JOURNAL OF PROTEOMICS 71 (2008) 346–356

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Protein identification pipeline for the homology-driven proteomics

Magno Junqueira[a], *Victor Spirin*[b], *Tiago Santana Balbuena*[a,c], *Henrik Thomas*[a], *Ivan Adzhubei*[b], *Shamil Sunyaev*[b], *Andrej Shevchenko*[a,*]

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

ARTICLE INFO

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

ABSTRACT

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.

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

* *Corresponding author.*
E-mail address: shevchenko@mpi-cbg.de (A. Shevchenko).

1874-3919/$ – see front matter
doi:10.1016/j.jprot.2008.07.003

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:

http://genetics.bwh.harvard.edu/msblast/index.html.

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 H_2O: ACN (v/v) containing 0.1% formic acid; solvent B was 20:80 H_2O: 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 5×10^5 charges for survey scan on the Orbitrap using one microscan and 5×10^4 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.

2.5. *Removal of background MS/MS from LC-MS/MS datasets*

Filtering at EagleEye web server removed non-annotated MS/MS spectra acquired from common background proteins [26]. Files in .mgf format obtained from several LC-MS/MS were combined in a single .zip file and uploaded to the server. Precursor mass tolerance and fragment mass tolerance were set to 0.01 Da and 0.6 Da, respectively; *p*-value cut-off was set at 0.01. The background library comprised 12,009 non-annotated tandem mass spectra acquired on a LTQ Orbitrap machine (see [26] for details). Filtered .mgf files downloaded from the server were subjected directly to MASCOT searches or de novo sequencing without further processing.

2.6. *Protein identification by MASCOT searches*

Tandem mass spectra from LC-MS/MS runs filtered by EagleEye were searched against a MSDB database (2,344,227 sequences entries; updated April, 2006) by MASCOT v.2.1 software (Matrix Science Ltd., London, UK) installed on a local 2 CPU server. Tolerances for precursor and fragment masses were set at 5 ppm and 0.6 Da, respectively; up to 2 missed cleavages were allowed; instrument profile: ESI-Trap; fixed modification: carbamidomethyl (cysteine); variable modification: oxidation (methionine) and acetylation of the N-terminal peptide of protein sequence entry are set as variable modifications. The confidence criteria for protein identification by MASCOT were set conditionally on the number of matched peptides and individual peptide ions scores. Hits were considered confident if produced by matching of at least three MS/MS spectra with scores higher than 50. Hits made by matching three peptides with the scores higher than 20 or only one peptide with the score higher than 50 were considered borderline.

2.7. *De novo sequencing and protein identification by MS BLAST searches*

MS/MS spectra were subjected to batch de novo sequencing by PepNovo program [27]. Up to 7 candidate peptide sequences for each interpreted tandem spectra were considered and only candidates with the sequence quality score of 6 or above were used for subsequent MS BLAST searches. PepNovo outputs were pasted directly into MS BLAST query window and searched against nr database using the *LC-MS/MS Presets* option of the MS BLAST web server. To increase identification confidence, we only considered high scoring segment pairs (HSPs) with the scores of 55 or above as evaluated by scoring scheme described by Habermann et al. *et al.* [7].

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 matching *m*/*z* of precursor and fragment ions and the *p*-value cut-off. The cut-off stands for the estimate of the fraction of high quality spectra that might be lost during filtering because of their random similarity to background spectra with the same *m*/*z* and z. High mass accuracy and resolution of the Orbitrap analyzer strongly limited the number of compared spectra and therefore almost no high quality spectra (as judged, for example, by their peptide ions scores) were lost under $p = 0.01$. Increasing *p*-value cut-off relaxed the matching specificity of background and experimental spectra. Therefore, while it increased the total number of removed background spectra, some genuine spectra, typically with low peptide ions scores, might be lost.

Typically, only a few minutes were required to screen a complete LC-MS/MS run of *ca.* 2000 spectra of multiply charged precursors against the library of more than 12,000 background spectra, although the actual processing time also depended on the server load and uploading/downloading speed. This,

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].

