Protein identification pipeline for the homology-driven proteomics

Magno Junqueira\textsuperscript{a}, Victor Spirin\textsuperscript{b}, Tiago Santana Balbuena\textsuperscript{a,c}, Henrik Thomas\textsuperscript{a}, Iuan Adzhubei\textsuperscript{b}, Shamil Sunyaev\textsuperscript{b}, Andrej Shevchenko\textsuperscript{a,*}

\textsuperscript{a}Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany
\textsuperscript{b}Division of Genetics, Department of Medicine, Brigham and Women's Hospital, New Research Building, 77 Ave. Louis Pasteur, Boston, MA 02115, United States
\textsuperscript{c}Plant Cell Biology Laboratory. Department of Botany, IB-University of São Paulo, CP11461, 05422-970 São Paulo, Brazil

ARTICLE INFO

Homology-driven proteomics is a major tool to characterize proteomes of organisms with unsequenced genomes. This paper addresses practical aspects of automated homology-driven protein identifications by LC-MS/MS on a hybrid LTQ Orbitrap mass spectrometer. All essential software elements supporting the presented pipeline are either hosted at the publicly accessible web server, or are available for free download.

© 2008 Elsevier B.V. All rights reserved.

Keywords:
MS/MS
De novo sequencing
MS BLAST
Sequence-similarity searches
Insect proteomics
Plant proteomics

1. Introduction

Conventional protein identification approaches presume the identity between sequences of peptide precursors fragmented in MS/MS experiments and sequences produced by in-silico digestion of protein database entries. Database searches with uninterpreted peak lists deduced from MS/MS spectra are, currently, the major approach for identification of proteins whose sequences are accurately represented in protein or EST databases (reviewed in [1–3]). However, this approach has limited capacity in identifying proteins whose sequences remain unknown (i.e. not present in a database), or heavily modified proteins, or proteins isolated from wild-bred species that often manifest strong sequence polymorphism (reviewed in [4–6]).

If both analyzed unknown proteins and reference proteins from related species belong to conserved protein families, a few identical peptides fragmented in MS/MS experiments might enable their direct cross-species identification by conventional database searching means [5]. However, local similarity between the two protein sequences does not necessarily imply their functional resemblance, especially if confined to a sequence domain, rather than extending over a full length protein sequence [7]. Furthermore, protein identification relying on sequence identities is inherently biased towards most conserved protein families and might not adequately reflect the true composition of a protein mixture [8].

Identification of unknown proteins typically relies on the similarity (rather than identity) of sequences of fragmented peptides and sequences of known homologous proteins from phylogenetically related species (reviewed in [4]). Sequence-similarity searches tolerate multiple mismatches between the analyzed and reference peptide sequences. One approach, termed sequence tag search, identifies peptides that share a stretch of identical sequence of only a few amino acid residues, if complemented by the masses of corresponding fragment ions [9]. If a part of each sequence tag is allowed to mismatch, the search would produce a large number of plausible, yet low statistically confident, hits. However, simultaneous consideration of multiple tags automatically
deduced from many MS/MS spectra in a single database search enhances the protein identification specificity [10,11].

De novo interpretation of peptide tandem mass spectra combined with dedicated sequence-similarity searching engines expands the organismal coverage of homology-driven proteomics. CIDentify [12], a Mass Spectrometry tailored version of gapped BLAST [13], MS BLAST [14], FASTS [15], MS-Homology [16], OpenSea [17], among others, have been successfully applied in numerous proteomics studies. They produce and score error-tolerant alignments of compared peptide sequences and, in principle, do not require long and identical sequence stretches to produce a confident protein hit. Mass Spectrometry driven BLAST (MS BLAST), a high-throughput web-accessible searching tool, simultaneously processes queries comprising redundant, degenerate and partially inaccurate peptide sequence candidates obtained by automated interpretation of tandem mass spectra [7,14,18–21].

LC-MS/MS analysis under data-dependent acquisition control produces thousands of tandem mass spectra of varying quality and information content. When applied to large MS/MS datasets, conventional sequence-similarity searches that rely on relatively accurate interpretation of a few selected spectra produce identifications with high false positive rates. It is therefore desirable to collapse the dataset down to a few essential, representative MS/MS spectra from target proteins and leave out spectra acquired from protein and chemical contaminants [21,22]. One way to cope with this problem would be to reduce the analysis sensitivity by targeting only most abundant precursors. However, since most abundant proteins often represent ubiquitous background commonly encountered in any biochemical isolation experiment (housekeeping proteins, heat shocks, metabolic enzymes etc. that are well represented in a database), sequence-similarity searches would hardly bring in any new proteins, but mainly recapitulate cross-species identifications produced by conventional stringent searches.

