERIMENTAL PROCEDURES). The framework is completely modular and easily customizable to different types of experiments. Below we provide proof of principle examples demonstrating the applicability of the intelligent agent to major areas of data acquisition in shotgun proteomics.

In normal operation the mass spectrometer is controlled by high and low level software created by the manufacturer. It allows the user to specify acquisition methods in the high level software on a personal computer, which are uploaded into the mass spectrometer prior to the run, where they provide the necessary settings for the low level control software running on CPUs inside the mass spectrometer. In the case of the Orbitrap family of instruments, this software is collectively called Xcalibur. The low level software records individual spectra in a data dependent fashion and reports the acquired data points back to the high level software on the PC. Xcalibur also allows visualization of the MS data both as it is acquired and from raw finalized data files.

The intelligent agent is implemented using the algorithms and software modules of MaxQuant and runs on the same PC as the Xcalibur software (Fig. 1*A*). This is possible because MaxQuant Real-Time uses moderate memory and CPU resources. It communicates with the low level acquisition software in the mass spectrometer through Thermo Fisher Scientific's "instrument OCX." Like the high level Xcalibur software it receives scan data as they are acquired. However, in contrast to normal operation of the mass spectrometer, MaxQuant Real-Time can base its decisions on the full capabilities provided by the MaxQuant computational proteomics environment. This can be done both on the real-time data as well as data from prior MS acquisition. Furthermore, they can also be directed by results from downstream bioinformatic analyses. During operation the intelligent agent triggers full scans and collects the resulting data from the instrument OCX as it becomes available to perform isotope pattern and SILAC pair assignment (Fig. 1*B*). When MaxQuant real-time locates candidates (either isotope patterns or SILAC partners), fragmentation scan events are scheduled on the machine. As soon as the fragmentation data becomes available, the spectrum is analyzed by the real-time search engine and its output can be used.

Direct Instrument Control by Real-Time MaxQuant—We wrote the intelligent agent in the programing language C#.NET making use of algorithms and functionality already implemented in MaxQuant (EXPERIMENTAL PROCEDURES). MaxQuant Real-Time extends these mainly by dealing with instrument control, feature detection from partial data, mapping of previous MS data onto the current run as well as incorporating results from bioinformatic analyses. Its user interface is restricted to gathering the parameters for acquisition and commencing the measurement.

A major challenge in real-time data analysis is the fact that not all relevant information is available. In our case, the complete elution profiles of the isotope patterns have not been recorded at the time that the directed acquisition decision has to be made. We addressed this issue by adapting the feature detection algorithms of MaxQuant (11). For the single MS spectrum the same centroiding and local minima detection algorithms can be used because the full information is present. For three-dimensional peak reconstruction and calling of isotope patterns, the required correlation values are relaxed. The same applies to the detection of SILAC or other stable isotope pairs. To ensure incorporation of maximal information, we add the data from new incoming spectra to the data structure of each three-dimensional peak and perform the isotope pattern and SILAC pair detection again. When a peak end is detected or if the peak cannot be assigned to an isotope pattern within an empirically determined time frame, it is dropped from the data structure.

The intelligent agent can also match information from previous acquisitions to the real-time information. This is a generalization of the inclusion or exclusion list concept. In addition to mass, charge state and retention time we incorporate minimum and maximum retention times, which are adjusted dynamically during the run, SILAC state and partner information as well as expected intensity, unique peptide identities as well as flags for directed data acquisition decisions (EXPERIMENTAL PROCEDURES).

Fig. 1*B* depicts the control structure of the intelligent agent. The parameter file is read at the start and contains all user input. The agent first sets global parameters of the mass spectrometer such as raw file name, target values for all scan

