Database independent proteomics analysis of the ostrich and human proteome

A. F. Maarten Altelaar[a,b], Danny Navarro[a,b], Jos Boekhorst[c], Bas van Breukelen[a,b], Berend Snel[c], Shabaz Mohammed[a,b,1], and Albert J. R. Heck[a,b,1]

[a]Biomolecular Mass Spectrometry and Proteomics Group, Utrecht Institute for Pharmaceutical Sciences and Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands; [b]Netherlands Proteomics Centre, Padualaan 8, 3584 CH Utrecht, The Netherlands; and [c]Theoretical Biology and Bioinformatics Group, Department of Biology, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH, The Netherlands

Edited by Richard N. Zare, Stanford University, Stanford, CA, and approved October 14, 2011 (received for review May 27, 2011)

Mass spectrometry (MS)-based proteome analysis relies heavily on the presence of complete protein databases. Such a strategy is extremely powerful, albeit not adequate in the analysis of unpredicted postgenome events, such as posttranslational modifications, which exponentially increase the search space. Therefore, it is of interest to explore "database-free" approaches. Here, we sampled the ostrich and human proteomes with a method facilitating de novo sequencing, utilizing the protease Lys-N in combination with electron transfer dissociation. By implementing several validation steps, including the combined use of collision-induced dissociation/electron transfer dissociation data and a cross-validation with conventional database search strategies, we identified approximately 2,500 unique de novo peptide sequences from the ostrich sample with over 900 peptides generating full backbone sequence coverage. This dataset allowed the appropriate positioning of ostrich in the evolutionary tree. The described database-free sequencing approach is generically applicable and has great potential in important proteomics applications such as in the analysis of variable parts of endogenous antibodies or proteins modified by a plethora of complex posttranslational modifications.

Generally, analyzing the cellular or tissue protein content poses a significant analytical challenge because of the complexity of the proteome, the high dynamic range in protein abundances, and the presence of posttranslational modifications (PTMs). Proteomic analysis by MS has taken on this challenge, utilizing the increasing sensitivity and peptide sequencing speed of MS instruments and getting closer to mapping complete proteomes (1, 2). Sequencing proteolytic peptides using fragmentation by collision-induced dissociation (CID) is common practice in high throughput proteome analyses. These fragmentation spectra are then searched against large protein sequence databases containing the predicted spectra, derived from curated genome/protein databases (3). The resulting peptide matches are scored based on the overlap between experimental and theoretical spectra (3, 4). To improve peptide identification, alternative fragmentation methods can be used, such as electron capture/transfer dissociation (ECD/ETD), which primarily result in peptide dissociation of the N-C_α bond (5, 6). ETD largely complements CID especially for longer peptides containing multiple basic residues (7).

MS-based proteomics has evolved into the preferred method for large-scale proteome analysis, although the dependence on gene or protein databases poses a limitation such as when, for example, no accurate proteome database exists, or when complex and/or multiple modifications are present that are not straightforwardly identified with conventional database search strategies (8, 9). To alleviate the dependence of MS-based proteome analysis on databases and thus assisting the analysis of protein PTMs, alternative strategies that do not depend on databases for protein identifications are useful. One such approach utilizes peptide de novo sequencing where the fragmentation spectra are not matched against databases but interpreted by applying established fragmentation rules to attain the peptide sequence (10–12). After de novo sequence analysis, the identified peptides can be submitted to a sequence alignment and similarity search using, for example, BLAST (13).

De novo sequencing is challenging partly due to the complicating presence of multiple and redundant fragment ion series. To reduce misassignments, approaches have been suggested based on improved mass accuracy or precision (14, 15), combining complementary fragmentation data (14, 16, 17) or the simplification of fragmentation spectra by using chemical derivatization to manipulate peptide behavior (18). Recently, we introduced a biochemical approach to simplify fragment spectra based on the protease Lys-N (19), which has enzymatic cleavage specificity for the N-terminal side of lysine residues. Using Lys-N, the majority of the proteolytic peptides contain a single N-terminal lysine residue and can be enriched by strong cation exchange (SCX) chromatography (20). These peptides generate predominantly sequence diagnostic "c" ion series upon ETD (19), resulting in easy to interpret sequence ladders, which opens up a unique avenue for de novo sequencing and, as we show here, markedly reduces the dependence on databases.

Results and Discussion

Here, we explored the performance of our de novo sequencing approach by investigating the proteomes of ostrich and human. An ostrich meat sample was therefore processed, and the protein content treated with Lys-N. The resulting peptides that contained a single lysine at the N terminus were enriched by SCX, after which this pool of peptides was subjected to MS-based sequencing. Intrinsic to the ETD process is the loss of one proton through charge reduction. We have previously demonstrated that in a single proton scenario, the basic moieties influence which fragment ions will be observed (19, 21). The proportion of c and z ions is dependent on the basicity present at each terminus. In the case of tryptic peptides, there are basic moieties present at both termini and so a population of both c and z ions will be observed where the dominance of each series depends on the gas phase basicity of the involved basic moieties. In the case of Lys-N 2+ peptides, both basic moieties are present at the N terminus, causing ETD spectra for these peptides to be dominated by c ions. Because the charge density of the enriched N-terminal lysine containing peptides is low, supplemental activation (22) is used in all ETD experiments. The overall workflow is represented

Author contributions: A.F.M.A., S.M., and A.J.R.H. designed research; A.F.M.A. performed research; A.F.M.A., D.N., J.B., B.v.B., and B.S. analyzed data; and A.F.M.A., S.M., and A.J.R.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Raw data has been uploaded into the publicly available database Tranche, https://proteomecommons.org/ (Accession: All MS raw files can be accessed using the hash code qV2hD0GPWViLQOm/kpVpCXZ61s+uznNZB96JCS1W8dAxwb7kDu+4nwhJLuhrvP/4wYNJDKq9M+RAXuCVf5pijM3KIfsAAAAAAAAFtw==).

