DOI 10.1002/pmic.200700003 *Proteomics* 2007, *7*, 2318–2329

RESEARCH ARTICLE

Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated *de novo* sequencing

Patrice Waridel[1]*, *Ari Frank*[2], *Henrik Thomas*[1], *Vineeth Surendranath*[1], *Shamil Sunyaev*[3], *Pavel Pevzner*[2] *and Andrej Shevchenko*[1]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[2] Department of Computer Science and Engineering, University of California, San Diego, CA, USA
[3] Birgham & Women's Hospital and Harvard Medical School, Boston, MA, USA

LC-MS/MS analysis on a linear ion trap LTQ mass spectrometer, combined with data processing, stringent, and sequence-similarity database searching tools, was employed in a layered manner to identify proteins in organisms with unsequenced genomes. Highly specific stringent searches (MASCOT) were applied as a first layer screen to identify either known (*i.e.* present in a database) proteins, or unknown proteins sharing identical peptides with related database sequences. Once the confidently matched spectra were removed, the remainder was filtered against a non-annotated library of background spectra that cleaned up the dataset from spectra of common protein and chemical contaminants. The rectified spectral dataset was further subjected to rapid batch *de novo* interpretation by PepNovo software, followed by the MS BLAST sequence-similarity search that used multiple redundant and partially accurate candidate peptide sequences. Importantly, a single dataset was acquired at the uncompromised sensitivity with no need of manual selection of MS/MS spectra for subsequent *de novo* interpretation. This approach enabled a completely automated identification of novel proteins that were, otherwise, missed by conventional database searches.

Received: January 2, 2007
Revised: March 19, 2007
Accepted: April 2, 2007

Keywords:
de novo sequencing / LC-MS/MS / MS BLAST / organisms with unknown genomes / sequence-similarity searches

1 Introduction

In bottom-up proteomics, proteins are digested in-solution or in-gel with proteolytic enzymes and recovered peptides fragmented in nano-ESI MS/MS, -LC-MS/MS, or -LC-MALDI experiments (reviewed in ref. [1–5]). Uninterpreted tandem mass spectra are subsequently searched against protein, or translated DNA, databases using dedicated software (reviewed in ref. [6–8]). Regardless of the employed algorithm, the software identifies peptide sequences (and, hence, corresponding database entries [9]) by comparing peaks in MS/MS spectra with m/z (or m/z and intensities) of peptide fragments, which have been precomputed from database sequences, assuming some enzyme cleavage specificity and peptide fragmentation models [10–13]. While this is an efficient approach towards the characterization of known proteins (*i.e.*, proteins, whose sequences are accu-

Correspondence: Dr. Andrej Shevchenko, Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany
E-mail: shevchenko@mpi-cbg.de
Fax: +49-351-210-2000

Abbreviations: **IT**, ion trap; **mgf**, MASCOT generic format; **MS BLAST**, MS driven BLAST; **XIC**, extracted ion current

* Current address: Protein Analysis Facility, University of Lausanne, 1015 Lausanne, Switzerland

rately represented in a database resource), the identification of proteins unrepresented in a database (further termed as unknown proteins) remains a challenging problem (reviewed in ref. [14, 15]). Conventional database searches could, in principle, recognize peptides that are identical both in the unknown protein and known homologous protein(s) from closely related species [16]. This, however, is not efficient for identifying proteins that either originate from species phylogenetically distant from corresponding reference organisms, or belong to poorly conserved protein families, or were collected from species with an undefined genetic background and manifest strong sequence polymorphism (see ref. [17–21] for representative case studies).

Sequence-similarity search is a powerful tool for the identification of proteins from organisms with unsequenced genomes [22–27]. Full peptide sequences, or short sequence stretches (also termed sequence tags) [28], could be deduced directly from MS/MS spectra with no recourse to database resources (reviewed in ref. [29]). Multiple peptide sequences (or sequence tags) are then searched against a database in an error-tolerant fashion and, in this way, even proteins having only marginal sequence similarity to reference database entries could be identified [23, 26, 27].

Dedicated sequence-similarity search engines, such as CIDentify [30], an MS tailored version of gapped BLAST [25], an MS driven BLAST (MS BLAST) [22], FASTS [24], MS-Homology [31], OpenSea [32], among others, have been successfully applied in various proteomics studies. They, however, most efficiently handle queries composed of a relatively small number (typically, 5–50) of peptide sequences typically deduced from the spectra of "hand-picked" precursors, which is nowhere close to the total number of informative MS/MS spectra produced by LC-MS/MS under data-dependent acquisition control [33, 34]. Thus, the entire approach usually encompasses a selection of relatively abundant peptides that often originate from most conserved and well-characterized protein families and, therefore, are readily identifiable by conventional means.

Furthermore, usually only a small fraction of acquired spectra matches target proteins, while the rest of the spectra is attributed to chemical [35] or peptide [36, 37] backgrounds. Computational methods have been developed to recognize background spectra by comparison with a reference library or with spectra from a blank LC-MS/MS run [38, 39]. While due to their high specificity, conventional database searches largely tolerate peptide background, it strongly impairs sequence-similarity identifications [40]. Hence, rapid data-dependent acquisition of MS/MS spectra has dramatically increased the success rate of conventional protein identification, since it has become possible to sequence many more peptides in a single LC-MS/MS run. However, because of exactly the same reason, it undermined the performance of sequence-similarity searches.

Here, we present an approach for the seamless integration of automated *de novo* sequencing and LC-MS/MS into a protein characterization pipeline. A single dataset of tandem mass spectra was acquired at the uncompromised sensitivity, and then a combination of database searching and spectra processing software tools was employed for its in-depth analysis. In a few case studies based on LC-MS/MS analyses, each producing *ca.* 6000 of low resolution linear ion trap (IT) tandem mass spectra, we demonstrated how conventional (stringent) searches by MASCOT [41], *de novo* interpretation of spectra by PepNovo [42], and sequence-similarity searches by MS BLAST [16, 22], bundled by simple data handling scripts, enabled the comprehensive interpretation of datasets acquired from digests of unknown proteins.

2 Materials and methods

2.1 Chemicals

Cleland's reagent (DTT) was obtained from Merck (Darmstadt, Germany) and other chemicals from Sigma-Aldrich (Munich, Germany). Modified porcine trypsin (*Trypsin Gold* grade) was purchased from Promega (Mannheim, Germany). HPLC solvents (*Lichrosolv* grade), formic acid and TFA were purchased from Merck.

2.2 Protein preparations

Protein samples were obtained in on-going characterization of the proteome of *Dunaliella salina*, a halotolerant unicellular green alga [20] (Katz, A., Waridel, P., Shevchenko, A., Pick, U., *Mol. Cell. Proteomics* 2007, in press). Protein spots, visualized by CBB R250 staining, were excised from Blue Native gels [43] and in-gel digested with trypsin as described previously [44].