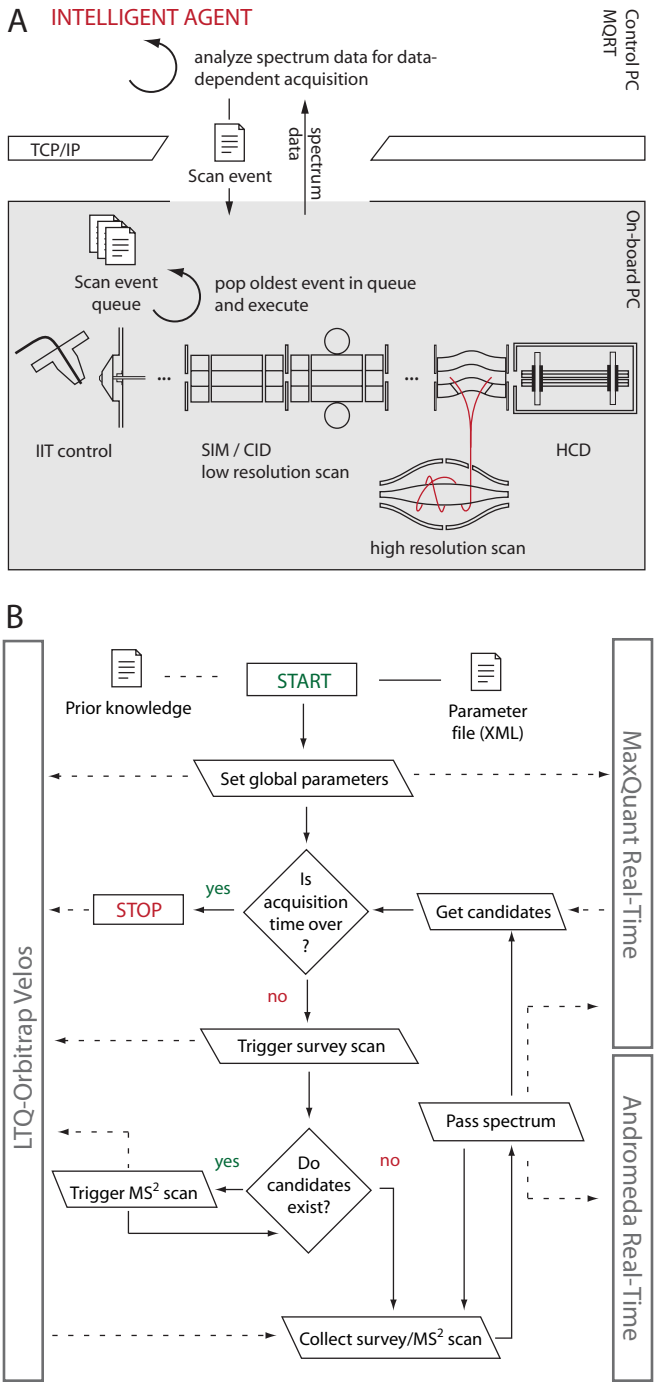


FIG. 1. **The complex environment the intelligent agent operates in.** *A*, The intelligent agent workflow, where scan event definitions are uploaded to the on-board PC, whereas all logic is located on the control PC (indicated with "INTELLIGENT AGENT" in red letters). The on-board PC maintains a queue from which the oldest scan event is retrieved and executed. The recorded spectrum data is downloaded to the PC for evaluation and storage. *B*, Flow-chart visualization of the intelligent agent. The parameters for setting up the measurement are stored in an XML-file and prior knowledge can optionally be stored in an extended feature map. Both are read upon start of the process. A full-scan is always triggered, from which MaxQuant Real-Time determines the isotope patterns (and if required the SILAC pairs) of interest. When candidates have been detected, the topN most intense are selected for sequencing and, if requested, SIM scan analysis is performed (not shown). The dashed arrows denote asynchronous and the solid lines synchronous command flow.

types as well as the tune-file to use and then it starts data acquisition. Throughout the run, the agent triggers MS scans, calculates isotope patterns and SILAC partners, determines valid precursor candidates and triggers MS/MS events. In case of a SIM strategy, the agent additionally determines whether a SIM scan is desirable on an isotope pattern or SILAC pair and what mass window to use. Information from prior acquisitions or from bioinformatic analysis is read from files produced in the MaxQuant computational proteomics environment. Note that the intelligent agent operates asynchronously with the data acquisition on the mass spectrometer, ensuring that there is no delay in triggering the scan events because of processing time.

MaxQuant Real-Time for Standard topN Runs—To investigate advanced data dependent strategies with the intelligent agent, we initially set out to create a basic topN method capable of analyzing a complex peptide mixture in depth. This strategy can be viewed as a litmus test for the employed full scan data analysis approach, as the extracted isotope patterns are used for any subsequent decision such as determining candidates to be sequenced. Incorrect decisions at this stage would result in low identification rates in post processing. After each survey spectrum the detected isotope patterns were sorted by intensity after each full scan to generate an initial list of at least doubly charged candidate precursor ions for fragmentation. As a minimal signal threshold, MS/MS scans were only initiated if at least a given percentage (typically set to 40%) of the desired target value could be collected in the maximum fill time. This approach assumes that peptides are detected early in the elution profile while they are still low in intensity and are thus located at the end of the topN queue. By the time the MS/MS experiment is actually performed the signal intensity is assumed to have increased. To guarantee the best possible sequencing efficiency for a SILAC pair, a fragmentation event was scheduled for the more intense partner of the pair.

To increase performance we investigated controlled resequencing, where an isotope pattern is resequenced after a limited amount of time. With the employed peak detection algorithms from MaxQuant it is possible to track the isotope pattern of a peptide during its complete elution time and record properties about its lifetime (in the chromatography conditions employed here this is on average 30 s). As mentioned previously, a fragmentation event is triggered early after which the timestamp of the MS/MS experiment is stored for that particular isotope pattern. If the isotope pattern is still present after an empirically determined resequencing delay and its intensity has increased compared with the one associated with the original sequencing event, a new fragmenta-