MW Peptide$_{(a)}$	DTA file. charge$_{(b)}$	TIC$_{(c)}$	TIC fraction$_{(d)}$	Score$_{(e)}$	Candidate sequences$_{(f)}$
2058.00	1858.1858.2	29140.185547	0.685148	15.2	-BSQNDDGSTTTVLTSGGYLSR-BSGANDDGSTTTVLTSGGYLSR -BSQGGDDGSTTTVLTSGGYLSR-BSGAGGDDGSTTTVLTSGGYLSR -BSQNDDGSTTTVLTSNYLSR-BSGANDDGSTTTVLTSNYLSR -BSQGGDDGSTTTVLTSNYLSR
1873.86	1942.1942.2	34632.109375	0.589080	14.4	-BHNLLEGGQEDFESSGAGK-BHGGLLEGGQEDFESSGAGK -BHNLLEGGQEDM+15.99ESSGAGK-HGGLLEGGQEDM+15.99ESSGAGK -BHNLLEGGQEDFESSQGK-BHNLLEGGGAEDFESSQGK -BHGGLLEGGQEDFESSQGK
1854.88	2174.2174.2	48517.125000	0.832615	13.8	-BWXXGSTTTVLTSNYLSR-BWXXGSTTTVLTSGGYLSR -BXXXXXGSTTTVLTSNYLSR-BXXXXGSTTTVLTSNYLSR -BXXXGSTTTVLTSNYLSR-BXXXXXGSTTTVLTSGGYLSR -BXXXXGSTTTVLTSGGYLSR
1798.88	2062.2062.2	84326.851563	0.690811	13.7	-BXXQGSTTTVLTSNYLSR-BXXGAGSTTTVLTSNYLSR -BXXQGSTTTVLTSGGYLSR-BXXGAGSTTTVLTSGGYLSR -GSTTTVLTSNYLSR-GSTTTVLTSGGYLSR-BATQGSTTTVLTSNYLSR
1842.88	2074.2074.2	51767.988281	0.596980	13.6	-BNDDGSTTTVLTSNYLSR-BNDDGSTTTVLTSGGYLSR -BGGDDGSTTTVLTSNYLSR-BGGDDGSTTTVLTSGGYLSR -DDGSTTTVLTSNYLSR-DDGSTTTVLTSGGYLSR
1990.95	1376.1376.2	123326.679688	0.689615	13.6	-BGAQVNDKVTGSMSTGLNGQK-BGAQVNDQVTGSMSTGLNGQK -BGAKVNDKVTGSMSTGLNGQK-BGAKVNDQVTGSMSTGLNGQK -BQQVNDKVTGSMSTGLNGQK-BQQVNDKVTGSMSTGLNGKK -BZQVNDQVTGSMSTGLNGQK
1801.94	2534.2534.2	122351.406250	0.769437	13.5	-BAGVDVLEGNEQM+15.99LNAAR-BAGVDVLEGNEQFLNAAR -BAGVDVLEGNEGAM+15.99LNAAR-BAGVDVLEGNEGAFLNAAR -BAGVDVLEGGGEQM+15.99LNAAR-BAGVDVLEGGGEQFLNAAR -BAGVDVLEGGGEGAM+15.99LNAAR

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.

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 upon 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.

Under conventional settings (applied by default if *LC-MS/MS Presets* box remains idle) MS BLAST was used for validating borderline hits produced by MASCOT searches [20,32]. The approach took advantage of the independent de novo interpretation of corresponding MS/MS spectra followed by sequence-similarity search with produced sequence candidates. If MS BLAST independently hit the same peptide as did MASCOT, the identification was considered as positive.

Upon submission of the peptide sequence query, Grid Gateway Interface (GGI) reported the current server workload and the number of pending and processed jobs. Note that, similarly to EagleEye, the server assigned the session identification number that was stored at the local workstation and was automatically recognized, once the browser accessed MS BLAST page again. In addition, the server assigned individual tracking numbers for all submitted job. While a current job was processed, MS BLAST submission page could be accessed via the provided link and another search query submitted. Job processing could be monitored by hitting *Refresh Status* button at the GGI page. User could quit the browser and access the search results anytime later from the same workstation (provided that *cookies* were enabled in the browser) using *Check Status* button. Search results could be also viewed from another computer by pasting the session identification number into the corresponding window at the GGI page and checking the box *Overwrite Default*.