Here we report an automated pipeline that combines the high sensitivity and dynamic range of LC-MS/MS analysis with sequence-similarity searches. It relies on a layered data mining strategy that targets de novo interpretations on MS/MS spectra that, otherwise, would remain unmatched by conventional database searching means. This paper addresses practical aspects of automated homology-driven protein identifications using large MS/MS datasets acquired on a hybrid LTQ Orbitrap mass spectrometer [23]. Importantly, all essential data processing software is either hosted at a publicly accessible web server, or is available for free download.

2. Materials and methods

2.1. Chemicals

Cleland’s reagent (DTT) and iodoacetamide were of analytical grade and purchased from Sigma-Aldrich (Munich, Germany); water and acetonitrile were of LC-MS grade from Fisher Scientific (Schwerte, Germany). Formic acid and trifluoroacetic acid were HPLC grade obtained from Merck (Darmstadt, Germany). Modified porcine trypsin (sequencing grade) was purchased from Promega (Mannheim, Germany).

2.2. Software and web resources

EagleEye software for filtering MS/MS queries is accessible at:
http://genetics.bwh.harvard.edu/cgi-bin/msfilter/eagleeye.cgi.

PepNovo software (version PepNovo2MSB) for high throughout de novo interpretation of ion trap MS/MS spectra is available for download at:
http://proteomics.bioprojects.org/Software/PepNovo.html.

MS BLAST sequence-similarity search engine is accessible at:

2.3. Protein samples and in-gel digestion

Protein samples were obtained from the insect Triatoma infestans and the Brazilian pine Araucaria angustifolia purified in on-going collaboration projects in the Laboratory of Biochemistry and Protein Chemistry – University of Brasilia, and Plant Cell Biology Laboratory – University of Sao Paulo, respectively. Samples were fractionated by two-dimensional gel electrophoresis and protein spots visualized by Coomassie (T. infestans proteins) or silver (A. angustifolia proteins) staining. Spots were excised and in-gel digested with trypsin as described in [24,25]. Tryptic peptides were extracted from the gel matrix by 0.1% formic acid and acetonitrile, dried down in a vacuum centrifuge, and stored at –20 °C prior to the analysis.

2.4. LC-MS/MS analysis

LC-MS/MS was performed on an Ultimate 3000 nanoLC system ( Dionex, Sunnyvale, USA), interfaced to a LTQ Orbitrap hybrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) via a robotic nanoflow ion source TriVersa (Advion Biosciences, Ithaca NY) equipped with a LC coupler and a chip with 4.1 μm nozzle diameter controlled by Chipsoft 6.4 software (Advion Biosciences). Ionization voltage was set to 1.7 kV and spacing between the chip and ion transfer capillary opening was maintained within 3 to 5 mm.

After in-gel digestion tryptic peptides were re-dissolved in 10 μL of 0.05% TFA and 4 μL loaded onto a trapping column packed with C18 PepMAP100 (Dionex) at the flow rate of 20 μL/min in 0.05% TFA. After 6 min washing, peptides were eluted into the nanocolumn C18 PepMAP100 (15 cm × 75 μm ID, 3 μm particles, Dionex) at the flow rate of 200 nL/min. Peptides were separated using the mobile phase gradient: from 5 to 20% of solvent B in 20 min, 20 to 50% B in 16 min, 50 to 100% B in 5 min, 100% B during 10 min, and back to 5% B in 10 min. Solvent A was 95:5 H2O: ACN (v/v) containing 0.1% formic acid; solvent B was 20:80 H2O: ACN (v/v) containing 0.1% formic acid.