[1]To whom correspondence should be addressed. E-mail: a.j.r.heck@uu.nl or s.mohammed@uu.nl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1108399108/-/DCSupplemental.

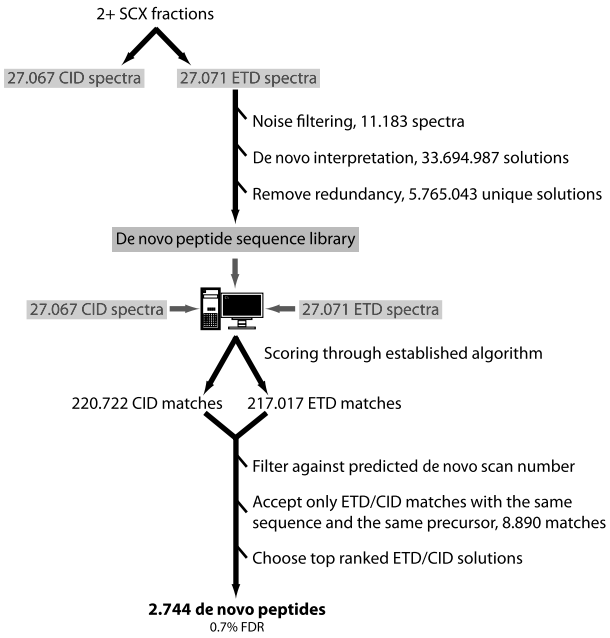


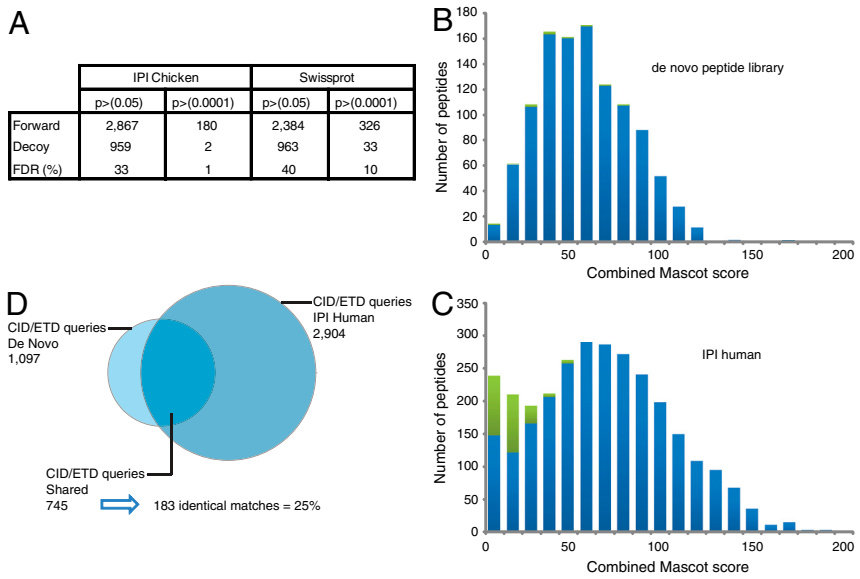
Fig. 1. Schematic overview of the de novo pipeline. After Lys-N protein digestion and SCX enrichment of peptides containing a single N-terminal lysine, a nanoliter flow liquid chromatography separation and a mass spectrometric analysis is performed. Each peptide is sequenced using the fragmentation techniques ETD and CID. The resulting spectra are, subsequently, read into the de novo algorithm, which performs noise filtering and de novo interpretation. From approximately 27,000 ETD spectra, this process resulted in more than 5.5 million possible sequence solutions for 11,183 spectra. The algorithm output consists of a library of all possible peptide solutions alongside the coordinates of the parent spectrum. To retrieve the best match between the ETD fragment spectrum and its reported de novo sequences, the de novo sequence library, alongside a decoy library of equivalent size, is uploaded into the identification engine Mascot that then performs its own matching process on the ETD data. To diminish false positive identifications further, the paired CID data are also matched by Mascot against the de novo solutions. The resulting matches are then filtered such that only results for an ETD spectrum obtained by Mascot originating from a de novo solution for the same spectrum are acceptable. The paired CID scan must also match the same solution(s). If multiple solutions match both the ETD and CID scans then the two scores for the same solution are combined and the highest-ranking result is taken forward. In this way, we can exploit the complementarity of the two fragmentation techniques. The entire process resulted in 8,890 de novo peptide solutions that represent an agreement between Mascot and the de novo algorithm as well as an agreement between ETD and CID. Collapsing the data further led to 2,744 unique nonredundant peptide sequences.

schematically in Fig. 1. We initially used the simplicity of the ETD fragmentation spectra of the single lysine containing peptides to deduce their sequences by automated de novo sequencing using an inhouse developed algorithm (23). The primary output of the algorithm consists of a de novo peptide library containing multiple putative solutions for each query. Subsequently, we uploaded the peptide library into Mascot to score the same query against the derived de novo sequences. The size of the de novo spectral library of possible solutions was considerable in size, approximately three times the size of International Protein Index (IPI) Human, due to the presence of a high number of putative solutions with high similarity. This ambiguity was caused largely by sequence gaps in the fragment ion series, leading to the algorithm generating all possible amino acid combinations that could explain the gap. To obtain an estimate of the quality of our results, we constructed a decoy peptide library by reversing the sequence of the putative solutions with the exception of the N and C termini (e.g., **K**ANYPEPTID**E** becomes **K**DITPEPYNA**E**) and conducted the decoy searches in a concatenated database fashion (24). Although decoy searches cannot determine the success rate of the de novo approach, it does give a handle on the quality of the spectra used.