2.3 Analysis by LC-MS/MS

Dried peptide extracts were redissolved in 15–25 μL of 0.05% v/v TFA, depending on the staining abundance of the protein spots, and 4 μL was loaded using a FAMOS autosampler on a nano-LC-MS/MS Ultimate system (Dionex, Amstersdam, The Netherlands) interfaced on-line to a linear IT LTQ mass spectrometer (Thermo Fisher Scientific, San Jose, CA). The mobile phase was composed of 95:5 H_2O:ACN v/v with 0.1% formic acid (solvent A) and 20:80 H_2O:ACN v/v with 0.1% formic acid (solvent B). Peptides were first loaded onto a trapping microcolumn C18 PepMAP100 (1 mm × 300 μm id, 5 μm, LC Packings) in 0.05% TFA at a flow rate of 20 μL/min. After 4 min, they were back-flush eluted and separated on a nanocolumn C18 PepMAP100 (15 cm × 75 μm id, 3 μm, LC Packings) at a flow rate of 200 nL/min in the mobile phase gradient: from 5 to 20% of solvent B in 20 min, 20–50% B in 16 min, 50–100% B in 5 min, 100% B during 10 min, and back to 5% B in 5 min; % B refers to the solvent B content in an A + B mixture.

Peptides were infused into the mass spectrometer *via* a dynamic nanospray probe (Thermo Fisher Scientific) and analyzed in positive mode. Uncoated needles Silicatip,

20 μm id, 10 μm tip id (New Objective, Woburn, MA) were used with a spray voltage of 1.8 kV, and the capillary transfer temperature was set to 200°C. In a typical data dependent acquisition cycle controlled by Xcalibur 1.4 software (Thermo Fisher Scientific), the four most abundant precursor ions detected in the full MS survey scan (*m/z* range of 350–1500) were isolated within a 4.0 amu window and fragmented. MS/MS fragmentation was triggered by a minimum signal threshold of 500 counts and carried out at the normalized collision energy of 35%. Spectra were acquired under automatic gain control (AGC) in one microscan for survey spectra and in three microscans for MS/MS spectra with a maximal ion injection time of 100 ms *per* each microscan. *M/z* of the fragmented precursors were then dynamically excluded for another 60 s. No precompiled exclusion lists were applied.

From *raw* files, MS/MS spectra were exported as *dta* (text format) files using BioWorks 3.1 software (Thermo Fisher Scientific) under the following settings: peptide mass range, 500–3500; minimal total ion intensity threshold, 1000; minimal number of fragment ions, 15; precursor mass tolerance, 1.4 amu; group scan, 1; minimum group count, 1.

We note that BioWorks 3.1 software named each *dta* file according to the original name of the *raw* file, the scan number and the assumed charge of the precursor ion. If the low mass resolution in the normal scan mode did not allow determination of the precursor charge state, then redundant *dta* files were created that only differed by the assumed precursor charge.

2.4 Database searches

Dta files were merged into a single MASCOT generic format (*mgf*) file and searched against a MSDB database (updated May 15, 2005, containing 2 011 572 entries), or a *D. salina* EST database (downloaded from NCBI, updated March 16, 2006, containing 3998 entries) by MASCOT v. 2.1 software (Matrix Science, London, UK) installed on a local 2 CPU server. Tolerance for precursor and fragment masses was 2.0 and 0.5 Da, respectively; instrument profile: ESI-Trap; fixed modification: carbamidomethyl (cysteine); variable modification: oxidation (methionine).

2.5 Filtering MS/MS spectra

Prior to batch *de novo* sequencing of individual spectra, *mgf* files were filtered to remove spectra originating from common peptide and nonpeptide backgrounds. Briefly, the applied filtering routine comprised three main components: a similarity measure between MS/MS spectra, a search algorithm and a statistical framework to identify significant matches between the compared spectra. The similarity measure was a normalized Pearson correlation [45] of the intensities corresponding to fragment ions with matching *m/z*. Given the precompiled background library (see below) and the assumed similarity measure, the statistical framework was based on a well studied "extreme value problem" [45]. Accordingly, a measure of the statistical confidence of a match between a sample spectrum and the corresponding spectrum from the background library, was a double exponential function fit to data obtained in a simulation experiment by way of considering the distribution of similarity scores for *ca* 2000 nonbackground *dta* files obtained from high quality spectra of *Saccharomyces cerevisiae* peptides. The filtering software was implemented in the Python programming language (Ubuntu 5.10 Linux Operating System, Python 2.3) and is available from the authors upon request.

During the filtering process, all spectra from singly charged precursors and "void" spectra (typically containing less than ten fragment ions, whose relative intensities were above 3% of the base peak intensity) were removed. The remaining spectra were screened against a background library containing, in total, 13 000 MS/MS spectra that were pooled together from several LC-MS/MS analyses of in-gel digests of blank gel pieces. Additionally, more than 800 spectra of common trypsin and keratin peptides, identified by MASCOT or MS BLAST searches, were handpicked from various LC-MS/MS runs and added to the library. To limit the library redundancy, each new spectrum was first screened against the current library and added only if no matching spectrum was recognized. For each sample spectrum, the filtering software at first identified, in the background library, the spectra that were acquired from precursors with the same (within 2.5 amu tolerance) mass. Then, each pair of spectra was independently examined by comparing both *m/z* (within 0.5 amu tolerance) and intensities of their fragment ions. The quality of each match was scored and the scores compared to the threshold expected for randomly matching spectra. Spectra having statistically significant similarities to corresponding library spectra were removed, along with their redundant variants that assumed alternative charge states of the same precursor. Note that the described procedure does not rely on any assumed identity of spectra, preliminary knowledge of their origin, or database sequence resources.

2.6 *De novo* sequencing and sequence-similarity searches

A basic version of *de novo* sequencing software PepNovo [42] was modified to produce several, maximally complete, albeit redundant, degenerate, and partially inaccurate sequence candidates *per* each interpreted spectrum. An MS BLAST compatible version of PepNovo is available on the UCSD Computational MS Research Group server at http://peptide.ucsd.edu. PepNovo estimated the *de novo* interpretation confidence by assigning a sequence quality score, which reflected the expected number of correct amino acids in the top sequence candidate. All sequences, whose scores exceeded a user-defined threshold (typically, the value of 5.0) [46], were merged into a single query string and submitted to an MS BLAST (MS driven BLAST) search against nr database (nrdb95) at the web-accessible server (http://genetics.bwh.harvard.edu/msblast/).

2.7 Confidence of database searching hits

In MASCOT searches against an MSDB database, protein hits were considered confident if database entries were matched with at least two peptides and a total peptide ions score of 100, while all individual peptides with scores below 20 were disregarded. Because of the small size of the *D. salina* EST database, we accepted as confident hits identified by at least one peptide whose ions score was above 40. The threshold ions scores suggested by MASCOT for confident single peptide identifications in MSDB and EST databases, were 53 and 31 ($p<0.05$), respectively. Nonetheless, all single-peptide hits (regardless of their peptide ion scores, even if they exceeded the threshold of 53), and hits matched by a few lower scoring peptides were considered borderline and independently validated as described in ref. [46]. In sequence-similarity searches, the statistical significance of hits was evaluated according to MS BLAST scoring scheme [16]. However, only HSPs with a score of 62 or above were considered. Borderline MASCOT hits independently validated by PepNovo/MS BLAST [46] were also accepted.

3 Results and discussion

3.1 The workflow

A proteomics workflow, integrating spectra processing with conventional and sequence-similarity searches in a layered manner [15], is outlined in Fig. 1. Stringent searches with the highest discriminating power were executed first and, apart from producing hits, also helped in rectifying the target spectral dataset for less specific sequence-similarity searches, as explained below.