LC-MS/MS data was acquired in data-dependent acquisition (DDA) mode controlled by Xcalibur 2.0 software (Thermo Fisher Scientific). The automatic gain control (AGC) was set to 5 × 10^6 charges for survey scan on the Orbitrap using one microscan and 5 × 10^6 charges for MS/MS on the ion trap analyzers using three microscans. A typical DDA cycle...
consisted of a MS scan within m/z 300–1600 performed under the target mass resolution of 60,000 (full width at half maximum) followed by MS/MS fragmentation of the four most intense precursor ions under the normalized collision energy of 35% in the linear trap. The dynamic ion selection threshold for MS/MS experiments was set to 500 counts and precursor isolation window was 4 amu. Activation parameter q was set to 0.25 and activation time of 30 ms was applied. Single charge precursors were automatically excluded from MS/MS acquisition, and m/z of precursors already fragmented was dynamically excluded for further 90 s. Each LC-MS/MS run was converted to .mgf file using extract_msn.exe utility from Xcalibur 2.0SR2 software (Thermo Fisher Scientific) under the following settings: minimum total ion intensity threshold: 500; minimum number of fragment ions: 5; minimum signal-to-noise ratio: 3; charge state recognition was enabled. Each .mgf file was named according to the original name of the .raw file.

3. Results and discussion

3.1. Pre-processing of LC-MS/MS queries prior to de novo sequencing

LC-MS/MS analysis of in-gel tryptic digests usually produces many MS/MS spectra originating from common background proteins — trypsin autolysis products, human and sheep keratins, antibodies or abundant protein components of cell media [20,25,28], which severely undermine the performance of sequence-similarity searches [20]. Note, that sequences of human and sheep keratins are rich in low complexity regions and therefore searches relying upon local sequence similarity (rather than complete identity of fragmented peptides) might confidently match a large number of functionally unrelated proteins, termed as “orphan” hits [26]. Because of protein contaminants diversity, stringent (e.g. MASCOT) searches could only remove a fraction of corresponding peptides, while more efficient approach is to filter all MS/MS spectra against a representative library of non-annotated background spectra. EagleEye software compares each MS/MS spectrum from the submitted query with non-annotated spectra from the background library having the same m/z and charge of the precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26]. Specificity and speed of EagleEye spectra screening was substantially higher compared to an earlier prototype [20] and therefore in this work we applied it prior to both MASCOT and sequence-similarity searches. Complete LC-MS/MS runs saved as .raw files were converted into .mgf (MASCOT generic format) files. For batch mode filtering, several .mgf files representing individual LC-MS/MS runs were combined into a single .zip archive and processed together under the same settings: the mass tolerance for precursor ion and then scores the dissimilarity between the two spectra [26].
however, did not limit the throughput because several LC-MS/MS runs could be submitted together and processed as a batch. Filtered .mgf files could be either retrieved from the server individually, or the full batch downloaded upon its completion. Once the job was submitted, the browser window could be closed and Grid Gateway Interface (GGI) that monitors job processing, could be accessed anytime later. The server identified users by a session ID number communicated to the browser.

From each submitted .mgf file, EagleEye created two .mgf files, one containing non-background spectra and another one containing background spectra from the same LC-MS/MS run. These .mgf files, along with two complementary files containing filtering settings and report on spectra matching statistics, could be viewed in a browser or downloaded via direct links provided at the CGI session page. Note that EagleEye filtering does not interfere with the content of individual MS/MS spectra and all metadata, such as comment lines, acquisition time etc., are preserved.

Case studies presented below demonstrate EagleEye filtering efficiency, while more examples are reported in [8,20,26,29,30]. We found that removing a large number of background spectra (30 to 50% of all MS/MS spectra of multiply charged precursors was a fair ballpark estimate) improved data processing speed, especially in complex routines involving multiple database searches and de novo sequencing. In sequence-similarity searches against a comprehensive (all-species) database it reduced very substantially (usually, by more than a few hundreds) the number of orphan hits that should, otherwise, be followed up by extensive manual validation. Taken together, EagleEye filtering was a key factor that enabled to combine sequence-similarity searches with LC-MS/MS analysis performed at the uncompromised acquisition speed and sensitivity.

3.2. Rapid de novo sequencing of LC-MS/MS datasets

To interpret LC-MS/MS runs de novo, we used PepNovo software developed by Frank and Pevzner [27,31]. The software was specifically tailored for interpreting linear ion trap tandem mass spectra and the output format of its version PepNovo2MSB conformed the input conventions of MS BLAST sequence-similarity searching engine [7,19].

Under default settings (controlled via command line) the software produced up to 7 redundant, degenerate and partially complete sequence candidates per each interpreted spectrum. Sequencing of a single spectrum usually took 0.15 s on a desktop PC. Typically, ca. 1500 spectra were acquired during 40 min LC-MS/MS run of an in-gel tryptic digest on LTQ Orbitrap machine and less than 50% of them remained after EagleEye filtering, so that complete de novo interpretation of the full dataset usually took less than 4 min. For each interpretation, the software assigned a quality score, which corresponded to the expected number of correctly called amino acid residues in the top sequence proposal (Fig. 1). Previous experiments suggested that, for MS BLAST identifications, it is usually practical to only consider sequence candidates having the score of 6.0 or above [20,32].