Once the search was completed, the server listed the submission among completed jobs and provided the link to MS BLAST output page, which was identical to the previously described [7,20]. MS BLAST hits were color-coded and confident hits reported at the top of the list.

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.

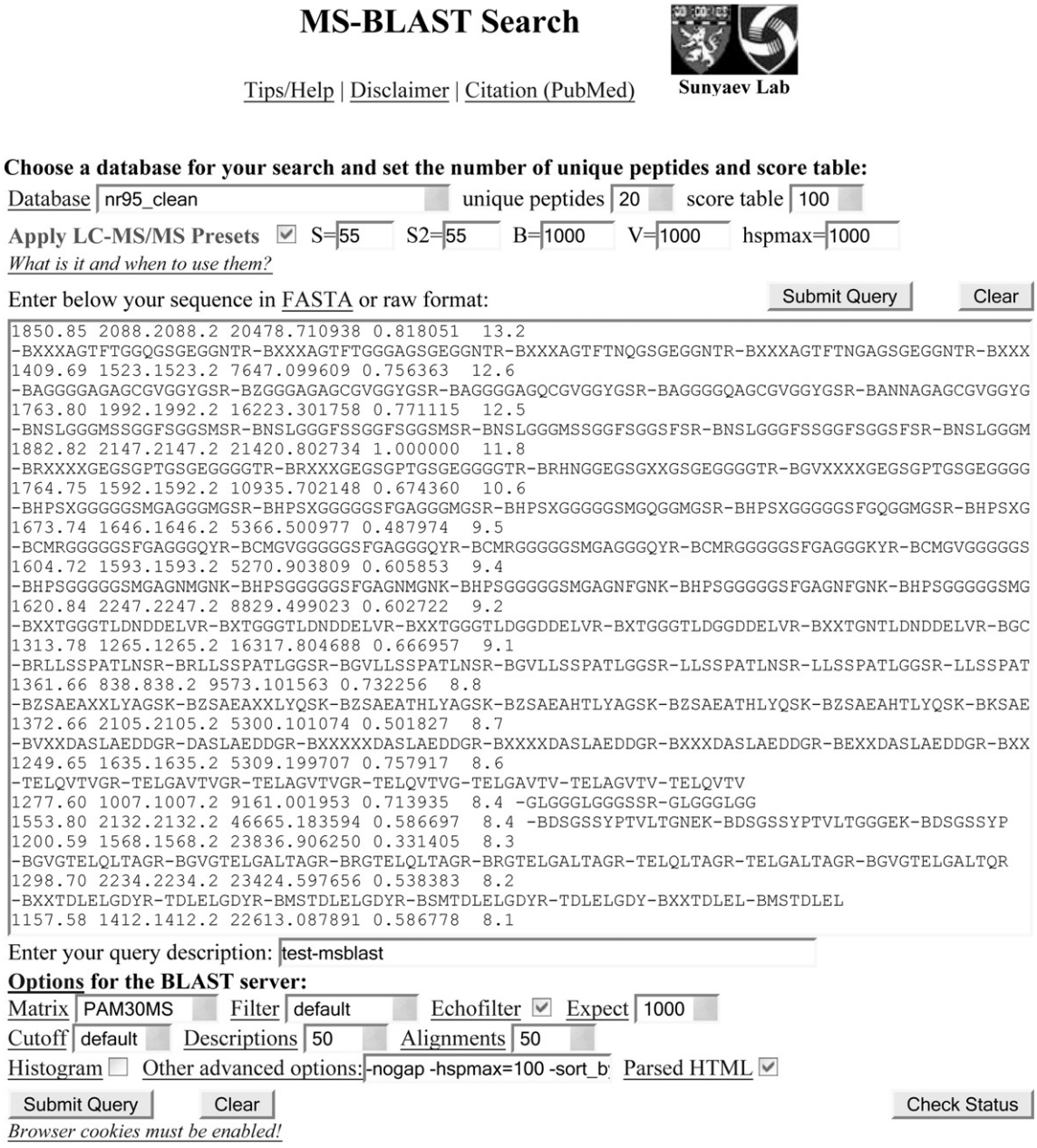

Fig. 2 – Web interface of the MS BLAST server. *LC-MS/MS Presets* check box activates the settings that make possible MS BLAST searches with large peptide sequence queries produced by automated interpretation of MS/MS spectra acquired by data-dependent LC-MS/MS. A part of the search string is shown in the submission window. MS BLAST utilizes degenerate, redundant and partially accurate sequence queries. Usually, up to 7 peptide sequence candidates per each interpreted MS/MS spectrum are included