Tandem mass spectra, acquired in a LC-MS/MS run on a linear IT LTQ instrument, were exported as individual files in *dta* format. A partially redundant combined pool of *dta*'s was converted into a single *mgf* file and submitted to the MASCOT search. If confident hits were produced, a special script removed all *dta* files pertinent to matched MS/MS spectra, including redundant *dta* file(s) with alternative charges of the same precursors. The remaining *dta's* representing spectra that, for some reason, did not match database sequences, were filtered against a library of background spectra, which was precompiled from *ca.* 13 000 non-annotated spectra acquired in several independent LC-MS/MS analyses of blank in-gel digests. Filtering was set to eliminate background spectra in a completely unbiased fashion, regardless of whether they match any database sequences, and to retain all spectra, other than recognized background. Typically, more than 75% of the total number of *dta*'s were removed and, at this stage, no quality assessment of spectra was performed. Filtering stringency was controlled by a user-selected *p*-value threshold, defined here as a probability of finding no similarity to any spectrum in the background library. Since matching of nonannotated sample and

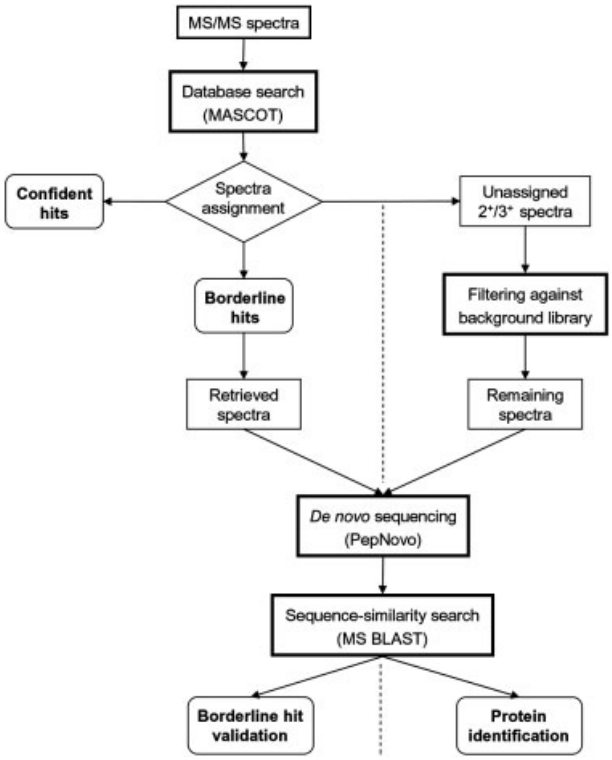


Figure 1. Layered protein identification workflow. After MASCOT search against a full MSDB database, confident hits were identified with a protein score $\geq$100 and number of peptides $\geq$2. Cross-species hits matching one or more peptides with a minimum score of 40 were considered borderline and were subsequently validated by *de novo* sequencing and MS BLAST searches. All spectra, excluding those acquired from singly charged precursors or already assigned to confident MASCOT hits, were filtered against a library of nonannotated background spectra, sequenced *de novo* and obtained candidate sequences submitted to MS BLAST searches.

background spectra is a probabilistic process, a certain fraction of genuine spectra could be falsely recognized as background and, subsequently, removed. Therefore, in the protein identification workflow (Fig. 1) filtering was employed downstream of stringent MASCOT searches, since, because of their higher specificity, the results were less affected by queried background spectra.

Because of a relatively small size, the pool of remaining spectra could be either submitted to MASCOT error-tolerant search (for example, assuming no enzyme cleavage specificity restrictions), or sequenced *de novo* in a batch mode by PepNovo software – this option was mainly used in this work. PepNovo interprets a single *dta* file in *ca.* 0.15 s, and assigns a quality score for the top sequence out of seven produced sequence candidates. Computational simulation experiments suggested the quality score of 5.0 as a conservative cut-off (Fig. 3 and ref. [46]). We note, however, that the threshold should be used with caution since the sequence quality score has not been normalized to the pep-

tide length and, aiming at *a priori* high scores would discriminate short, yet accurately sequenced peptides. Insufficient quality of MS/MS spectra is not the only reason for low scoring interpretations. PepNovo was optimized for sequencing doubly charged precursor ions and is less efficient in sequencing triply charged precursors. Therefore, PepNovo sequence quality scores could not substitute spectra quality estimates developed for conventional stringent searches [47, 48] since they are biased towards interpretable (rather than abundant) patterns of fragment ions. All sequence candidates selected from all interpreted MS/MS spectra were merged into a single query string and submitted for MS BLAST sequence-similarity search. Typically, these searches either revealed new proteins, or detected new peptides from proteins already identified by MASCOT.

In parallel, PepNovo and MS BLAST were also used for validating borderline hits produced by MASCOT searches as described previously [46, 49]. Briefly, if *de novo* sequencing by the MS BLAST search independently hit the same peptide, as was previously found by the MASCOT search, it was considered as positively identified.

3.2 Seamless integration of *de novo* sequencing with MS BLAST searches

An efficient and fast *de novo* sequencing program tailored for interpreting low-resolution MS/MS spectra is a key element of the proposed workflow. PepNovo software was particularly successful in interpreting IT spectra and scored favorably when compared to other *de novo* sequencing programs [42, 50]. Importantly, PepNovo interprets a single MS/MS spectrum in *ca*. 0.15 s and, in batch mode, operates with spectra in *mgf* format, which simplifies the processing of LC-MS/MS data. In this work, we used a new version of PepNovo, which was tailored for interpreting MS/MS spectra obtained on linear IT instruments and whose output format conforms to MS BLAST conventions [16, 40] (Fig. 2).

Next, we asked if the parallel consideration of multiple, partially redundant candidate sequences produced by Pep Novo interpretation of the same MS/MS spectrum increased the success rate of MS BLAST identifications, compared to nonerror-tolerant searches with only the top candidate sequence – a frequently used method of cross-species protein identification. To this end, a simulation dataset was compiled from 71 high quality MS/MS spectra. Each of them unequivocally hit a single peptide sequence of, on average, 14 amino acid residues upon a MASCOT database search with the ions score above the value of 63. In each spectrum, peaks with relative intensities below 1% of the base peak intensity were declared noise and their absolute intensity left unchanged, whereas the absolute intensity of other fragment peaks was gradually reduced [46]. All unmodified and computationally manipulated spectra were sequenced *de novo*. Up to seven candidate sequences, along with the corresponding PepNovo quality score, were registered for each interpreted spectrum. The entire range of quality scores was divided into five bins (Fig. 3) and spectra, interpreted with the corresponding scores, were picked for each score bin. The full dataset used for the simulation comprised 71 unmodified spectra and their 131 computationally altered "clones". Note that we used "spoiled clones" of initially high quality spectra, rather than some native low scoring spectra since, for each sequenced spectrum, we should precisely know the "source" peptide sequence to judge if its *de novo* interpretation was correct. Within each score bin, we determined fractions of interpreted spectra, in which (i) the top candidate sequence was complete and correct; (ii) one of the listed candidate sequences was complete and correct; and (iii) candidate sequence(s) contained a noninterrupted stretch of at least eight accurate amino acid residues (here termed as a tag). Finally, we combined all candidate sequences produced from a given spectrum into MS BLAST query string, searched it against a comprehensive database and checked if the correct "source" peptide sequence was confidently hit. The distribution of fractions of *de novo* interpreted spectra were plotted separately for each quality score bin (Fig. 3).