![Fig. 1 – Part of the output of PepNovo batch mode interpretation of MS/MS spectra from the LC-MS/MS run. For each interpreted spectrum PepNovo provides: neutral mass of the peptide precursor (a); spectrum name, including the precursor charge state (b); total ion current (TIC) in MS/MS spectrum (c); TIC fraction covered by expected fragments of the top candidate sequence (d); sequence quality score representing the expected number of correct amino acids in the top candidate sequence (e); candidate sequences (f), formatted according to MS BLAST conventions: B = R or K (generic trypsin cleavage site); Z = Q or K (if indistinguishable in low resolution MS/MS spectra); L = L or I; M + 15.99 = methionine sulfoxide residue; X = undetermined amino acid residue.](image-url)
PepNovo output sequence candidates as a tab-delimited text file (Fig. 1), which could be pasted directly into the MS BLAST query window. Note that in low mass resolution MS/MS spectra, isobaric amino acid residues of phenylalanine and mono-oxidized form of methionine could not be reliably distinguished. Since their mismatching is heavily penalized by the BLAST sequence alignment algorithm, PepNovo software was set to output both variants of de novo interpretation, while command line option span1 of WU-BLAST search engine utilized by MS BLAST compensated increased redundancy of search queries [19].

3.3. MS BLAST searches with peptide sequence queries produced by LC-MS/MS

Filtered .mgf queries were interpreted de novo by PepNovo software and the entire output directly submitted to MS BLAST search at the web server. Advanced (command line) options and major MS BLAST conventions are explained in detail in [19]; MS BLAST scoring scheme and its validation was described in [7].

Note that default MS BLAST settings (as seen upon accessing MS BLAST job submission web page) were optimized for sequence-similarity identifications that rely on a small number (usually, less than 50) of relatively accurate peptide sequence candidates, typically obtained by manual interpretation of a few MS/MS spectra acquired from most abundant precursors. These settings maximize the sensitivity of sequence-similarity identification and strictly follow MS BLAST scoring scheme. They are, however, not directly applicable to ca. 100-fold larger queries produced by LC-MS/MS and therefore additional restrictions were applied by activating LC-MS/MS Presets box at the query input page (Fig. 2).

MS BLAST is based on WU-BLAST search engine [33]. Its command line settings S and S2 specify the threshold scores for, respectively, the highest and all other High Scoring Segment Pairs (HSPs) reported for each hit. Reasonably high thresholds (default settings were S=55 and S2 = 55) prevented MS BLAST from reporting many weakly matching HSPs that, otherwise, plagued its scoring scheme [7] and increased the false positive rate. However, if necessary, weaker sequence alignments might still be reported by lowering S and S2 thresholds.

While analyzing mixtures of unknown proteins it is hard to guess how many sequenced precursors might belong to the target protein. LC-MS/MS Presets specified the expected number of peptides at 20. In our experience of LC-MS/MS analysis of mixtures of known proteins, the number peptides matched to the individual protein sequence is typically lower and therefore more conserved threshold scores were applied while evaluating MS BLAST output [7]. B, V and hspmax parameters are, respectively, the number of reported alignments, descriptions and HSPs. By default, they were set at the arbitrary value of 1000, which sufficed for processing queries obtained by de novo interpretation of 500–600 tandem mass spectra. Note that unnecessarily high settings slowed down the search and only increased the number of reported non-confident alignments. Therefore, they were only used if the search engine produced a warning message that B, V or hspmax limits were exceeded.

Filtering of low complexity sequences is a built-in option of the WU-BLAST engine and was engaged by setting it to Default in the corresponding menu. It was instrumental in eliminating low complexity sequence stretches that are common in human and sheep keratin peptides, if corresponding MS/MS spectra passed, for any reason, through EagleEye filtering. Low complexity segments were substituted by zero-scoring X symbols while full input sequence queries were reported at the top of MS BLAST output. Albeit low complexity filtering further reduced the number of keratin-related hits, it might accidentally eliminate bona fide proteins. If data analysis indicated that a major component might be missed despite good quality of input MS/MS data, low complexity filter should be turned off and MS BLAST search repeated under B, V and hspmax settings exceeding 2000.