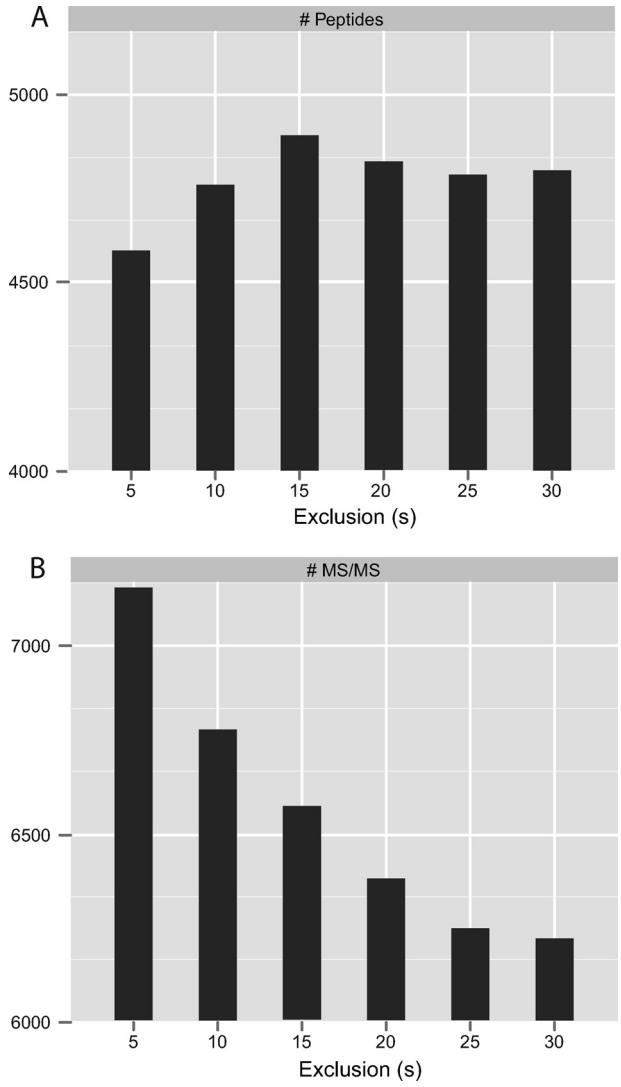


FIG. 2. **Resequencing can be beneficial for peptide identifications.** *A*, Controlled resequencing of peptides with different times shows that re-sequencing of those peptides whose abundance is increased after 15 s yields the best performance. *B*, Total number of MS/MS scans as a function of resequencing delay. Short delays yield more fragmentation events due to short ion injection times but less identified peptides (panel *A*), as many highly abundant peptides keep being resequenced.

tion event is triggered. The restriction at the intensity level ensures that no fragmentation scans are triggered that are likely to be less successful than the previous. Fig. 2*A* shows that, based on the number of identified peptides, a re-sequence delay of 15 s is optimal in the context of the employed chromatography conditions. With longer controlled exclusion times less fragmentation scans are recorded (Fig. 2*B*), which can be attributed to re-sequencing of highly abundant peptides at lower delay times effectively reducing the cycle-time with shorter inject times.

Fig. 3*A* compares the number of collected scans between an Xcalibur run and a MaxQuant real-time directed acquisition. Overall, Xcalibur acquires more scans for the same gradient length because of the communication overhead incurred by the intelligent agent during the upload of the separate scan definitions. This is reflected in a cycle time increase from 1.5 s to 1.9 s for a top10 cycle, indicating an overhead of ~36 ms per scan event, in the current implementation (Fig. 3*B*). However, the intelligent agent more often achieves a top10 cycle compared with the Xcalibur run, resulting in similar identification rates between the two approaches despite the 26% cycle time penalty (Fig. 3*C*). When the communication overhead can be eliminated we expect to see that the intelligent agent will perform better than the Xcalibur method as it will be able to achieve a higher scan frequency (giving better peak definition) and sequencing ability.

Real-time Database Search—We constructed a specialized version of the Andromeda search engine (13) to enable real-time identification of the collected fragmentation spectra. The resulting output is used to inform the real-time decisions of the intelligent agent, but is disregarded during offline analysis with MaxQuant, which can access complete information over multiple RAW files for statistically correct results (Fig. 4*A*). The specialized version was required to keep the identification process within the time-constraints placed by the applied top10 CID fragmentation mode, allowing for a maximum of 170 ms processing time per fragmentation spectrum at a lower average cycle-time of 1.7 s (or even less than 80 ms on a linear quadrupole Orbitrap instrument (22)). To ensure that these time restrictions are met, we limited the normal search-options to: (1) a maximum of one missed cleavage, (2) carbamidomethylation on cysteine as a single fixed modification, (3) two variable modifications consisting of oxidation on methionine and acetylation on the Protein N-term, and (4) the heavy SILAC labels Arg10 and Lys8. Additionally, the database structure used by Andromeda (13) was re-ordered by introducing redundancy to facilitate instantaneous retrieval of the required data (*i.e.* each of the mass ordered peptide item also contains the full sequence information). The search was performed against a human IPI database (version 3.68), supplemented with a list of common contaminants and concatenated with its own reversed version. We found that these settings cover more than 95% of the peptides encountered in the used HeLa SILAC digest. The resulting database file of 815 Mb contained 12,211,914 peptides of which 4,427,989 had unique sequences (including the reverse peptides). Speed critical parts of the algorithm were optimized resulting in search times of MaxQuant real-time less than 30 ms per spectrum in the human proteome (average of 20 ms). This indicates that more liberal search criteria are still feasible. When SILAC information for the precursor peptide is available (*i.e.* number of lysine or arginine residues in the peptide amino acid sequence), incompatible hits from the database are filtered out. The inclusion of a reversed database in the real-time search allowed for estimation of a score cutoff value at a given false discovery rate. As the total identified dataset is