Figure 3 suggests that considering more candidate sequences and their error-tolerant matching to database entries improved the success rate of protein identification, although, regardless of the used method of sequence matching, it was below 20% for low scoring candidate sequences. Interestingly, even highly scored top sequences were seldom fully accurate. Since quality scores were not normalized to the full peptide length (see the discussion above), higher scores were expected for larger peptides, although their spectra are usually complex and (in case of IT) are affected by low m/z cut-off. Therefore, their full interpretation is seldom completely accurate and it is not surprising that a smaller fraction of these spectra also produced fully accurate tags. At the same time, a few miscalled amino acid residues within a largely accurate long peptide sequence did not affect the identification performance of MS BLAST.

We note that sequences with high quality scores might also be obtained from low abundant peptide precursors or from spectra with abundant chemical noise. The TIC and TIC fraction reported by PepNovo for each interpreted spectrum could be used, with some caution, either for preselecting the most representative precursors or for postsearch validation of assignments based on sequence alignments of the borderline confidence. Note that reported TIC does not directly reflect the precursor abundance, since the spectrum might not be taken at the apex of the chromatographic peak. There is, nevertheless, some correlation between spectrum TIC and precursor intensity, since, because of statistical sampling, a more abundant precursors would more likely produce spectra with higher TIC.

3.3 Unbiased filtering of background MS/MS spectra is essential for sequence-similarity identifications

Removing background spectra prior to batch *de novo* sequencing is important for the successful identification of

1. Peptide MW	2. Scan number .charge	3. TIC	4. TIC fraction	5. Score	6. Candidate sequences
2068.48	2748.2748.2	57899	0.628	15.4	-BSMGLSNSESVDVDATSDLLK-BSMGLSGGSESVDVDATSDLLK -BXXXLSNSESVDVWTSDLLK-BXXLSNSESVDVWTSDLLK -BXXGLSNSESVDVWTSDLLK
2066.82	2750.2750.2	74794	0.611	15.3	-BFAGLSNSESVDVDATSDNLK-BM+16AGLSNSESVDVDATSDNLK -BFAGLSNSESVDVDATSDGGLK-BM+16AGLSNSESVDVDATSDGGLK -BFQLSNSESVDVDATSDNLK
2699.87	2760.2762.2	66717	0.595	15.1	-CSADSCAGGALNNAYGGAA-CSADSCAGGALNGGAYGGAA -CSADSCAGGALGGNAYGGAA-CSADSCANALNNAYGGAA -CSADSCANALGGNAYGGAA
2134.73	3004.3004.2	73604	0.513	14.9	-BXXXXXSSTDLGWSYYGVALTK-BXXXXSSTDLGWSYYGVALTK -BXXXSSTDLGWSYYGVALTK-BXXXXXSSTDLGXXSYYGVALTK -BXXXXSSTDLGXXSYYGVALTK
2016.24	2243.2243.2	12035	0.632	14.8	-APLDDGGLTNTEAGDVK-APLDDGGLTGGTEAGDVK -APLDDGGLTNTEQDVK-APLDDGGLTGGTEQDVK-APLDDGGL
1837.52	3592.3592.2	43769	0.589	14.2	-BSSEPVVGGGGQLM+16TAM+16ER-BSSEPVVGGGGQLFTAM+16ER -BSSEPVVGGGGQLM+16TAFER-BSSEPVVGGGGQLFTAFER -BSSEPVVGGGGGALM+16TAM+16ER
2277.45	3172.3172.2	47011	0.717	14.2	-M+16NAQYNTLDGTDVSM+16K-M+16NAQYNTLDGTDVSFK -FNAQYNTLDGTDVSM+16K-M+16NAQYGGTLDGTDVSM+16K -M+16NAQYGGTLDGTDVSFK
2017.33	2233.2233.2	16534	0.591	14.2	-APLDDGGLTNTEAGDVK-APLDDGGLTGGTEAGDVK -APLDDGGLTNTEADGVK-APLDDGGLTGGTEADGVK -APLDDGGLTGGTEQDVK
2081.80	3153.3153.2	46632	0.581	14.1	-BXXXXDLTVVGGNQLFAAYER-DLTVVGGNQLFAAYER -DLTVVGGNGALFAAYER-DLTVVGGGGQLFAAYER -DLTVVGGGGGALFAAYER
2133.51	3009.3009.2	79830	0.524	13.9	-LSTDLGWSYYGVALTK-LSTDLGXXSYYGVALTK-LSTDLGWSYYRALTK -LSTDLGXXSYYRALTK-LSTDLGWS

with B = R or K (tryptic cleavage)
Z = Q or K
L = L or I
M+16 = methionine oxidation
X = undetermined amino-acid

Figure 2. An example of the output of batch *de novo* interpretation of MS/MS spectra by PepNovo. For each interpreted spectrum, PepNovo reported: intact mass of the fragmented peptide (column 1); name of the *dta* file, including the scan number and assumed charge (column 2); TIC of the MS/MS spectrum (sum of fragment absolute intensities) (column 3); TIC fraction covered by expected fragments of the top candidate sequence (column 4); sequence quality score, representing the expected number of correct amino acids in the top (first) candidate sequence (column 5); candidate sequences (typically, limited to seven) (column 6). Note that F and M (oxidized) residues, as well as Q and K, are isobaric and are not distinguishable in low-resolution MS/MS spectra. In these instances, optional sequences were included into the query. All candidate sequences obtained by interpreting all submitted MS/MS spectra (column 6) with the quality score (column 5) that typically exceeded the value of 5.0 were merged into a single MS BLAST query string and searched against the nr database. MS BLAST web interface disregards all numbers and nonconventional symbols and, therefore, the entire output (or only the selected sequences) can be directly pasted into the query window.

proteins by sequence-similarity searches at the high sensitivity. The characterization of a Coomassie stained spot with an apparent molecular weight of 55 kDa, which contained membrane proteins from *D. salina*, is presented here as a representative example. The spot was excised from a 2-D Blue Native gel, in-gel digested with trypsin and the recovered peptides sequenced by LC-MS/MS. First, all MS/MS spectra were interpreted *de novo* and the entire pool of candidate sequences was submitted to an MS BLAST search (Fig. 4A). The search produced a multitude of formally confident alignments with typical background proteins, mainly keratins, trypsins, and serine proteases from a variety of species. In the output, the three top nonbackground proteins were only listed at 137th, 190th, and 233rd positions.

Enlarging the query with sequences, apparently unrelated to target proteins, does not directly affect the scores of reported HSPs. This, however, increases the significance threshold scores [16] and, hence, indirectly impairs the confidence of hits with only a few aligned HSPs. Note, that keratin sequences are rich in low complexity regions that are also common in many proteins in a database. Therefore, sequence-similarity searches with keratin sequences produce a large number of formally confident hits that are, at first glance, apparently unrelated to keratins.