into the search query. Precursor masses, scan numbers, sequence quality scores and other parameters simplifying handling of de novo output, are ignored by the server. MS BLAST server can process a query comprising up to 150,000 amino acid residues, which would be equivalent to a BLAST search with the sequence of *ca.* 16.5 MDa protein chimera.

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 a 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

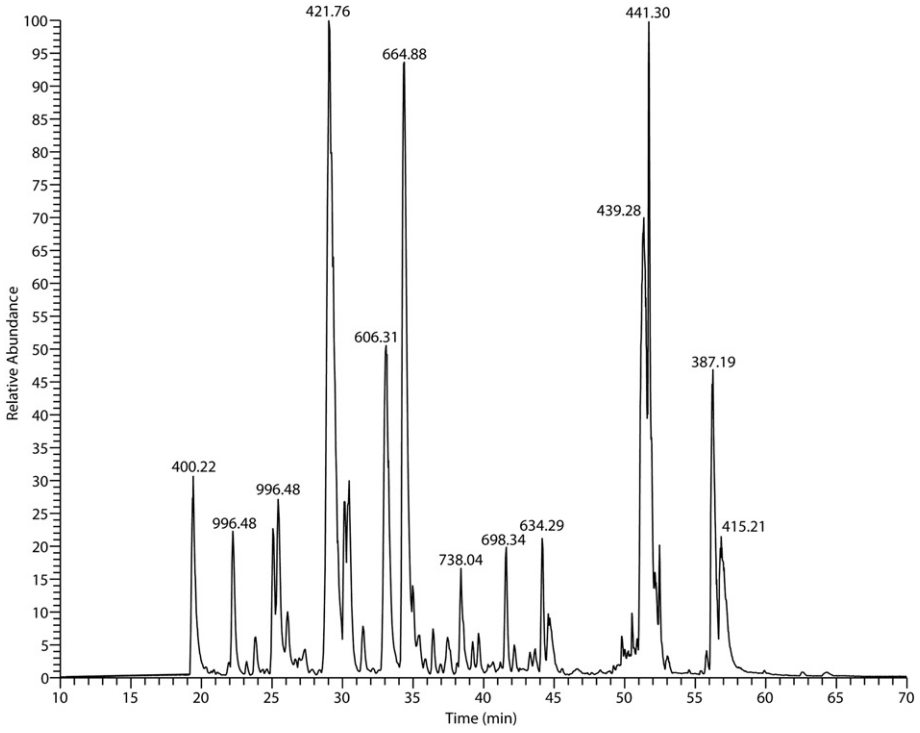

Fig. 3 – Base peak LC-MS/MS chromatogram of in-gel tryptic digest of a silver stained spot with apparent MW of 19 kDa and p*I* of 8.0 excised from a 2D gel of *Triatoma infestans* saliva. The analysis produced, in total, 2210 MS/MS spectra acquired from doubly- and triply-charged precursor ions. Peaks at the chromatogram are designated with base peak *m*/*z*.