3.4. Identification of insect and plant proteins from species with unsequenced genomes

Here we describe the practical application of the automated de novo sequencing — sequence-similarity searching pipeline for the LC-MS/MS identification of unknown proteins from insect and plant species.

First we seek to identify a potent platelet inhibitor from the saliva of T. infestans, a blood-sucking bug transmitting the parasite Trypanosoma cruzi that causes Chagas disease [34]. T. infestans proteome is poorly represented in a database (26 sequences currently available in NCBI nr) and, not surprisingly, conventional database searches fail to identify even most abundant proteins.
A sample of *T. infestans* saliva was separated on 2D gel and proteins visualized by Coomassie staining. LC-MS/MS analysis of the in-gel digest of one of these spots produced 2210 MS/MS spectra, 1888 (2+) and 322 (3+), considering that only multiply charged ions were targeted by MS/MS in DDA experiments (Fig. 3).

Fig. 4 presents MS/MS spectrum acquired from a doubly charged precursor m/z 922.933 (panel A) along with several candidate sequences obtained by its PepNovo interpretation (inset B). Altogether, 2665 peptide sequence candidates obtained by de novo interpretation of 422 MS/MS spectra, were merged into an MS BLAST query string. MS BLAST search produced several HSPs (Table 1) that confidently identified a protein homologous to Triabin 33, a platelet inhibitor from *T. infestans* saliva. Since MASCOT search hit no peptides from the Triabin 33 sequence, we assumed that the analyzed spot, most likely, contained its homologue. Hence, the sequence of AAQ68064 Triabin 33 from *T. infestans* suggests that the corresponding peptide (shown in Fig. 4) should...
Fig. 3 – Base peak LC-MS/MS chromatogram of in-gel tryptic digest of a silver stained spot with apparent MW of 19 kDa and pI of 8.0 excised from a 2D gel of Triatoma infestans saliva. The analysis produced, in total, 2210 MS/MS spectra acquired from doubly- and triply-charged precursor ions. Peaks at the chromatogram are designated with base peak m/z.

Fig. 4 – De novo interpretation of the MS/MS spectrum from precursor ion with m/z 922.933 by PepNovo software. The interpretation of the spectrum (panel A) acquired on a linear ion trap analyzer produced several candidate sequences (inset B) with the sequence quality score of the top candidate of 13.6. Along with candidate sequences from other fragmented precursors, they were submitted to MS BLAST search that produced the sequence alignment presented in Table 1. Peaks in the spectrum (panel A) are designated according to the fragment type and m/z, computed from the aligned peptide sequence.
have the sequence: (K)NGDGSTTTVTINSYISR (calculated mass=1785.8613 Da), which differed from the actually observed m/z 1843.866 by 58.007 Da. The best matching PepNovo candidate sequence (Table 1, alignment at the top) was (K)NDDGSTTTVTINSYISR. Its calculated mass of 1843.8668 Da conforms with Asp−Gly substitution, which is also supported by continuous b-ion series. Hence, we demonstrated that automated de novo sequencing of LC-MS/MS spectra followed by MS BLAST searches identified a protein, which was missed by conventional (MASCOT) data mining approach.

Another example, the identification of a protein from developing embryos of A. angustifolia presents a more complex scenario, in which several proteins were identified by MASCOT and MS BLAST searches in a complementary manner [20,32].

A. angustifolia is an economically important endangered native conifer [35], whose genome has not been sequenced. Currently, NCBI nr database contains five protein entries from this species.

Two silver stained spots with apparent molecular weights of 50 kDa and 62 kDa were excised from a two-dimensional gel of the preparation obtained at the late stage of zygotic embryo development. Proteins were in-gel digested with trypsin and analyzed by LC-MS/MS as described above (Table 2). MASCOT searches were performed against a MSDB database.

Eleven proteins were identified by MASCOT and MS BLAST searches in both spots (Table 3), while only five of them were unambiguously identified by MASCOT. Note that MASCOT searches identified no proteins in 62 kDa spot. We then applied de novo sequencing to the filtered MS/MS datasets and submitted candidate sequences to MS BLAST, according to the workflow presented in Fig. 5. Sequence-similarity

Table 2 - EagleEye filtering of MS/MS queries acquired from 50 and 62 kDa proteins against a background library of 12,095 MS/MS spectra