The same pool of spectra was then processed according to the workflow shown in Fig. 1. Upon the MASCOT search, spectra from matching peptides (including peptides from keratins and trypsin) were removed and the remainder was

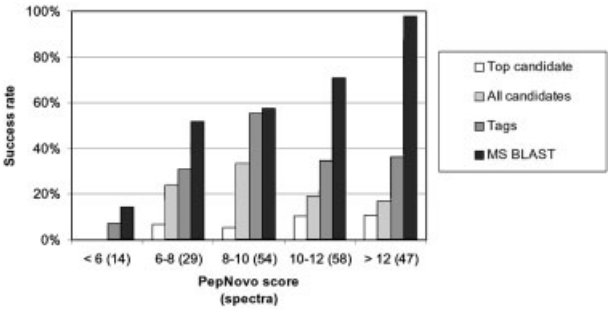


Figure 3. Computational simulation of the identification success rate of *de novo* sequencing combined with MS BLAST searches. Bars represent fractions of spectra, in which PepNovo produced, with the sequence quality score within the specified range, a complete and accurate peptide sequence as a top candidate (top candidate); or it was found among all produced candidates (all candidates); or a stretch of at least eight correctly called amino acid residues was found (Tags); or the corresponding peptide was hit by MS BLAST upon error-tolerant database searching (MS BLAST). The total number of spectra within each score bin is specified in parenthesis and was assumed to be 100%. Note that the sum of corresponding fractions can exceed 100%, since, for example, all spectra with accurately predicted sequences ("top candidates" and/or "all candidates") were also identified by MS BLAST.

filtered against a background library. A preprocessed dataset, comprising only 16% of the total number of *dta* files, was sequenced *de novo* and the obtained peptide sequence candidates were submitted to the MS BLAST search (Fig. 4B). Filtering largely eliminated background proteins and, therefore, nonbackground protein hits were observed at the 1st, 2nd, 7th, and 9th positions in MS BLAST output. Only 12 hits were reported as confident and only five of them were, apparently, related to nonremoved trypsin precursors. The top hit was a 48 kDa membrane protein from a closely related alga *Chlamydomonas reinhardtii*, which was not identified by MASCOT since not a single sequenced peptide was identical between the two, apparently homologous, proteins (Fig. 4C). Another two top listed hits were chloroplast proteins from *Chlamydomonas* sp., which were found by MASCOT, yet also identified by MS BLAST with other peptide sequences, not present in a protein database. Interestingly, a few spectra originating from trypsin autolysis products still survived the filtering apparently because, at that moment, no matching spectra were present in the background library.

Why are low abundant trypsin/keratin contaminants that used to be essentially harmless for MASCOT searches, so detrimental to MS BLAST identifications? The reason is that

A

Color Key: red = positive hit; green = borderline hit; black = negative result

Summary: High Total Sequences producing High-scoring Segment Pairs:	High Score	Total Score
/:sptrembl|Q14664|Q14664 Keratin 10.//:trembl|J04029|HSKE...	99	2254
/:swiss|P13645|K1CJ_HUMAN Keratin, type I cytoskeletal 10...	99	2236
/:sptrembl|new|AAH63697|AAH63697 Keratin 1.//:tremblnew|BC...	126	2202
/:sptrembl|Q6IFW6|Q6IFW6 Type I keratin KA10.	96	2136
/:swiss|P06394|K1CJ_BOVIN Keratin, type I cytoskeletal VI...	96	2131
/:swiss|P35908|K22E_HUMAN Keratin, type II cytoskeletal 2...	91	2112
/:tremblnew|AY318944|AY318944_1 gene: "Ker10"; product: "...	96	2068
/:tremblnew|AY318943|AY318943_1 gene: "K2e"; product: "ep...	80	1981
/:trembl|L00193|MMKTEPI2_1 Mouse epidermal keratin type I...	90	1966
/:tremblnew|AY318945|AY318945_1 gene: "Ker1"; product: "e...	97	1914
/:trembl|M19156|HSKRT10A_1 gene: "KRT10"; Human acidic k...	99	1910
/:trembl|BT007974|BT007974_1 product: "Homo sapiens kerat...	109	1909
/:tremblnew|BC039169|BC039169_1 gene: "KRT16"; product: "...	109	1879
/:sptrembl|Q6IMF3|Q6IMF3 Type II keratin Kb1.	89	1846
/:swiss|Q61781|K1CN_MOUSE Keratin, type I cytoskeletal 14...	109	1833
/:trembl|V00830|MMKERA_1 Mouse mRNA encoding epiderm...	90	1816
/:sptrembl|Q6IFU9|Q6IFU9 Type I keratin KA16.	109	1794
/:sptrembl|Q6IFV1|Q6IFV1 Type I keratin KA14.	109	1785
/:sptrembl|Q9Z331|Q9Z331 Keratin 6 beta.//:trembl|AB01204...	82	1777
/:sptrembl|Q9EQD6|Q9EQD6 Keratin intermediate filament 16...	109	1763

B

Color Key: red = positive hit; green = borderline hit; black = negative result

Summary: High Total Sequences producing High-scoring Segment Pairs:	High Score	Total Score
/:sptrembl|Q75NZ1|Q75NZ1 Low-CO2 inducible protein LCIC./...	81	578
/:sptrembl|Q75NZ2|Q75NZ2 Low-CO2 inducible protein LCIB./...	87	393
/:swissnew|P00761|TRYP_PIG Trypsin precursor (EC 3.4.21.4...	87	293
/:trembl|X15505|HSTRYIII_1 Human mRNA for pancreatic tryp...	71	204
/:sptrembl|Q7Z5F4|Q7Z5F4 Protease serine 4 isoform B.//:t...	71	204
/:swissnew|P35030|TRY3_HUMAN Trypsin III precursor (EC 3....	71	204
/:swiss|P10898|PSBC_CHLRE Photosystem II 44 kDa reaction...	80	120
/:swiss|Q29463|TRY2_BOVIN Trypsin, anionic precursor (EC ...	74	120
/:swiss|P37255|PSBB_CHLRE Photosystem II P680 chlorophy...	94	94
/:sptrembl|Q7Q0E3|Q7Q0E3 AgCP9400 (Fragment).//:trembl|A...	76	76
/:sptrembl|Q748D6|Q748D6 UDP-N-acetylglucosamine--N-acet...	75	75
/:sptrembl|Q71KN6|Q71KN6 Photosystem II P680 chlorophyll ...	74	74
/:sptrembl|Q9GGY6|Q9GGY6 Photosystem II CP43 apoprotein...	70	110
/:swiss|P41624|PSBB_PINTH Photosystem II P680 chlorophyll...	72	72
/:swiss|P27200|PSBB_PROHO Photosystem II P680 chlorophy.l.	71	71
/:swissnew|P56307|PSBB_CHLVU Photosystem II P680 chloro...	71	71
/:sptrembl|Q82PR2|Q82PR2 Putative dehydrogenase.//:trembl...	70	70
/:sptrembl|Q8NHM4|Q8NHM4 Trypsinogen C.//:trembl|U66061...	66	154

C

^ = sptrembl|Q75NZ1|Q75NZ1 Low-CO2 inducible protein LCIC.//:tremblnew|AB168094|AB168094_1 gene: "LciC"; product: "low-CO2 inducible protein LCIC"; Chlamydomonas reinhardtii LciC mRNA for low-CO2 inducible protein LCIC, complete cds. //:trembl|AB168094|AB168094_1 gene: "LciC"; product: "low-CO2 inducible protein LCIC"; Chlamydomonas reinhardtii LciC mRNA for low-CO2 inducible protein LCIC, complete cds.
Length = 443