Because we analyzed each peptide by ETD and CID, it was possible to use the CID data as part of a validation strategy. We scored the CID spectra against the ETD putative solutions derived by de novo sequencing using Mascot. The results mostly agreed with the solutions provided by the de novo algorithm, but often differed in terms of their ranking when there were multiple highly similar solutions. The occurrence of incomplete sequence information obtainable from either the CID and/or ETD spectra frequently caused this disagreement. It is well documented that CID and ETD can be highly complementary in covering parts of a peptide sequence (14, 16, 25). Therefore, for every precursor we combined the ETD and CID ion score of the agreeing matches in the top 10 rankings, and took forward the highest-ranking combined score. Essentially, CID helps eliminate the gaps present in the ETD data and allows the multiple solutions for the ETD data to collapse, often to a single solution. This resulted in the identification of 2,744 de novo peptides, originating from approximately 27,000 ETD/CID spectra pairs, generating an annotation rate of >10% for our de novo sequencing approach.

To assess the performance of our de novo approach we performed two direct comparisons between identifications obtained from de novo sequencing and a conventional database-searching approach. First, we matched our Lys-N ETD ostrich dataset against one of the closest related species with a sequenced genome, the chicken, (26) as well as against the entire UniProt database, using the Mascot search engine (Fig. 2*A*). We observed an exceptionally high false discovery rate (FDR) value for the chicken result that we argue can largely be explained by the low similarity between ostrich and chicken. Adjusting the Mascot acceptance criteria to obtain an FDR <1% reduced the number of unique peptide identifications to 180, of which most belonged to highly conserved proteins such as Myosin. When we searched the experimental data against the entire UniProt database, the obtained results were worse still (Fig. 2*A*). These results show that for the experimental ostrich data, random matches largely dominate the IPI chicken or complete UniProt database search results. As pointed out by Vallender et al. (27) genetically close related species still frequently display amino acid changes in every protein. Birds have rapidly diverged and ostrich and chicken are distant family members, (26) as reflected in our poor protein identification numbers. Our de novo sequencing approach, in contrast, is able to generate a large number of peptide sequences from the fragmentation spectra independent of databases.

Next, we assessed the performance of the de novo approach against a conventional database-searching strategy using a sample originating from a highly accurate genome using experimental MS data from a single SCX fraction of Lys-N digested human HEK293 cells. To allow a fair comparison between both methods we filtered the ETD and CID data in both approaches, based on the highest combined score. When we plotted the number of de novo peptides identified in the human experiment against the combined CID/ETD score, we obtained a "Boltzmann-like" distribution (Fig. 2*B*), which is somewhat atypical and can be related to the use of paired identifications. Standard database search strategies usually generate a bimodal score distribution with the dominant false positive distribution present at the low scoring side of the graph that is present (albeit at a lower level than normal) when using the database search strategy (24, 28, 29) (Fig. 2*C*). Evaluating the decoy results for the human de novo data indicated an FDR of 1%, where only 10 decoy hits were observed after the filtering process (Fig. 2*B*). The low FDR level result suggests that only a few of the spectra are matched against decoy sequences within our de novo dataset. Both strategies generated, in common, solutions for 745 CID/ETD pairwise queries (Fig. 2*D*). Focusing on these queries, the de novo and database strategies agreed fully on 183 sequences, representing a success



	IPI Chicken		Swissprot	
	p>(0.05)	p>(0.0001)	p>(0.05)	p>(0.0001)
Forward	2,867	180	2,384	326
Decoy	959	2	963	33
FDR (%)	33	1	40	10

Fig. 2. Performance characteristics of the Lys-N ETD de novo sequencing strategy. (*A*) Peptide identification results from a Mascot search of the Lys-N peptides derived from ostrich against the IPI chicken and the combined UniProt databases. The database search against IPI chicken resulted in the identification of 2,867 forward peptides ($p < 0.05$) and 959 decoy peptides (using the Mascot decoy strategy), leading to an FDR of 33.5%. Forcing the IPI chicken identification procedure to reach an FDR <1% required stringent Mascot identification criteria (*p* value <0.0001 and score > 44), reducing the number of unique peptide identifications 15-fold to 180. The database search against the UniProt database identified 2,384 forward peptides ($p < 0.05$) and 963 decoy peptides, resulting again in an FDR of 40%. Adjusting the p value to <0.0001 resulted in an FDR of 10.1% at a Mascot score cutoff of >55, and the identification of only 44 unique peptides. (*B*) Combined CID/ETD Mascot score plotted against the number of unique peptides identified (decoy hits depicted in green) for a single SCX fraction of Lys-N digested human HEK293 cells using the de novo sequencing identification strategy and (*C*) using a conventional IPI-human-based database search strategy. The results from the IPI human database showed increased false positive hits at lower Mascot ion scores (ion score < 30), whereas this number was significantly lower in the de novo strategy. (*D*) Comparison between the de novo workflow, as applied to the ostrich data, and a conventional database search against IPI human for a human HEK293 sample. The de novo workflow resulted in 1,097 paired CID/ETD queries generating the identification of 1,029 unique peptide sequences from this single SCX fraction. The common database search generated 1,492 unique peptides with an FDR of 1%. Both strategies had 745 CID/ETD pairwise queries in common, of which 183 peptide sequences agreed fully, representing a minimum success rate of 25%.

rate of 25%, and if we consider "raw peak lists" we obtain a success rate of 6.3%. As reported previously by our group, the remaining 75% contain a certain amount of ambiguity in their reported sequence, which may incorporate the true result. Additionally, the receiver operating characteristics (ROC) curves for the de novo and database-based search show a similar trend (Fig. S1), providing further confidence in our de novo approach.