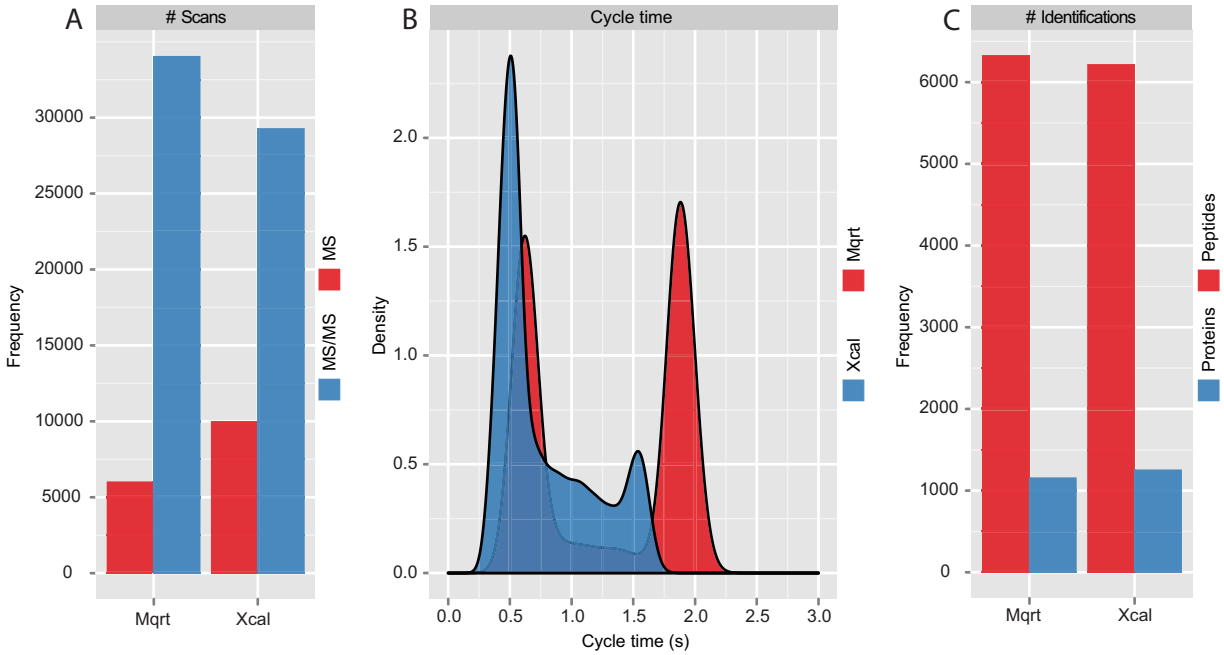


FIG. 3. **Characteristics of the top10 strategy implemented by the intelligent agent.** *A*, Direct comparison with an Xcalibur directed measurement with similar settings. Total number of scans is higher for Xcalibur, but the intelligent agent more often achieves a full top10 cycle. *B*, The difference in the number of scans can be attributed to communication overhead for each scan event of ~36 ms, resulting in a longer cycle time for the intelligent agent (difference between the right peaks). *C*, The total number of identifications is comparable between the two approaches.

unavailable in real-time this is limited to a local FDR, for which the previous 500 scores are tracked in a first-in first-out queue with a flag to indicate reverse or forward peptide hits. The score of each identification is inserted into the queue and the local threshold calculated for the selected FDR. We applied this approach at an FDR of 1 and 5% (Fig. 4*B*), resulting in an average cut-off score of 60 and 40, respectively. Note that these scores are only used for real-time data driven decision and not for the final data analysis. As such, in real-time applications, higher peptide FDR ratios may be desirable to minimize false negatives. Interestingly, the local FDR has a downward trend, which we hypothesize to be attributable to the longer peptides eluting toward the end of the gradient, whose identification could be carried by the precursor mass to a larger extent. To prove this, we scrambled the order of the MS/MS spectra in the data file, which indeed resulted in a horizontal trend for the local FDR.

Real-time Mass Calibration—Although postanalysis of Orbitrap generated data can result in extremely high mass accuracy in the ppb range (12), during acquisition mass calibration can drift by several ppm due to factors such as changes in ambient temperature. This limits all mass measurement driven strategies including simple inclusion and exclusion lists. Peptide identification on-the-fly is uniquely positioned to solve this problem by correcting the current mass deviation of the mass spectrometer in real-time. We calculate the mass deviation of the real-time detected precursor mass from the true values in a moving window of 50 highly confident peptide identifications (local peptide FDR of 1%) and apply a local mass correction based on the median mass offset. For the chromatography conditions used here, the initial 50 high confidence peptide identifications are generally collected within the first few minutes during which time a search tolerance of 20 ppm is used and no correction is applied (similar to the first pass search window employed in MaxQuant (20)). As soon as the list is filled, the precursor mass values are continuously corrected by the moving median prior to database search with a reduced search tolerance of 6 ppm. To demonstrate the robustness of the procedure, we artificially introduced a mass offset of an additional 10 ppm (original deviation avg. 3 ppm; Fig. 4*C*), which was likewise successfully removed (Fig. 4*D*).

As a straightforward application, the above described real-time mass recalibration addresses an important limitation from the concepts of inclusion and exclusion lists. These normally have to be specified with large mass tolerances to accommodate mass drift during the measurement, which is now removed by the real-time search engine (*e.g.* the number of candidates matching inclusion list entries is reduced to less than half by moving the tolerance from 20 to 6 ppm in a 4 min window). This type of mass calibration however does not affect the RAW file, which still need to be corrected in a postprocessing step. Offline calibration in any case is preferable for the final data analysis because postprocessing software has no time-limit and has access to complete data (*e.g.* fully assembled isotope patterns and the complete mass and retention time range) enabling powerful algorithms to be ap-