Total Score: 578

```
                    0      90     180    270    360
                    |      |      |      |      |     |443
sptrembl|Q75NZ1|Q75NZ |______________________________
Local hits (HSPs)   |      --  --  -  --    ---
```

```
Score = 81 (42.2 bits)
Identities = 11/14 (78%), Positives = 12/14 (85%)

Query: 1355 GFGFTGDNTLAFSN 1368
          GFGFTGDNT+A  N
Sbjct:  104 GFGFTGDNTIAMTN 117

Score = 75 (39.2 bits)
Identities = 10/14 (71%), Positives = 12/14 (85%)

Query:  169 FFPTALAVDDYMNR 182
          +FPTAL VDD+M R
Sbjct:   84 YFPTALGVDDFMAR 97

Score = 73 (38.2 bits)
Identities = 9/16 (56%), Positives = 13/16 (81%)

Query: 1117 BTLTDDLEHLLAKAVD 1132
          +T+ DDLE L+ KAV+
Sbjct:  267 RTISDDLEYLIEKAVN 282

Score = 65 (34.2 bits)
Identities = 10/15 (66%), Positives = 10/15 (66%)

Query: 2547 AVFTGVXXXNWAADL 2561
          AV TGV   NWAA L
Sbjct:  289 AVITGVEIHNWAAHL 303

Score = 59 (31.2 bits)
Identities = 6/10 (60%), Positives = 9/10 (90%)

Query: 3664 FSFPHLALDS 3673
          F FPH+A++S
Sbjct:  175 FAFPHIAINS 184
```

Figure 4. Unbiased filtering of background spectra enabled the MS BLAST identification of an unknown protein from *D. salina*. (**A**) The output of MS BLAST search with peptide sequence candidates produced by *de novo* interpretation of the entire pool of acquired tandem mass spectra (5511 *dta* files); (**B**) the fraction of *dta*'s unmatched by MASCOT was filtered against a background library. The remaining 874 *dta* files (16% of the total number of *dta*'s) were interpreted *de novo* and sequences with the quality score above the value of 5.0 (in total, 131 precursors) were submitted for MS BLAST search; (**C**) HSPs of the top hit from MS BLAST output in panel B. Only the most significant alignments are presented (HSPs with scores above 55). Note that this protein was not identified by MASCOT since its sequence is not in a database and none of its sequenced peptides was identical to known homologous proteins.

de novo interpretation of MS/MS spectra produces several (in this work we considered up to seven) partially redundant sequence candidates, essentially "cloning" the same sequence into multiple copies. MS BLAST engine treats them equally and tries to match them to different sequence segments of the same protein, or different proteins in a database. Since keratins are rich in low complexity sequence stretches, similarity searches with their multiple, partially different, variant sequences trigger an avalanche of hits, largely based on similarly looking HSPs. This, however, does not happen in MASCOT searches because a much higher stringency of the match is required and any deviation from the expected fragment patterns heavily penalized.

As shown in Fig. 4B, stringent (MASCOT) searches were unable to completely remove peptide spectra matching background proteins. In contrast, the identification-independent filtering, that is based on the rapid comparison of uninterpreted fragment ion patterns between the examined and library spectra, recognized and removed background spectra, regardless of whether they matched anything in a database. The representative examples included, but were not limited to, orifice fragmentation products, products of unspecific enzyme cleavage or polymorphic sequences. Continuous updating the library by adding newly recognized peptide spectra of trypsin, keratins, and other common (or, on user's request, experiment-specific) contaminants eventually should eliminate background spectra (almost) completely from any probed query. We note, however, that the number of falsely retained background spectra also depends on the user-defined filtering stringency. Under stringent filtering settings, less background spectra would be retained, albeit with a higher chance of losing *bona fide* spectra from target proteins due to their random similarity to background.

3.4 Validation of MASCOT cross-species identifications with borderline statistical confidence

Conventional proteomics methodologies are capable of cross-species identification of unknown proteins by matching identical peptides in known homologous proteins. However, such peptides are relatively rare and their identification typically relies on matching only a few peptide sequences, often with borderline statistical confidence. Here, we demonstrate how *de novo* sequencing and MS BLAST searches provided independent validation of borderline cross-species MASCOT hits.

In the above sample of *D. salina* proteins, the MASCOT search identified a plausible homologue of the ATP synthase from another alga, *Bigelowiella natans*. However, this identification relied upon a single exactly matching peptide (Fig. 5A) and, in line with current proteomics guidelines [51], it should be considered as borderline. To validate this hit, the *dta* file corresponding to the matched spectrum was then interpreted *de novo* (Fig. 5B) and partially redundant candidate sequences submitted to the MS BLAST search (Fig. 5C), which produced a statistically confident hit to the overlapping sequence stretch in a related database entry. We note that peptide sequences of the MASCOT hit and *de novo* candidates differed at their *N*-termini and, currently, it is not possible to judge which peptide sequence was correct, since the full sequence of *D. salina* protein remains unknown. This, however, did not affect the confidence of MS BLAST hit assignment, which relies upon an independent scoring scheme that only considers the local similarity of sequence stretches aligned within the HSP.

3.5 The characterization of a complex mixture of unknown proteins by a layered proteomics workflow

The workflow (Fig. 1) was applied to the complete analysis of a 55 kDa spot excised from a 2-D Blue Native gel separating *D. salina* membrane proteins (Table 1). Out of the ten proteins identified in this sample, five were identified by MASCOT searches. Two were known *D. salina* proteins (proteins 1 and 2 in Table 1) and another three were identified by statistically confident cross-species matches to multiple identical peptide sequences in proteins from related algal species (proteins 3, 4, and 5). Three more identifications were statistically borderline (proteins 6, 7, 8, see also Fig. 5) and were subsequently validated by a combination of PepNovo sequencing and MS BLAST searches, as described above. When all *dta*'s corresponding to matching peptides were removed and the rest was submitted to *de novo* sequencing followed by the MS BLAST search, we identified two more proteins (proteins 9 and 10), which were missed by the MASCOT search. Importantly, the examination of the corresponding extracted ion current (XIC) traces in LC-MS/MS profiles suggested that, in fact, protein 9 could be the major component of the mixture. Additionally, MS BLAST searches also revealed several new peptides from proteins already identified by MASCOT (proteins 3 and 4) thus improving the sequence coverage and confidence of cross-species identifications.

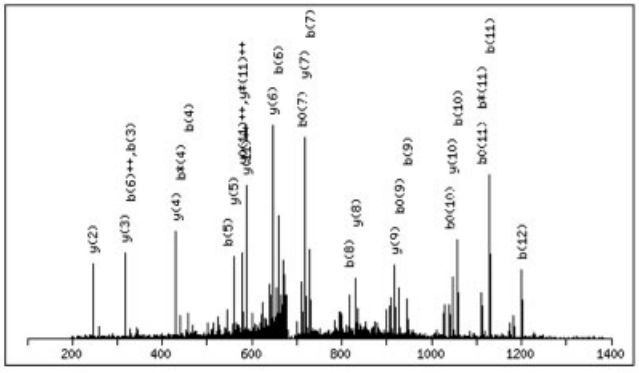
This analysis is representative of a series of 32 samples from *D. salina*, isolated by Blue Native gel electrophoresis (Table 2) and suggested the high relevance of the multi-layer datamining strategy, compared to previously reported identifications based on *de novo* sequencing by nano-ESI MS/MS [20]. Stringent (MASCOT) and sequence-similarity searches proved to be complementary identification tools and the latter increased the number of confident hits by more than 15%, including several major protein components in analyzed samples.