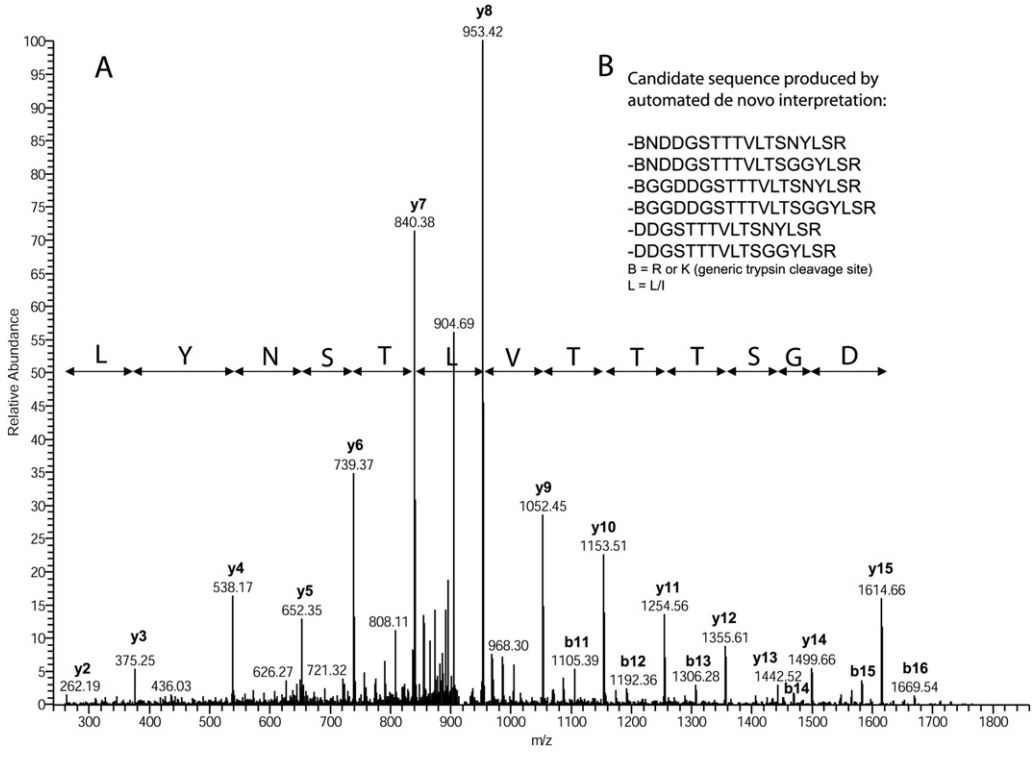

Fig. 4 – De novo interpretation of the MS/MS spectrum from precursor ion with *m*/*z* 922.933 by PepNovo software. The interpretation of the spectrum (panel A) acquired on a linear ion trap analyzer produced several candidate sequences (inset B) with the sequence quality score of the top candidate of 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.

Table 1 – High scoring segment pairs (HSPs) produced by automated de novo sequencing and MS BLAST searches