Due to the absence of a database, we cannot obtain the exact level of accuracy for the ostrich experiment, although in principle, it will be similar to that obtained for the human experiment because the strategy is independent of the analyzed species. We can, however, estimate the quality of the data by calculating the level of peptide coverage through the presence of b, y, and c ions. A detailed summary of the information for each peptide [SCX fraction, scan number, peptide sequence, peptide length, number of fragments observed, coverage (%), coverage in the sequence, fragment ions annotated, combined score, combined score normalized for peptide length, ETD score, ETD rank, CID score, CID rank, and delta score] can be found in Table S1. Furthermore, plots of peptide length and combined score versus coverage (%) (Fig. S2) show a global view of the information captured for each peptide by our de novo approach. We observed a minimum of 50% of fragment sequence coverage for the data and 923 peptides with complete backbone coverage. Moreover, because we know the phylogenetic location of ostrich, (26) we attempted to obtain the degree of accuracy in our experiment by conducting a phylogenetic analysis. We infer the position of the ostrich in the tree of life using the de novo proteomics analysis and compare this tree to the taxonomic tree as provided by the National Center for Biotechnology Information (NCBI) (30) (Fig. S3). One of the classical ways of positioning a species in the tree of life is the generation of a phylogenetic tree based on a multiple sequence alignment of one-to-one orthologs in all species of interest (31). As there is no complete genome sequence available for the ostrich, we infer orthology by globally aligning ostrich peptides against several species chosen to represent multiple levels of evolutionary relationship within the tree of life at different distances from the ostrich. We chose two avian species, chicken and zebra finch, and two species that are more distant, lizard and human (Fig. 3*A*). Chicken is supposedly the closest relative to ostrich for which the genome is sequenced and the zebra finch is more distant but still on the same avian branch. A maximum likelihood tree of the concatenated multiple sequence alignments of these peptides with their orthologs (Fig. 3*B*) has exactly the same topology as the tree provided by NCBI taxonomy (Fig. S3). The topology of the resulting tree from the negative control (Fig. 3*C*) is different from the topology of the tree based on real data, incorrectly placing ostrich with lizard. This shows that the topology of the original tree is in fact dependent on the primary sequence of the de novo peptides, and does not merely reflect amino acid composition. The correctly generated phylogenetic tree of life based just on the de novo data clearly corroborates our approach. Functional analysis of the peptides that could be mapped to chicken proteins further underlines the biological significance of our results: as expected, the most significantly overrepresented gene ontology (GO) slim category is "generation of precursor metabolites and energy" (GO-ID 6091; $p < 1e\text{-}04$) and includes peptides mapped to the chicken proteins muscle phosphofructokinase (ENSGALP00000021650) and malate dehydrogenase (ENSGALP00000014374).

The simple ladder sequences generated in our de novo approach provide benefits in the confident site determination of posttranslational modifications as illustrated (in Fig. 3*D*) by the fragmentation spectra of the ostrich phosphopeptide KGILAADESTGSIA from fructose-biphosphate aldolase A. We could distinguish with ease this peptide in two different phosphorylated forms using Lys-N and ETD. An additional illustrative example is the identification of two peptides containing an acetylated lysine (indicated in red, Fig. 3*E*). BLAST searching revealed that