Out of a total of 75 proteins identified in 32 spots, 55 proteins were unique (Table 2). Eight proteins (15%) were identified by matching to known *D. salina* sequences. Confident cross-species identifications by MASCOT accounted for almost a half of all identified proteins (25 hits, 45%), matching sequences from algae (13 proteins), plants (3 proteins),

A

MS/MS Fragmentation of **SIQESLASELAAR**
Found in **Q7XYQ8_CHLS6**, ATP synthase gamma subunit.-
Chlorarachnion sp. (strain CCMP 621) (Pedinomonas minutissima).

Match to Query 2412: 1374.403448 from(688.209000,2+)
BN_9.3535.3535.2.dta
From data file D:\My Documents\Data\Dunalellia\DunBN\Dun_BN_III\Mascot\BN_9.mgf



Monoisotopic mass of neutral peptide Mr(calc): 1373.72
Fixed modifications: Carbamidomethyl (C)
Ions Score: 68 **Expect:** 0.0016
Matches: 29/132 fragment ions using 58 most intense peaks

B

Precursor MW	*Scan number .charge*	*TIC*	*TIC fraction*	*Score*	*Sequence candidates*
1374.42	3535.3535.2	159429	0.689	11.1	-BEAGAESLASELAAR-BEQAESLASELAAR

C

^ = sptrembl|Q7XYQ8|Q7XYQ8 ATP synthase gamma subunit.//:trembl|AY267622|AY267622_1 product: "ATP synthase gamma subunit"; Bigelowiella natans ATP synthase gamma subunit mRNA, complete cds; nuclear gene for plastid product. //:gp|AY267622|32307462 ATP synthase gamma subunit [Bigelowiella natans]
Length = 390

Total Score: 64

```
                          0     80    160   240   320
                          |     |     |     |     |     | 390
sptrembl|Q7XYQ8|Q7XYQ     |_____________________________
Local hits (HSPs)         |                           _

Score = 64 (33.5 bits)
Identities = 10/10 (100%), Positives = 10/10 (100%)

Query:   7  ESLASELAAR 16
            ESLASELAAR
Sbjct: 336  ESLASELAAR 345
```

Figure 5. *De novo* sequencing and an MS BLAST search validated a borderline cross-species hit produced by the MASCOT search. (**A**) A MASCOT search produced a candidate hit with one matched peptide to the protein from the related organism (**B**) The *dta* file corresponding to the spectrum in panel A was retrieved by a script and its *de novo* interpretation produced two candidate sequences with the quality score of 11.1; (**C**) They were merged into an MS BLAST query and the search hit the same peptide from the related alga *B. natans*. According to the MS BLAST scoring scheme, the hit was confident albeit the N-terminal piece of the sequence, for some reason, mismatched. Note that the N-terminal stretch of amino acid residues in all three variant sequences is isobaric and that the true sequence of the fragmented peptide is currently unknown.

bacteria (8 proteins) and fungi (1 protein). Sequence-similarity searches with MS BLAST added a substantial number of novel identifications (6 hits, 11%) by matching correspondent protein homologues from algae (3 proteins), plants (1 protein) and bacteria (2 proteins). Importantly, the examination of peptide signal intensities in XIC traces showed that two of them were, most likely, the major components of the samples. Finally, MASCOT searches against an EST database containing about 4000 *D. salina* sequences more than doubled the number of intra-species identifications (although relatively short EST sequences do not provide the functional annotation of hits directly). Additionally, validating MASCOT identifications by MS BLAST rescued one, otherwise a false negative, borderline hit and unequivocally rejected another 11 borderline hits.

These results demonstrate the efficiency of the workflow that combines stringent and sequence-similarity searches for the proteomic characterization of organisms with unsequenced genomes. The relative contribution of complementary database interrogation strategies depends on the species of origin, particularly on the number of sequences available in protein or EST databases, and the phylogenetic distance between the studied species and reference organism(s), completely or partially covered by genomic sequenc-

Table 1. Proteins identified in a single spot on a 2-D Blue Native gel separating *D. salina* membrane proteins

Hit number	Protein	Species	MW (kDa)	MASCOT score (peptides)	MS BLAST score (HSPs, score>55)	Highest XIC of matched peptides	Comments
1	High affinity nitrate transporter	*D. salina*	59	450 (10)		3E+05	Known *D. salina* proteins
2	Transferrin-like protein Ttf-1	*D. salina*	140	137 (4)		1E+05	
3	Photosystem II reaction centers CP47 apoprotein	*C. reinhardtii*	56	297 (6)	94(1)	6E+05	Cross-species MASCOT hits
4	Photosystem II chlorophyll a-binding protein psbC	*Chlamydomonas eugametos*	50	215 (5)	120(1)	5E+05	
5	P700 chlorophyll a-apoprotein A2	*Chlamydomonas moewusii*	82	149 (3)		7E+04	
6	Putative plastidic ATP/ADP-transporter (fragment)	*Prototheca wickerhamii*	24	72 (1)	*80(1)*	3E+05	Borderline hits validated by MS BLAST
7	ATP synthase gamma subunit	*B. natans*	42	68 (1)	*64(1)*	2E+05	
8	Probable multispanning membrane protein	*Arabidopsis thaliana*	74	44 (1)	*67 (1)*	8E+04	
9[a]	**Low-CO2 inducible protein LCIC**[a]	***C. reinhardtii***	**48**	**–**	**578(5)**	**2E+06**	New proteins identified by MS BLAST
10	UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide)pyrophosphoryl-undecaprenol *N*-acetylglucosamine transferase	*Geobacter sulfurreducens*	40	–	75(1)	9E+04	

a) The most abundant protein (as judged by corresponding XIC traces) is shown in bold.

Table 2. Overview of protein identifications by MASCOT and PepNovo–MS BLAST in the preparations of *D. salina* membrane proteins separated into 32 individual spots by Blue Native gel electrophoresis

Search	Species	Database	Identifications	
			Hits	(%)
MASCOT	*D. salina*	MSDB	8	15
		EST	15	27
	Other	MSDB	25	45
MASCOT/MS BLAST[a]		MSDB/ nrdb95	1	2
MS BLAST		nrdb95	6	11
Total			**55**	**100**

a) Borderline hits from MASCOT searches validated by *de novo* sequencing and MS BLAST.

ing [16]. It is important, however, that complementary database searches operate with the same dataset of MS/MS spectra and no adjustment of the data acquisition routine is required.

4 Concluding remarks

We demonstrated that a combination of a fast and accurate *de novo* sequencing software and MS BLAST searches enabled sequence similarity-driven proteomic interpretation of large LC-MS/MS datasets acquired on a rapid scanning, low mass resolution linear IT instrument. A layered database mining workflow improves substantially the characterization of proteomes of organisms with unsequenced genomes. Yet, we have reason to believe that it might also have important implications for proteomics in fully sequenced organisms, as it validates borderline hits produced by conventional database searches and has the potential for unbiased screening for PTMs, sequence polymorphism and unrecognized splicing variants.

It is important that, regardless of the availability of database sequences, a single LC-MS/MS dataset is always acquired and, subsequently, used for conventional (stringent) and sequence-similarity searches. There is no need for chemical derivatization or isotopic labeling of analyzed peptides, or for repetitive LC-MS/MS analysis under specific settings, which, for example, would target the data acquisition at the most abundant ions or employ zoom scans. Once acquired, the complete spectral dataset can be post-processed

according to the user's needs. More interpretation or data processing layers could be included, such as, for example, searches against species-restricted databases, or searches with an alternative set of variable PTMs, or screening against a spectra library produced in specific control experiments. Whenever the search produces confident hits, the corresponding *dta*'s should be subtracted, thus compacting the query down to essential and informative spectra that do not match anything in a database in a stringent manner and, hence, require error-tolerant interpretation. The entire data processing routine could be automated and integrated into any proteomics pipeline adopted in the laboratory.