Peptide mass, Da	Candidate sequences[a,b]	HSPs reported by MS BLAST	BLAST score
1842.88	BNDDGSTTTVLTSNYLSR-BNDDGSTTTVLTSGGYLSR -BGGDDGSTTTVLTSNYLSR-BGGDDGSTTTVLTSGGYLSR -DDGSTTTVLTSNYLSR-DDGSTTTVLTSGGYLS	Query: 572 BNDDGSTTTVLTSNYLSR 589 +N DGSTTTV+TSNY+SR Sbjct: 61 KNGDGSTTTVITSNYISR 78	113
1704.89	-BWLATDNQNYALLQR-BWLATDNKNYALLQR -BWLATDNQGGYALLQR-BWLATDGGQNYALLQR -BWLATDGGQGGYALLQR-BWLATDGGKNYALLQR -BSVLATDNQNYALLQR	Query: 2258 SVLATDNQNYALLQR 2272 SVLATDNQNYA+LQR Sbjct: 117 SVLATDNQNYAILQR 131	105
1715.9	-PTSGQGNLLVLQTAK-PTSGGAGNLLVLQTAK -PTSGQGGGLLVLQTAK-PTSGGAGGGLLVLQTAK -PTSGQGNLLVLGATAK-BSCPTSGQGNLLVLQTAK -BSCPTSGGAGNLLVLQTAK	Query: 18423 CPTSGQGNLLVLQT 18436 CP SGQGN+LVLQT Sbjct: 132 CPKSGQGNILVLQT 145	86
2046.9	-LGSQNDDGSTTTVLTSN-LGSQNDDGSTTTVLTSGG -LGSQGGDDGSTTTVLTSN-LGSGANDDGSTTTVLTSN -LGSKGGDDGSTTTVLTSN-LGSQGGDDGSTTTVLTSGG -LGSGANDDGSTTTVLTSGG	Query: 6040 SQNDDGSTTTVLTSN 6054 S N DGSTTTV+TSN Sbjct: 60 SKNGDGSTTTVITSN 74	82
2089.13	-BEXXNLLVLQTATHGVNPGVK-BEXXGGLLVLQTATHGVNPGVK -BETVNLLVLQTATHGVNPGVK-BEXXXXLLVLQTATHGVNPGVK -BEXXXLLVLQTATHGVNPGVK-BEXXNLLVLQTATHRNPGVK -BEXXGGLLVLQTATHRNPGVK	Query: 16929 NLLVLQTATHGVNPGVK 16945 N+LVLQT GVNPGVK Sbjct: 139 NILVLQTTESGVNPGVK 155	81
1884.94	-BXXAVNNQVTGLSTGLNGGAK-BXXAVNNQVTGLSTGLNGQK -BXXXAVNNQVTGLSTGLNGGAK-BXXXAVNNQVTGLSTGLNGQK -BXXAVNNGAVTGLSTGLNGQK-BXXXAVNNGAVTGLSTGLNGQK -BQGAVNNGAVTGLSTGLNGGAK	Query: 5015 BXXAVNNQVTGLSTGLNGQK 5034 + AVNN VT STGL GQK Sbjct: 78 RGVAVNNKVTCTSTGLSGQK 97	76
1436.7	-BGLGWNLDSWFSR-BGLGWGGLDSWFSR -BGLGWNLDSXXFSR-BGLGWNLDSXXM+15.99SR -BGLGXXNLDSWFSR-BGLGXXGGLDSWFSR -BGLGXXNLDSWM+16SR	Query: 25877 GWNLDSWFSR 25886 GW + SWFSR Sbjct: 162 GWSIGSWFSR 171	63
2254.09	-BNFDAETYFSLPFYQXXVNK-BNFDAETYFSLPFXXXXXVGGK -BNFDAETYFSLPFXXXXVGGK-BNFDAETYFSLPFXXXXXVNK -BNFDAETYFSLPFXXXXVNK-BNFDAETYFSLPFYXXXXVGGK -BNFDAETYFSLPFYXXXVGGK	Query: 15862 NFDAETYFS 15870 NFDA TYFS Sbjct: 29 NFDAATYFS 37	61

[a] Only candidate sequences with PepNovo score of 6 or above that produced HSPs with the score of 55 or higher were considered.
[b] MS BLAST conventions: B = R or K (generic trypsin cleavage site); Z = Q or K; L = L or I; M+16 = methionine sulfoxide; X = undetermined amino acid residue, assigned zero score in the substitution matrix.

have the sequence: (K)NGDGSTTTVITSNYISR (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)NDDGSTTTVITSNYISR. 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

Spot	MS/MS spectra		Keratin/trypsin related hits[a]		MS BLAST search time shortened by,
	Removed	Retained	Before filtering	After filtering	
50 kDa	698	1417	395	42	30 min
62 kDa	871	1182	378	124	25 min

[a] "Keratin\trypsin" hits are database entries explicitly annotated as trypsins or keratins (from any species).