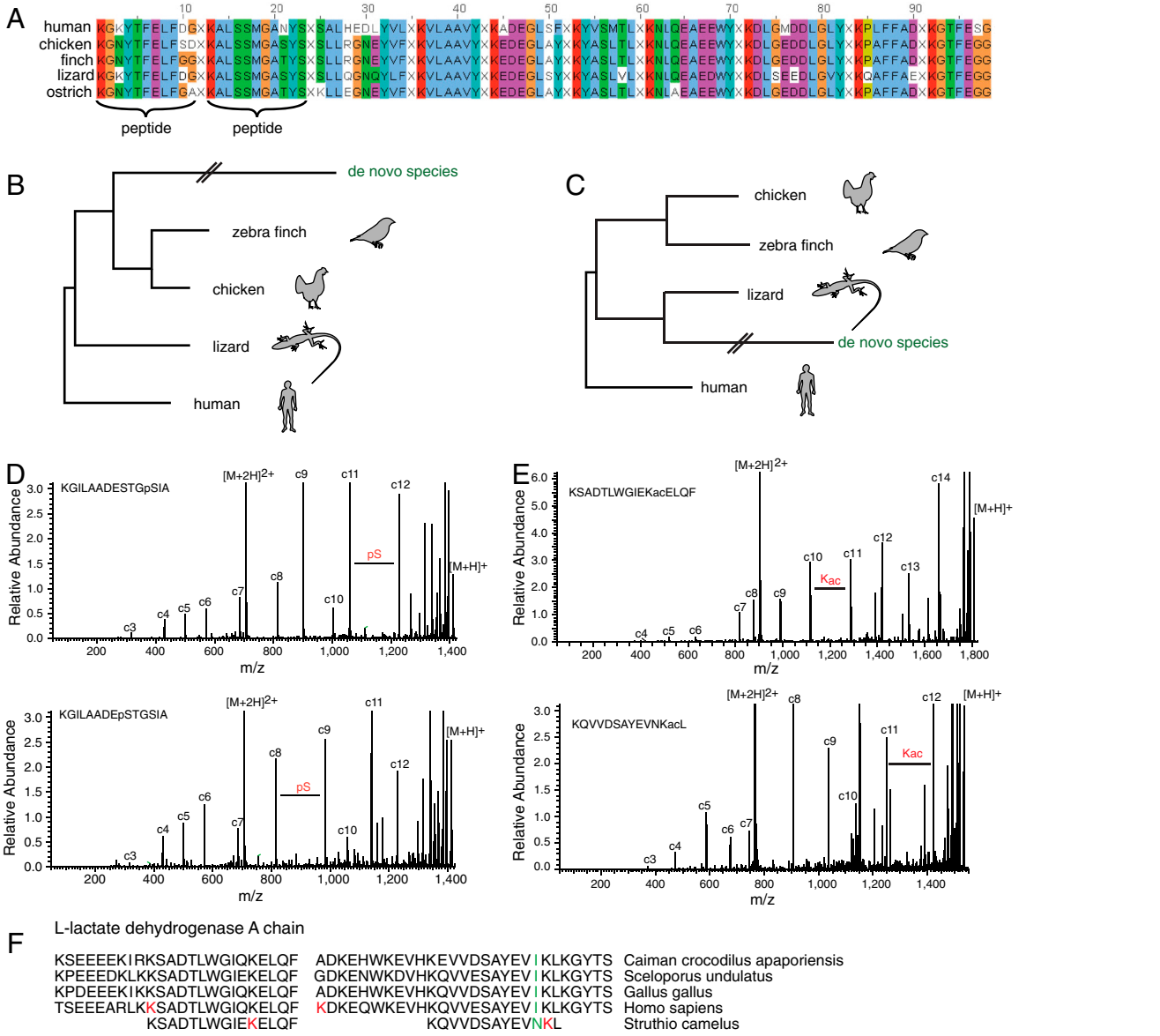


Fig. 3. Phylogenetic analysis of the peptide dataset generated by the de novo approach. (*A*) Ensembl Compara-based alignment of concatenated de novo identified peptides (peptides are separated by X) with the sequence of four selected species at different evolutionary distance, i.e., chicken, zebra finch, lizard, and human. Because the de novo approach cannot distinguish between isoleucine and leucine, I is changed to L. All ostrich derived de novo peptides that map uniquely to a protein in one or more of these species are selected and of these peptides, the subset that mapped to a protein with exactly one orthologous sequence in the other species was used for the alignment. (*B*) Maximum likelihood tree of the concatenated multiple sequence alignments of the selected peptides with their orthologs. There was high bootstrap support for the distinct avian branch (67%), as well as for the distinct differentiation within the avian branch (85%). (*C*) Negative control. Maximum likelihood tree of the concatenated multiple sequence alignments of the selected peptides with their orthologs after randomizing the order of the residues in the peptides of our species of interest. There was high bootstrap support (85%) for incorrectly placing ostrich with lizard. (*D*) ETD MS/MS spectra of two different forms of the ostrich phosphopeptide KGILAADESTGSIA clearly revealed the site localization capabilities of this approach with a clear distinction between the isobaric phosphopeptides KGILAADESTGpSIA and KGILAADEpSTGSIA. (*E*) ETD MS/MS spectra of two lysine acetylated peptides from the protein ostrich L-lactate dehydrogenase A chain, whose human homologue is known to be heavily decorated with lysine acetylations and (*F*) sequence alignment of these peptides from ostrich with several other species showing the acetylated lysines (in red), the high conservation of the lysine found to be acetylated in our study, and a novel point mutation (in green).

both these peptides originate from the ostrich protein L-lactate dehydrogenase A chain, whose human homologue was recently reported to be highly acetylated at numerous lysine residues (32). The unique lysine acetylation sites reported here (Fig. 3*F*) are localized on highly conserved lysines, in line with a recent report claiming high degrees of conservation for lysine acetylation (33).

In all these examples, a traditional database search would have had difficulty resolving the correct peptides. The phosphorylated ostrich peptide sequences, although conserved in human (34) are completely absent from the IPI chicken database and the lysine acetylated peptides from ostrich contain point mutations when compared to their chicken and other homolog sequences. Moreover, the second acetylated peptide (KQVVDSAYEVNKacL) contains additionally a unique, single nucleotide exchange, isoleucine to asparagine mutation (indicated in green, Fig. 3*F*), providing further illustrative evidence of the potential of the presented approach.