In the rapidly evolving field of proteomics, it is important that data interpretation pipelines maintain a modular organization that utilizes a common data format and allows unrestricted and independent operations with individual MS/MS spectra, while program elements could be added or replaced, according to specific user demands.

The authors are grateful to Dr. Bianca Habermann and members of Shevchenko laboratory for useful discussions and experimental support, and to Dr. Uri Pick and Dr. Adriana Katz (Weizmann Institute, Israel) for fruitful collaboration on D. salina proteomics; Dr. Ivan Adzhubey (Brigham and Women's Hospital) for expert set up of the MS BLAST server. We are indebted to Ms. Judith Nicholls for a critical reading of the manuscript. The work was supported by grants PTJ-BIO/0313130 from BMBF to A.S. and 1R01GM070986-01A1 from NIH NIGMS to S.S. and A.S.

5 References

[1] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[2] Yates, J. R. III, Gilchrist, A., Howell, K. E., Bergeron, J. J., *Nat. Rev. Mol. Cell. Biol.* 2005, *6*, 702–714.

[3] Elias, J. E., Haas, W., Faherty, B. K., Gygi, S. P., *Nat. Methods* 2005, *2*, 667–675.

[4] Domon, B., Aebersold, R., *Science* 2006, *312*, 212–217.

[5] Delahunty, C., Yates, J. R. III, *Methods* 2005, *35*, 248–255.

[6] Sadygov, R. G., Cociorva, D., Yates, J. R. III, *Nat. Methods* 2004, *1*, 195–202.

[7] Fenyo, D., *Curr. Opin. Biotechnol.* 2000, *11*, 391–395.

[8] Shadforth, I., Crowther, D., Bessant, C., *Proteomics* 2005, *5*, 4082–4095.

[9] Rappsilber, J., Mann, M., *Trends Biochem. Sci.* 2002, *27*, 74–78.

[10] Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H. *et al.*, *Anal. Chem.* 2003, *75*, 1155–1163.

[11] Zhang, Z., *Anal. Chem.* 2004, *76*, 3908–3922.

[12] Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P. *et al.*, *Nat. Biotechnol.* 2004, *22*, 214–219.

[13] Gibbons, F. D., Elias, J. E., Gygi, S. P., Roth, F. P., *J. Am. Soc. Mass Spectrom.* 2004, *15*, 910–912.

[14] Liska, A. J., Shevchenko, A., *Proteomics* 2003, *3*, 19–28.

[15] Liska, A. J., Shevchenko, A., *Trends Anal. Chem.* 2003, *22*, 291–298.

[16] Habermann, B., Oegema, J., Sunyaev, S., Shevchenko, A., *Mol. Cell. Proteomics* 2004, *3*, 238–249.

[17] Guercio, R. A., Shevchenko, A., Lopez-Lozano, J. L., Paba, J. *et al.*, *Proteome Sci.* 2006, *4*, 11.

[18] Fullekrug, J., Shevchenko, A., Simons, K., *BMC Biochem.* 2006, *7*, 8.

[19] Shevchenko, A., Leal de Sousa, M. M., Waridel, P., Bittencourt, S. T. *et al.*, *J. Proteome Res.* 2005, *4*, 862–869.

[20] Liska, A. J., Shevchenko, A., Pick, U., Katz, A., *Plant Physiol.* 2004, *136*, 2806–2817.

[21] Liska, A. J., Popov, A. V., Sunyaev, S., Coughlin, P. *et al.*, *Proteomics* 2004, *4*, 2707–2721.

[22] Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A. *et al.*, *Anal. Chem.* 2001, *73*, 1917–1926.

[23] Sunyaev, S., Liska, A. J., Golod, A., Shevchenko, A., *Anal. Chem.* 2003, *75*, 1307–1315.

[24] Mackey, A. J., Haystead, T. A. J., Pearson, W. R., *Mol. Cell. Proteomics* 2002, *1*, 139–147.

[25] Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. A. *et al.*, *J. Biol. Chem.* 2001, *17*, 28327–28339.

[26] Frank, A., Tanner, S., Bafna, V., Pevzner, P., *J. Proteome Res.* 2005, *4*, 1287–1295.

[27] Tabb, D. L., Saraf, A., Yates, J. R. III, *Anal. Chem.* 2003, *75*, 6415–6421.

[28] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[29] Standing, K. G., *Curr. Opin. Struct. Biol.* 2003, *13*, 595–601.

[30] Taylor, J. A., Johnson, R. S., *Rapid Commun. Mass. Spectrom.* 1997, *11*, 1067–1075.

[31] Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C. *et al.*, *Mol. Cell. Proteomics* 2005, *4*, 1194–1204.

[32] Searle, B. C., Dasari, S., Turner, M., Reddy, A. P. *et al.*, *Anal. Chem.* 2004, *76*, 2220–2230.

[33] Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzihradszky, K. F. *et al.*, *Mol. Cell. Proteomics* 2005, *4*, 1189–1193.

[34] Mayya, V., Rezaul, K., Cong, Y. S., Han, D., *Mol. Cell. Proteomics* 2005, *4*, 214–223.

[35] Schlosser, A., Volkmer-Engert, R., *J. Mass Spectrom.* 2003, *38*, 523–525.

[36] Parker, K. C., Garrels, J. I., Hines, W., Butler, E. M. *et al.*, *Electrophoresis* 1998, *19*, 1920–1932.

[37] Shevchenko, A., Chernushevich, I., Wilm, M., Mann, M., *Mol. Biotechnol.* 2002, *20*, 107–118.

[38] Yates, J. R. III, Morgan, S. F., Gatlin, C. L., Griffin, P. R. *et al.*, *Anal. Chem.* 1998, *70*, 3557–3565.

[39] Gentzel, M., Kocher, T., Ponnusamy, S., Wilm, M., *Proteomics* 2003, *3*, 1597–1610.

[40] Shevchenko, A., Sunyaev, S., Liska, A., Bork, P. *et al.*, *Methods Mol. Biol.* 2003, *211*, 221–234.

[41] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[42] Frank, A., Pevzner, P., *Anal. Chem.* 2005, *77*, 964–973.

[43] Schagger, H., *Methods Cell. Biol.* 2001, *65*, 231–244.

[44] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[45] Feller, W., *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, London, Sydney 1966.

[46] Wielsch, N., Thomas, H., Surendranath, V., Waridel, P. *et al.*, *J. Proteome Res.* 2006, *5*, 2448–2456.

[47] Savitski, M. M., Nielsen, M. L., Zubarev, R. A., *Mol. Cell. Proteomics* 2005, *4*, 1180–1188.

[48] Bern, M., Goldberg, D., McDonald, W. H., Yates, J. R. III, *Bioinformatics* 2004, *20*, 149–154.

[49] Taylor, J. A., Johnson, R. S., *Anal. Chem.* 2001, *73*, 2594–2604.

[50] Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C. *et al.*, *J. Proteome Res.* 2006, *5*, 3018–3028.

[51] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