Table 3 – Identification of proteins from *Araucaria angustifolia* embryos by a combination of MASCOT and de novo sequencing followed by MS BLAST

Protein MW, kDa	Protein name	Organism	MASCOT			MS BLAST		
			Acc. number	Score[a]	Peptides[b]	Acc. number	HSPs[c]	Total score[d]
Independent identifications by MASCOT and MS BLAST								
50	Elongation factor	*O. sativa*	Q851Y8	503	9	Q8W2C4	12	801
50	Tryptophan synthase	*A. thaliana*	P25269	399	9	P14671	6	485
50	Aspartate aminotransferase	*O. sativa*	Q6KAJ2	348	8	P37833	3	196
50	HSP70	*O. sativa*	Q6L509	358	7	Q9I8F9	8	643
50	Enolase	*S. oleracea*	Q9LEE0	341	7	Q9LEE0	4	321
MASCOT borderline hits validated by de novo sequencing								
62	RuBisCO binding protein, alpha subunit	*A. thaliana*	Q8L5U4	236	4	P21238	10	714
50	Acetyl-CoA C-acyltransferase	*O. sativa*	Q6K3Z3	218	3	Q570C8	15	1072
50	Cysteine desulfurase	*A. thaliana*	Q93WX6	100	2	Q2QTQ1	8	585
50	Alcohol dehydrogenase	*P. banksiana*	Q43020	124	2	Q4JIY8	4	297
50	Heterogeneous ribonucleoprotein	*O. sativa*	Q84ZR9	81	1	Q84ZR9	4	294
MS BLAST identification								
62	Disulfide-isomerase	*O. sativa*	–	–	–	Q43116	12	798

[a] Combined peptide ions score of matched peptides [34].
[b] Number of unique peptides matched by MASCOT.
[c] Number of HSPs with scores above 55.
[d] Sum of all HSPs scores.

searches independently confirmed, respectively, one and four borderline hits in the analysis of 50 kDa and 62 kDa proteins, including incluing 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

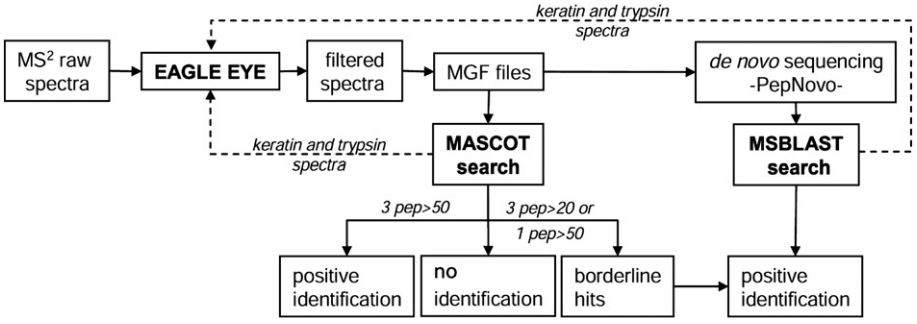

Fig. 5 – Protein identification workflow that uses a combination of MASCOT searches and de novo sequencing followed by MS BLAST searches. First, all spectra were filtered against a background spectra library by EagleEye software, which removed a large number of background MS/MS spectra irrespective of 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 subjected 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 EagelEye 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