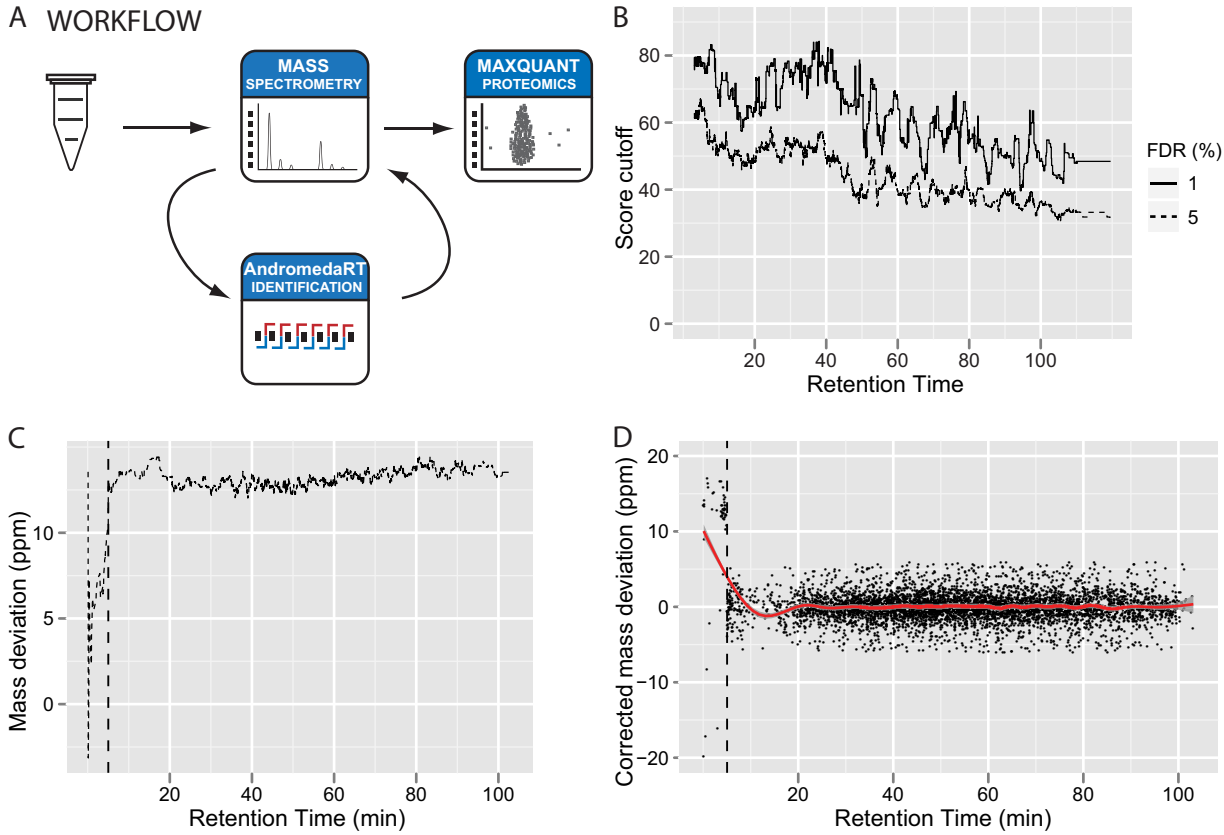


FIG. 4. **Real-time search engine.** *A*, Upper row in the flow chart represents the standard analysis workflow. The real-time search engine (AndromedaRT) identifies MS/MS spectra "on-the-fly." The intelligent agent is continuously receiving identified spectra, enabling dynamic measurements based on knowledge of peptide identity. MaxQuant analysis afterward is done with full knowledge of one or multiple measurements, ensuring complete and statistically valid analysis of the results. *B*, Dynamic calculation of the Andromeda score cutoff resulting in a local FDR of 1% (solid line) or 5% (dashed line). This value is continuously calculated from a population of the last 500 identifications maintained in a first-in first-out queue. *C*, Characterization of the mass deviation with high confidence identifications. The initial search-window is 20 ppm, which can be reduced to 6 ppm when at least 50 nonreverse identifications have been collected (dashed vertical line). From this point the found mass deviation is used to correct the mass deviation of the real-time detected precursor *m/z* prior to the real-time search. The trace represents experimental masses with an added 10 ppm offset. *D*, The mass deviation is centered on 0 (median −0.02 ppm) using the correction factor calculated from the previous 50 high confidence identifications, demonstrating the successful application of the results from the search engine.

plied that result in mass precision in the sub-ppm range (20). In comparison to lock masses on the mass spectrometer, this method is more robust especially for complex mixtures where the lock mass can easily be absent in particularly feature rich full scans.

Prioritized Data Acquisition for Protein Classes of Interest—The capability of MaxQuant Real-Time to reliably mass and retention time align an LC/MS analyses during the data acquisition opens up a plethora of creative targeted data acquisition schemes by the intelligent agent. We here showcase two examples: the prioritized data acquisition for a protein class of choice in the re-analysis of a HeLa lysate and the targeted identification of peptides likely representing modified versions of already identified peptides. In the first example, a first MaxQuant Real-Time driven sample analysis was performed according to a standard "shotgun proteomic" acquisition paradigm with fragmentation spectra acquisition for the topN most intense peptide signals. Offline data analysis using the MaxQuant data analysis pipeline including bioinformatic analysis in Perseus yielded a list of proteins associated with gene ontology (GO) categories. Proteins assigned the GO term "kinase activity" (GO:0016301) were automatically selected and linked back to the corresponding chromatographic peaks through peptide identifications and individual fragmentation spectra. The m/z, charge state as well as elution time properties of these kinase-linked chromatographic features were used to generate an extensive priority list. In a technical replicate analysis of the same sample MaxQuant Real-Time then preferentially targeted kinase-derived peptides for re-sequencing. Subsequent post-acquisition analysis of both replicates in MaxQuant led to the identification of 4978 peptide sequences making up 1186 protein groups. A total of 51 protein groups were associated to the GO term, representing 212 peptide sequences that were added to the inclusion list