<table>
<thead>
<tr>
<th>Spot</th>
<th>MS/MS spectra Removed</th>
<th>Retained</th>
<th>Keratin/trypsin related hits* Before filtering</th>
<th>After filtering</th>
<th>MS BLAST search time shortened by,</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 kDa</td>
<td>698</td>
<td>1417</td>
<td>395</td>
<td>42</td>
<td>30 min</td>
</tr>
<tr>
<td>62 kDa</td>
<td>871</td>
<td>1182</td>
<td>378</td>
<td>124</td>
<td>25 min</td>
</tr>
</tbody>
</table>

* Keratin/trypsin” hits are database entries explicitly annotated as trypsins or keratins (from any species).
searches independently confirmed, respectively, one and four borderline hits in the analysis of 50 kDa and 62 kDa proteins, including a rubisco binding protein that controls the assembly and activity of RuBisCo enzyme, (Table 3), which is responsible for carbon fixation from atmospheric carbon dioxide.

In 62 kDa spot MS BLAST also identified a disulfide-isomerase protein with 12 HSPs matched to the protein sequence from *Oryza sativa*. Importantly, this protein was not identified by MASCOT search.

The workflow presented here is a simple, fast and robust approach for the identification of proteins from organisms with unsequenced genomes [5,20]. MASCOT conveniently identified most conserved proteins, sharing several identical peptides with known database sequences, while the rest was processed with less stringent sequence-similarity searches. The approach has been successfully applied in several proteomics projects in unsequenced insect [29] and plant [8] species.

Although the case studies presented above encompassed the identification of gel-separated proteins, the approach is generic and could be equally applied to far more complex protein mixtures. Indeed, at all stages of data processing (EagleEye filtering, de novo sequencing by PepNovo, MASCOT and MS BLAST searches) individual mass spectra (or peptide sequence candidates) were considered independently and the

![Fig. 5](https://example.com/fig5.png)

**Fig. 5** – Protein identification workflow that uses a combination of MASCOT searches and de novo sequencing followed by MS BLAST searches. First, all spectra were filtered against a background spectra library by EagleEye software, which removed a large number of background MS/MS spectra irrespective of their quality, annotation and origin. Filtered data file in .mgf format was submitted to MASCOT searches against a comprehensive MSDB database. If 3 or more unique peptides were matched by MS/MS spectra with ions score exceeding 50, these identifications were considered positive. If 3 peptides were matched with ions scores above, but 20 but below 50, or only one peptide was matched with the score above 50, hits were considered borderline and subjected to further validation by de novo sequencing. In parallel, the same .mgf file with filtered spectra was subject to batch de novo sequencing followed by MS BLAST search with obtained candidate sequences. For identifications solely based on sequence similarity and for independent validation of MASCOT borderline hits, MS BLAST scoring scheme was applied.
total size of the query (reflected by the full number of submitted MS/MS spectra) did not play a major role: it mostly affected the processing time, rather than the search outcome.

4. Conclusions and perspectives

We presented a pipeline for homology-driven proteomics by LC-MS/MS and sequence-similarity protein identifications. Single LC-MS/MS dataset acquired at the uncompromised sensitivity typically yielded several thousands low mass resolution tandem mass spectra. Upon removal of background MS/MS spectra by EagleEye software, the dataset was used for conventional stringent searches (MASCOT) and de novo sequencing by PepNovo followed by MS BLAST sequence-similarity searches. This enabled robust identification of known proteins, proteins highly homologous to known proteins and relatively non-conserved proteins in a single analysis. Importantly, the approach only relied on publicly available software tools, with two key elements of the pipeline — EagleEye and MS BLAST run on a publicly accessible server. Therefore, no or minimal changes in adopted laboratory routines would be required for implementing sequence-similarity searches in any interested proteomics laboratory.

Because of its availability, relative ease of use and data processing speed, we envision that this pipeline paves the way to accurate deciphering of unknown proteomes of organisms that were not adequately covered by genomic sequencing and might have interesting implications in the broad field of plant and animal biology. It is equally conceivable that de novo sequencing followed by the similarity analysis should become a common requirement for presenting identification datasets obtained from organisms with insufficiently characterized genomes and/or if strong effect of protein sequence polymorphism is expected.

Acknowledgements

We are grateful to Prof. Pavel Pevzner and Dr. Ari Frank (Department of Computer Science and Engineering, UCSD) for expert help with interfacing PepNovo software to MS BLAST. This work was, in part, supported by NIH NIGMS grant 1R01GM070986-01A1 to S. Sunyaev and A. Shevchenko.

References


[33] Gish W. WU-BLAST2.0; 1996.