[1] Forner F, Foster LJ, Toppo S. Mass spectrometry data analysis in the proteomics era. Current Bioinformatics 2007;2:63–93.
[2] Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal Chem 2004;76:1664–71.
[3] Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discov Today 2004;9:173–81.
[4] Liska AJ, Shevchenko A. Expanding organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. Proteomics 2003;3:19–28.
[5] Liska AJ, Shevchenko A. Combining mass spectrometry with database interrogation strategies in proteomics. Trends Anal Chem 2003;22:291–8.
[6] Standing KG. Peptide and protein de novo sequencing by mass spectrometry. Curr Opin Struct Biol 2003;13:595–601.
[7] Habermann B, Oegema J, Sunyaev S, Shevchenko A. The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. Mol Cell Proteomics 2004;3:238–49.
[8] Katz A, Waridel P, Shevchenko A, Pick U. Salt-induced changes in the plasma membrane proteome of the halotolerant alga *Dunaliella salina* as revealed by Blue-Native gel electrophoresis and nanoLC-MS/MS analysis. Mol Cell Proteomics 2007;6:1459–72.
[9] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 1994;66:4390–9.
[10] Tabb DL, Saraf A, Yates 3rd JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 2003;75:6415–21.
[11] Sunyaev S, Liska AJ, Golod A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal Chem 2003;75:1307–15.
[12] Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 1997;11:1067–75.
[13] Huang L, Jacob RJ, Pegg SC, Baldwin MA, Wang CC, Burlingame AA, et al. Functional assignment of the 20S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. J Biol Chem 2001;17:28327–39.
[14] Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, et al. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal Chem 2001;73:1917–26.
[15] Mackey AJ, Haystead TAJ, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics 2002;1:139–47.
[16] Chalkley RJ, Baker PR, Huang L, Hansen KC, Allen NP, Rexach M, et al. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. new developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol Cell Proteomics 2005;4:1194–204.
[17] Searle BC, Dasari S, Turner M, Reddy AP, Choi D, Wilmarth PA, et al. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal Chem 2004;76:2220–30.
[18] Liska AJ, Popov AV, Sunyaev S, Coughlin P, Habermann B, Shevchenko A, et al. Homology-based functional proteomics by mass spectrometry: application to the *Xenopus* microtubule-associated proteome. Proteomics 2004;4:2707–21.
[19] Shevchenko A, Sunyaev S, Liska A, Bork P, Shevchenko A. Nanoelectrospray tandem mass spectrometry and sequence similarity searching for identification of proteins from organisms with unknown genomes. Methods Mol Biol 2003;211:221–34.
[20] Waridel P, Frank A, Thomas H, Surendranath V, Sunyaev S, Pevzner P, et al. Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated de novo sequencing. Proteomics 2007;7:2318–29.
[21] Grossmann J, Fischer B, Baerenfaller K, Owiti J, Buhmann JM, Gruissem W, et al. A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments. Proteomics 2007;7:4245–54.

[22] Gentzel M, Kocher T, Ponnusamy S, Wilm M. Preprocessing of tandem mass spectrometric data to support automatic protein identification. Proteomics 2003;3:1597–610.
[23] Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, Strupat K, et al. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal Chem 2006;78:2113–20.
[24] Shevchenko A, Wilm M, Vorm O, Mann M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal Chem 1996;68:850–8.
[25] Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 2006;1:2856–60.
[26] Junqueira, M, Spirin, V, Balbuena, TS, Waridel, P, Surendranath, V, Kryukov, G, Adzhubei, I, Thomas, H, Sunyaev, S, Shevchenko, A. Separating the wheat from the chaff: unbiased filtering of background tandem mass spectra improves protein identification. J Proteome Res: in press. doi:10.1021/pr800140v.
[27] Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem 2005;77:964–73.
[28] Parker KC, Garrels JI, Hines W, Butler EM, McKee AH, Patterson D, et al. Identification of yeast proteins from two-dimensional gels: working out spot cross-contamination. Electrophoresis 1998;19:1920–32.
[29] Charneau S, Junqueira M, Costa CM, Pires DL, Fernandes ES, Bussacos AC, et al. The saliva proteome of the blood-feeding insect *Triatoma infestans* is rich in platelet-aggregation inhibitors. Int J Mass Spectrom 2007;268:265–76.
[30] Gache V, Waridel P, Luche S, Shevchenko A, Popov AV. Purification and mass spectrometry identification of microtubule-binding proteins from *Xenopus* egg extracts. Methods Mol Med 2007;137:29–43.
[31] Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA. De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res 2007;6:114–23.
[32] Wielsch N, Thomas H, Surendranath V, Waridel P, Frank A, Pevzner P, et al. Rapid validation of protein identifications with the borderline statistical confidence via de novo sequencing and MS BLAST searches. J Proteome Res 2006;5:2448–56.
[33] Gish W. WU-BLAST2.0; 1996.
[34] Barrett MP, Burchmore RJ, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, et al. The trypanosomiases. Lancet 2003;362:1469–80.
[35] Stefenon VM, Gailing O, Finkeldey R. Genetic structure of *Araucaria angustifolia* (Araucariaceae) populations in Brazil: implications for the in situ conservation of genetic resources. Plant Biol (Stuttg) 2007;9:516–25.