Notably, our human data revealed a number of spectra resulting in high scoring de novo identification results for which the IPI based database search gave no or very poor results (Fig. 4 and Fig. S4). Alignment searches of the de novo identified sequences using BLAST did not reveal any similarities with known proteins. We generated synthetic peptides of several of these de novo pep-

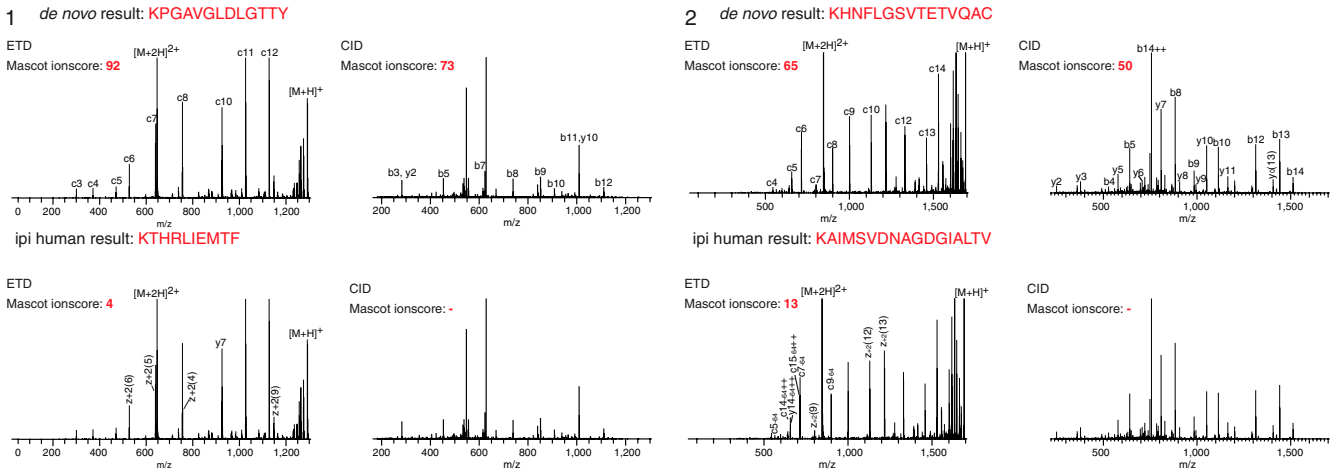


Fig. 4. Two examples of ETD and CID fragment spectra originating from the human HEK293 sample where the de novo solution scored significantly higher than the database result. The illustrated examples of high quality ETD and CID spectra produce de novo sequences that are neither present in the IPI human database nor show significant similarity with sequences from other species, as confirmed by BLAST searching.

tides unknown by IPI; the resulting ETD fragment spectra were largely indistinguishable from the HEK293 experimental peptide spectra (Fig. 5), suggesting these sequences are real but are not present in current high quality databases. When we used a combined score cutoff for our de novo identifications, we observed 202 peptide sequences that were not identified in the database approach.

In summary, we demonstrate here a de novo approach that exploits the simple ladder sequences generated by ETD dissociation of single lysine Lys-N peptides for peptide sequencing. De novo sequencing of these fragment ladders leads to a remarkable 25% success rate of completely correct peptide sequences. The remaining 75% of peptides identified by our de novo approach contain a certain amount of ambiguity in amino acid assignment, (23, 35) which may be improved using mass spectrometers with superior ETD supplemental activation (36) and by performing ECD/ETD at high mass accuracy (14, 37). Although genome sequencing has become relatively easy in recent years, we argue that our de novo approach will be a valuable tool in, for instance, the analysis of hypervariable parts of proteins as presented in antibodies, and proteins obtained from extinct species (8, 38). As it becomes apparent that several key proteins are highly decorated by various co-occurring PTMs, the application of our approach will be most beneficial in the analysis of such complex posttranslational modifications.

Materials and Methods

Sample Preparation. Approximately 500 mg of ostrich muscle tissue (obtained, as a steak, at the local butcher) was homogenized in 8 M urea using a Dounce Homogenizer. The resulting tissue suspension was further homogenized by microtip sonication (Labsonic M; Sartorius AG), after which the sample was centrifuged at 14,000 × *g* for 30 min at 4 °C. A fraction of the supernatant (corresponding to 481 μg of protein, as determined by Bradford) was placed on a 1D SDS/PAGE gel and run approximately 1 cm into the gel. The resulting single band was excised and cut into four pieces and washed with ammonium bicarbonate. Subsequently, the content of the gel pieces was reduced and alkylated with DTT and iodoacetamide, respectively. After further washing the gel was incubated with Lys-N (metalloendopeptidase derived from *Grifola frondosa*; Seikagaku Corporation) (protein:enzyme 85:1) overnight at 37 °C. Peptides were extracted using 5% formic acid.

HEK293T cells were lysed by resuspension in lysis buffer containing 50 mM ammonium bicarbonate, 8 M urea, EDTA-free protease inhibitor cocktail (Roche), 1 mM potassium fluoride, 1 mM sodium orthovanadate, and 5 mM potassium phosphate. The mixture was vortexed for 20 min on ice. After centrifugation at 1,000 × *g* for 10 min at 4 °C to pellet unbroken cells and debris, the HEK293 lysate was reduced and alkylated with DTT and iodoacetamide, respectively, and digested with Lys-N at an enzyme/substrate ratio of 1:85 in 8 M urea/50 mM NH4HCO3, pH 8.0, overnight at 37 °C (39).

Strong Cation Exchange Chromatography. SCX was performed essentially as described previously, (20) using an Agilent 1100 HPLC system (Agilent Technologies) with two C18 Opti-Lynx (Optimized Technologies) guard columns and a polysulfoethyl A SCX column (PolyLC; 200 × 2.1 mm inner diameter, 5 μm, 200 Å). A total of 50 SCX fractions (1 min each; i.e., 50 μL elution volume) were collected and dried in a vacuum centrifuge.