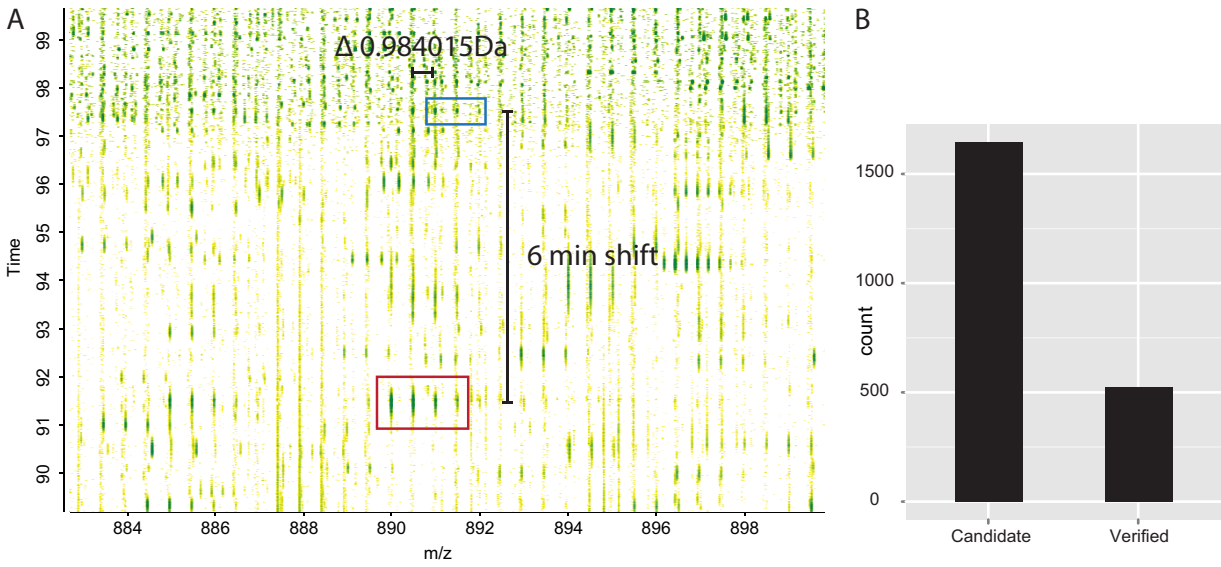


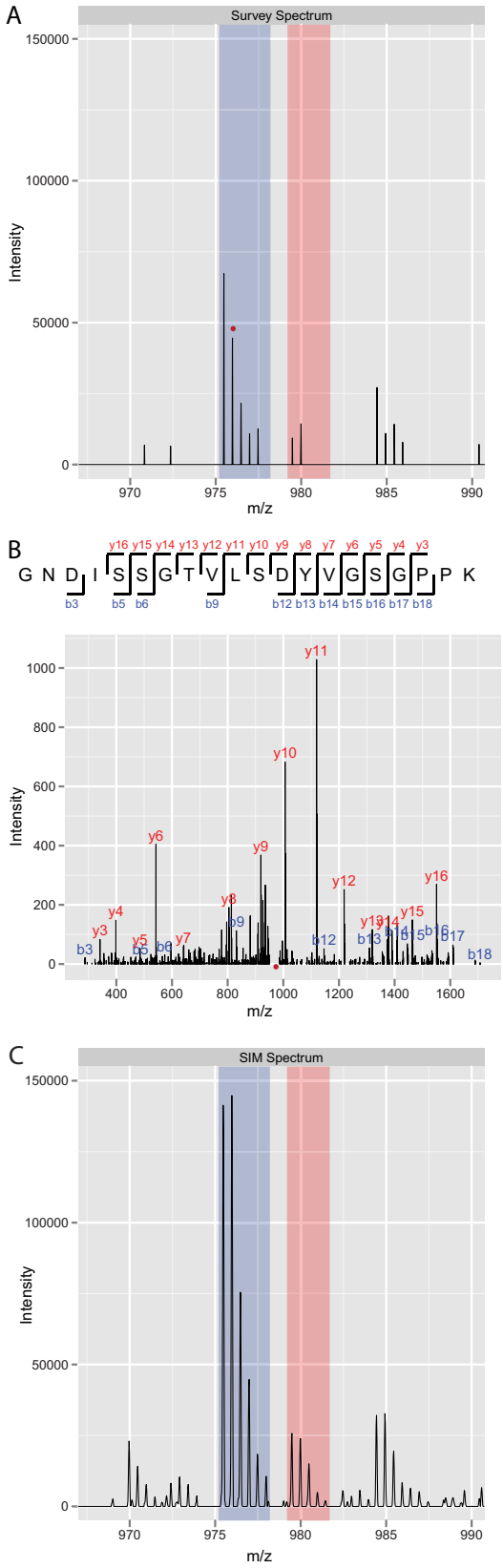
FIG. 5. **Prioritizing the data acquisition of peptides/proteins of interest in real-time.** *A*, Heat map of eluting peptides, visualizing SILAC peptide pairs that potentially differ by deamidation (red box, unmodified; blue box, deamidated). Peptides that were identified in real-time as unmodified and that were followed by isotope patterns representing potentially de-amidated versions, triggered MS/MS events on these isotope patterns. *B*, Identifying peptides with specific post-translational modifications. Note that the deamidated isotope pattern has three rather than four recognized peaks because of its lower intensity.

fed to MaxQuant Real-Time for retargeting. Retargeting was successful in 203 of the 212 cases (96%) and 175 had sufficient scores for re-identification (85%). Although shown here for the automatic re-analysis of kinases, this strategy is applicable to any GO category or indeed for any collection of proteins of interest.

In shotgun proteomics runs, modified peptides often appear as pairs of peptides with a defined mass offset at different elution times. In a second example of a directed analysis but this time using the results of a real-time search, we directed the intelligent agent to verify the identity of peptide pairs with a defined mass and retention time difference, possibly representing a modification. Here we applied this principle to locate peptides with a deamidation of asparagine (Δmass = 0.984015 Da; Fig. 5*A*). Candidate isotope pattern were prioritized for fragmentation if they varied in mass by a difference ± 6 ppm of an unmodified real-time identification of a peptide that had eluted no more than 8 min before. We focused on SILAC pairs only, where the SILAC state was required to be the same for both the unmodified and modified peptide. From the real-time data we identified 1647 deamidation candidates. From the deamidation candidates, 505 proved to actually be deamidated versions of the originally identified peptides as determined from their Andromeda identification (Fig. 5*B*). Again, this example illustrates a generic strategy, which in this case could be applied to any other modification, too.

Targeted SIM Scans for Peptides of Interest—The maximum measurable difference between the strongest and the weakest signal in a spectrum, also termed *dynamic range*, presents a major challenge in mass spectrometric measurements. Isotope clusters for low abundance ions are difficult to reconstruct and a limited dynamic range negatively affects the depth of identification and the precision of quantitation. In trapping instruments, targeted SIM scans offer a remedy to this problem by filling the trap only with a small mass range. Here we wished to quantify signals for which the lower abundant SILAC partner is below the detection limit in survey spectra, by selecting the individual precursor ions in real-time. When MaxQuant Real-Time reported a singlet isotope cluster with a sufficiently high intensity for fragmentation, a SIM scan was scheduled after the MS2 scan with a dynamically constructed *m/z* range. This *m/z* range is either determined based on identification yielding the SILAC state of the triggered MS2 scan (when the score is above the FDR score cutoff), or based on the detected charge state, the known SILAC labels (allowing for the presence of two labels due to one missed cleavage). The window is chosen to encompass a possible lower or higher mass partner. Additional selection window "padding" of –2 and +8 Th on the left and right edges of the mass range, respectively, compensated for the nonlinear isolation efficiency in the linear trap of the hybrid mass spectrometer (empirically derived).

Fig. 6 demonstrates by way of example that this strategy of triggering a SIM spectrum based on real-time information provides quantitative information otherwise not present in the data file and therefore inaccessible to post-acquisition data analysis by MaxQuant. Based on survey scan information, a well-defined isotope cluster of a doubly charged precursor ion at *m/z* 975.44 was identified (Fig. 6*A*). In post-acquisition data analysis, the fragmentation spectrum identified the peak as the peptide with sequence GNDISSGTVLSDYVGSGPPK and light SILAC label (Fig. 6*B*). However, the corresponding heavy



SILAC partner peptide cannot be reconstructed from the regular MS survey spectra covering the entire elution time, because its signal is below the dynamic range. This SIM scan boosted the injection time significantly compared with the full scan (30.7 ms to 252.3 ms), providing selective dynamic range improvement for the potential SILAC pair with a 17-fold increase in the number of ions for the peptide of interest collected. This is not reflected in the provided intensity values (approximately a twofold difference), which are automatically scaled to ions-per-second for the Orbitrap platform. The actual number of ions can be calculated with the ion inject time provided in the meta-data for each scan, which were 30.7 ms and 252.35 ms resulting in 2066 ions and 36,590 ions at maximum intensity for the FULL scan and the SIM scan respectively. Indeed, the resulting spectrum (Fig. 6*C*) clearly extracts the missing SILAC partner from the background and allows retrieval of quantitative ratio information.

From the collected RAW file, a total of 7,123 MS2 spectra were collected during a MaxQuant Real-Time driven 140 min LC-MS analysis of a SILAC-controlled immunoprecipitate. Postacquisition data processing by MaxQuant led to the association of 1639 of these scan events with peptide sequences (1353 peptide identifications at 1% FDR). After postacquisition data processing, 66% (896) of the resulting peptide identifications were not associated with SILAC ratios and 832 of these were detected by MaxQuant Real-Time as singlet and subjected to a custom SIM scan acquisition. The approach yielded 204 rescued ratios not deducable from survey scan data due to dynamic range limitations. The additional time required for executing the SIM scans was 50 min (or 35.7% of the total retention time); therefore the SIM strategy on this mass spectrometric platform is not optimal for complex mixtures. However, the sample used here is not of high complexity allowing for time consuming operations that ensure higher quality quantification for selected precursors.

Conclusions and Outlook—Shotgun proteomics produces a very large number of peptides to be fragmented and quantified, many more than can be handled even by modern mass spectrometers (5). Here we have introduced the concept of an intelligent agent, which provides a framework to allocate measurement resources to features of interest in a dynamic way. Intelligent agents are already widely used in many areas

FIG. 6. **"Dig Out" of missing SILAC partners through selective ion monitoring (SIM).** *A*, Example spectrum where low signal intensity and limited dynamic range render SILAC partner identification impossible from survey spectrum information (mass range selected according to the scheduled SIM scan), but still yielding an Andromeda search result (precursor *m/z* marked with red dot). *B*, CID fragmentation spectrum identifying the "light" SILAC partner with fragment annotation by the Andromeda search engine (red dot indicates *m/z* of precursor ion). *C*, Real-time data analysis allows the intelligent agent to dynamically schedule a SIM scan that extracts the corresponding partner's isotope cluster from background (precursor *m/z* marked with a red dot).

of information technology and our results show that this is also a promising approach in proteomics. Data dependent decisions have been performed in LC MS/MS for many years—usually implemented at the low level in the instrument vendor's software for data acquisition. The intelligent agent concept subsumes these separate implementations and collects them into a coherent structure.