Liquid Chromatography (LC) MS/MS Experiments. The dried SCX fractions were resuspended in 40 μL of 10% formic acid and 10 μL of each of the SCX fractions 7–16 were subjected to nano-LC-MS/MS analysis, performed on an Agilent 1200 HPLC system (Agilent Technologies) connected to an Orbitrap XL Mass Spectrometer equipped with an ETD source (Thermo Scientific), as previously described (40). Parent ions were fragmented by CID and ETD in data dependent mode with an automatic gain control value of 5.00e + 04 and a maximum injection time of 500 ms. ETD fragmentation was performed with supplemental activation, fluoranthene was used as reagent anion, and ion/ion reaction time in the ion trap was charge state dependent with 50 ms reaction for 2+ ions, 33.3 ms for 3+ ions, and 25 ms for 4+ ions.

Data Processing and Standard Database Searching and Protein Identification. Raw MS data were converted to peak lists using Proteome Discoverer, version 1.0 (Thermo Scientific). Spectra were searched against the IPI Human database version 3.36 (69,012 sequences; 29.002682 million residues), IPI chicken version 3.40 (25,685 sequences; 12,430,339 residues) and the UniProt data-

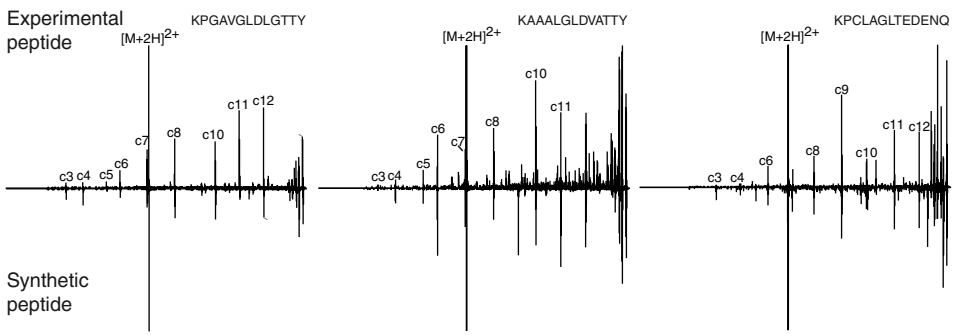


Fig. 5. Comparative assessment of three ETD fragment spectra originating from the human HEK293 sample and their synthetic de novo constructed peptide sequences. Synthetic peptides of the de novo predicted peptide sequences, unknown by IPI, were constructed and analyzed by ETD, the resulting fragment spectra are largely indistinguishable from the HEK293 experimental peptide spectra.

base version 56.2 (398,181 sequences; 143.572911 million residues) using Mascot software version 2.2.0 (Matrix Science), with Lys-N cleavage specificity. The database search was made with the following parameters: a peptide tolerance of ±10 ppm, a fragment tolerance of ±0.6 Da, allowing two missed cleavages, carbamidomethyl as fixed modification, oxidation, phosphorylation, and acetylation (protein N terminus) as variable modifications. Mascot interpretation was accepted for N-terminal acetylation and phosphorylation site assignment.

Protein De Novo Identification. Protein identification was performed by an in-house developed algorithm for de novo sequencing of Lys-N generated peptides fragmented by ETD, as described before (23). In short, the program consists of two components. The first consists of a collection of preprocessing algorithms to perform noise filtering, and the second contains a heuristic algorithm that performs the actual peptide identification, tuned to specifically handle the ETD data generated Lys-N proteolytic peptides. Essentially, the algorithm builds sequences constructed by searching solely for c-type ion fragments. The algorithm takes into account possible neutral losses frequently observed in ETD and the possibility of missing c ions, due to, for example, the presence of proline residues. The algorithm output consists of a library of all possible peptides solutions alongside the raw spectra details.

Phylogenetic Analysis. Predicted proteomes of human (*Homo sapiens*), lizard (*Anolis carolinensi*), chicken (*Gallus gallus*), and zebra finch (*Taeniopygia guttata*) were downloaded from Ensembl (41). All isoleucine residues were changed to leucines, because of our mass spectrometry data cannot distinguish between these residues. The longest splice variant of each gene was used. De novo peptides were mapped against these proteomes using glsearch (42). Peptides were filtered using the following criteria (all criteria need to be met): *(i)* Hits should have an e value of 1e-03 or lower and a sequence identity of at least 70%; *(ii)* if a peptide has multiple hits in a species, there should be a unique best hit; *(iii)* if a peptide has hits in multiple species, the best hits need to be in the same orthologous group. Orthologous groups and the alignments of these groups were downloaded from Ensembl Compara (41). The parts of the alignments of orthologous groups that were hit by a de novo peptide were selected based on the start and stop coordinates of the glsearch high-scoring segment pairs. These subalignments had to meet the following criteria: (i) each of the species of interest (human, lizard, chicken, and finch) should have exactly one protein in the orthologous group; (ii) if there are any gaps in the sub alignment, they need to be in the same column. All sub alignments and corresponding de novo peptides were concatenated and a phylogenetic tree was generated using Phyml (43). For the negative control, randomized peptides were generated by shuffling residues on a per peptide basis, leaving the N-terminal lysine residue in place. The procedure was repeated 100 times. A phylogenetic tree was generated for each of these datasets as described above and the consensus tree was identified using Consense (44). Functional overrepresentation was determined using Fisher's exact test. P values were adjusted for multiple testing through Bonferroni correction.