In this paper we have described a first implementation of such an agent, mainly using proof of principle applications. In particular, we re-implemented the functionality of the commonly used top10 fragmentation methods and extended this with controlled re-sequencing based on the intensity development of the eluting peptides. To provide the maximum amount of information to the agent, we implemented a real-time search engine based on an extremely fast version of the Andromeda search engine. Clearly there are numerous possibilities for directing measurements in real-time once the likely identity of the peptide is known. Here we used the real-time search engine to improve the mass accuracy during measurements several-fold, which in itself improves the ability of the agent to make mass based decisions. The intelligent agent can also incorporate prior information to direct measurements in real-time. As examples, we described a streamlined scheme to re-measure peptides associated with a set of proteins of interest and to interrogate potentially modified peptides. Finally, we provided an example of dynamically allocating SIM windows to selectively boost retrieval of quantitative information of stable isotope pairs.

Clearly, at this stage most of these examples could also have been implemented on a case-by-case basis in the manufacturer software. However, we suggest that it is useful to conceptualize and implement a separate entity that is under the control of the scientist. Unlike a pre-programmed logic embedded in the instrument, the intelligent agent is extremely flexible and extensible and it can draw on external, prior knowledge. As more and more of such auxiliary knowledge is available, and as the intelligent agent is more capable of directing the instrument, it will increasingly be able to focus the measurement to the most important part of the proteome under investigation. The software is currently implemented for the LTQ-Orbitrap Velos instruments and through the OCX framework, which entails certain limitations in terms of overhead and robust day to day operation. However, we expect that increased firmware and hardware support could result in dramatically improved performance in shotgun proteomics in the future, especially on novel mass spectrometric platforms such as the quadrupole Orbitrap instrument (Q Exactive (22)). The software for MaxQuant Real-Time and our implementation of the LTQ-Orbitrap Velos control software is available upon request. However, the OCX library is a property of Thermo Scientific.

Acknowledgments—We thank scientists at Thermo Fisher Scientific, especially Eric Hemmenway, Oliver Lange, and Andreas Kuehn, and our colleagues at the Max Planck Institute, for help and fruitful discussions.

* This project was supported by the European Commission's 7th Framework Program PROteomics SPECificat ion in Time and Space (PROSPECTS, HEALTH-F4-2008–201648).

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

¶ Current address: Shenzhen Engineering Laboratory for Proteomics, BGI-Shenzhen, Shenzhen 518083, China.

‖ Current address: Weill Cornell Medical College in Qatar, Qatar Foundation, Education City, Doha, State of Qatar.

** Equal contribution.

Data availability: Supplementary data is available with this publication at the MCP web site. Raw MS files were uploaded to Tranche (www.proteomecommons.org) as "Graumann *et al.* Intelligent Agent". Hash code to access the RAW files: hozVUJzhjj7z9Txr68oWlAFGDcxuM8oO9sifWgWD1jAlfK1w4mRlpnOXBzy/YQkhJviKv34huijgchc5pVD2/whXfyMAAAAAAAAPZQ = = .

REFERENCES

1. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
3. Yates, J. R., 3rd, Gilchrist, A., Howell, K. E., and Bergeron, J. J. (2005) Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **6,** 702–714
4. Walther, T. C., and Mann, M. (2010) Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190,** 491–500
5. 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
6. Rudomin, E. L., Carr, S. A., and Jaffe, J. D. (2009) Directed sample interrogation utilizing an accurate mass exclusion-based data-dependent acquisition strategy (AMEx). *J. Proteome Res.* **8,** 3154–3160
7. Schmidt, A., Claassen, M., and Aebersold, R. (2009) Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr. Opin. Chem. Biol.* **13,** 510–517
8. Beck, M., Claassen, M., and Aebersold, R. (2011) Comprehensive proteomics. *Curr. Opin. Biotechnol.* **22,** 3–8
9. Swaney, D. L., McAlister, G. C., and Coon, J. J. (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **5,** 959–964
10. Russell, S. J., and Norvig, P. (2009) *Artificial Intelligence: A Modern Approach*, 3rd Ed., Prentice hall, Englewood Cliffs, NJ
11. 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
12. Cox, J., and Mann, M. (2009) Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* **20,** 1477–1485
13. 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
14. Wenger, C. D., Boyne, M. T., 2nd, Ferguson, J. T., Robinson, D. E., and Kelleher, N. L. (2008) Versatile online-offline engine for automated acquisition of high-resolution tandem mass spectra. *Anal. Chem.* **80,** 8055–8063
15. Dignam, J. D., Lebovitz, R. M., and Roeder, R. G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11,** 1475–1489
16. Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W., van Schaik, F. M., Varier, R. A., Baltissen, M. P., Stunnenberg, H. G., Mann, M., and Timmers, H. T. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131,** 58–69
17. Graumann, J., Hubner, N. C., Kim, J. B., Ko, K., Moser, M., Kumar, C., Cox, J., Schöler, H., and Mann, M. (2008) Stable isotope labeling by amino

acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell. Proteomics* **7,** 672–683

18. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem*. **75,** 663–670
19. Stafford Jr., G. C., Kelley, P. E., Syka, J. E. P., Reynolds, W. E., and Todd, J. F. J. (1984) Recent improvements in and analytical applications of advanced ion trap technology. *Int. J. Mass Spectrom. Ion Proc.* **60,** 85–98
20. Cox, J., and Mann, M. (2011) Software lock mass by two dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* **22,** 1373–1380
21. Ihaka, R., and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5,** 16
22. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** M111.011015