ACKNOWLEDGMENTS. We thank Richard van Schaik and H. Th. Mark Timmers for synthesis of the peptides. We would like to thank Dr. Rebecca Rose for a careful reading of the manuscript. We kindly acknowledge The Netherlands Proteomics Centre, embedded in the Netherlands Genomics Initiative, for financial support. Support from the Centre for Translational Molecular Medicine is appreciated. Continued collaborative support from Thermo Fisher, Bremen is acknowledged.

1. McLafferty FW (2011) A century of progress in molecular mass spectrometry. *Annu Rev Anal Chem* 4:1–22.
2. Cox J, Mann M (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem* 80:273–299.
3. Sadygov RG, Cociorva D, Yates JR, 3rd (2004) Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat Methods* 1:195–202.
4. Pappin DJ, Hojrup P, Bleasby AJ (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 3:327–332.
5. Zubarev RA, Kelleher NL, McLafferty FW (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J Am Chem Soc* 120:3265–3266.
6. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci USA* 101:9528–9533.
7. Toorn van de HWP, Mohammed S, Gouw JW, Breukelen van B, Heck AJR (2008) Targeted SCX based peptide fractionation for optimal sequencing by collision- and electron transfer-induced dissociation. *J Proteomics Bioinform* 1:379–388.
8. Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC (2007) Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry. *Science* 316:280–285.
9. Organ CL, et al. (2008) Molecular phylogenetics of mastodon and Tyrannosaurus rex. *Science* 320–499.
10. Spengler B (2004) De novo sequencing, peptide composition analysis, and composition-based sequencing: A new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. *J Am Soc Mass Spectrom* 15:703–714.
11. Frank A, Pevzner P (2005) PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal Chem* 77:964–973.
12. Mo L, Dutta D, Wan Y, Chen T (2007) MSNovo: A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal Chem* 79:4870–4878.
13. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
14. Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA (2005) Proteomics-grade de novo sequencing approach. *J Proteome Res* 4:2348–2354.
15. Chi H, et al. (2010) pNovo: De novo peptide sequencing and identification using HCD spectra. *J Proteome Res* 9:2713–2724.
16. Bertsch A, et al. (2009) De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis* 30:3736–3747.
17. Datta R, Bern M (2009) Spectrum fusion: Using multiple mass spectra for de novo peptide sequencing. *J Comput Biol* 16:1169–1182.
18. Marekov LN, Steinert PM (2003) Charge derivatization by 4-sulfophenyl isothiocyanate enhances peptide sequencing by post-source decay matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J Mass Spectrom* 38:373–377.
19. Taouatas N, Drugan MM, Heck AJ, Mohammed S (2008) Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase. *Nat Methods* 5:405–407.
20. Taouatas N, et al. (2009) Strong cation exchange-based fractionation of Lys-N-generated peptides facilitates the targeted analysis of post-translational modifications. *Mol Cell Proteomics* 8:190–200.
21. Hennrich ML, et al. (2009) Effect of chemical modifications on peptide fragmentation behavior upon electron transfer induced dissociation. *Anal Chem* 81:7814–7822.
22. Swaney DL, et al. (2007) Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal Chem* 79:477–485.
23. van Breukelen B, et al. (2010) LysNDeNovo: An algorithm enabling de novo sequencing of Lys-N generated peptides fragmented by electron transfer dissociation. *Proteomics* 10:1196–1201.
24. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4:207–214.
25. Molina H, Matthiesen R, Kandasamy K, Pandey A (2008) Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Anal Chem* 80:4825–4835.
26. Hackett SJ, et al. (2008) A phylogenomic study of birds reveals their evolutionary history. *Science* 320:1763–1768.
27. Vallender EJ, Mekel-Bobrov N, Lahn BT (2008) Genetic basis of human brain evolution. *Trends Neurosci* 31:637–644.
28. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74:5383–5392.
29. Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 4:787–797.
30. Geer LY, et al. (2010) The NCBI BioSystems database. *Nucl Acids Res* 38:D492–496.
31. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.
32. Choudhary C, et al. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* 325:834–840.
33. Weinert BT, et al. (2011) Proteome-wide mapping of the Drosophila acetylome demonstrates a high degree of conservation of lysine acetylation. *Sci Signal* 4:ra48.
34. Mayya V, et al. (2009) Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci Signal* 2:ra46.
35. Kim S, Bandeira N, Pevzner PA (2009) Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Mol Cell Proteomics* 8:1391–1400.
36. Ledvina AR, et al. (2009) Infrared photoactivation reduces peptide folding and hydrogen-atom migration following ETD tandem mass spectrometry. *Angew Chem Int Ed Engl* 48:8526–8528.
37. Samgina TY, et al. (2008) De novo sequencing of peptides secreted by the skin glands of the caucasian green frog *Rana ridibunda*. *Rapid Commun Mass Spectrom* 22:3517–3525.
38. Pevzner PA, Kim S, Ng J (2008) Comment on "Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry". *Science* 321:1040 author reply 1040.
39. Taouatas N, Heck AJ, Mohammed S (2010) Evaluation of metalloendopeptidase Lys-N protease performance under different sample handling conditions. *J Proteome Res* 9:4282–4288.
40. Mischerikow N, Altelaar AF, Navarro JD, Mohammed S, Heck AJ (2010) Comparative assessment of site assignments in CID and electron transfer dissociation spectra of phosphopeptides discloses limited relocation of phosphate groups. *Mol Cell Proteomics* 9:2140–2148.
41. Hubbard TJP, et al. (2009) Ensembl 2009. *Nucl Acids Res* 37:D690–D697.
42. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444–2448.
43. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
44. Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166.