JOURNAL OF PROTEOMICS 72 (2009) 137–144

available at www.sciencedirect.com

www.elsevier.com/locate/jprot

Review

Tools for exploring the proteomosphere

Andrej Shevchenko[a,*], *Cristina-Maria Valcu*[b,1], *Magno Junqueira*[a]

[a]*Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany*
[b]*Section of Forest Genetics, Technical University Munich, Am Hochanger 13, 85354 Freising, Germany*

ARTICLE DATA

ABSTRACT

Homology-driven proteomics aims at exploring the proteomes of organisms with unsequenced genomes that, despite rapid genomic sequencing progress, still represent the overwhelming majority of species in the biosphere. Methodologies have been developed to enable automated LC-MS/MS identifications of unknown proteins, which rely on the sequence similarity between the fragmented peptides and reference database sequences from phylogenetically related species. However, because full sequences of matched proteins are not available and matching specificity is reduced, estimating protein abundances should become the obligatory element of homology-driven proteomics pipelines to circumvent the interpretation bias towards proteins from evolutionary conserved families.

Contents

1. Introduction

Proteomics has become an integral part of molecular and cell biology (reviewed in [1–3]). One, however, should admit that, despite remarkable progress in both mass spectrometry technologies and bioinformatics, the significance of proteomics methodologies for biological research has been granted by the pace, precision and scope of sequencing and interpretation of genomes.

In a typical LC-MS/MS or MALDI MS/MS experiment thousands of tandem mass spectra are acquired from the majority of detectable peptide precursor ions. Computational post-processing converts MS/MS spectra into peak lists, which are then submitted to database searches by the dedicated

* *Corresponding author.*
E-mail address: shevchenko@mpi-cbg.de (A. Shevchenko).
[1] Present address: Department of Pediatric Cardiology and Congenital Heart Disease, German Heart Centre, Technical University Munich, Lazarettstraße 36, 80636 München, Germany.

1874-3919/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2009.01.012

software (reviewed in [4–7]). Irrespectively of its algorithmic basis, the software fulfils two major functions: first, it provides a numerical estimate of the likeness of a particular MS/MS spectrum to its *in silico* representation computed from the database peptide sequence using empirical fragmentation models. Second, it evaluates their similarity in respect to random matches and, hence, computes the estimate of probability that this match is correct. Despite numerous algorithmic, statistical and bioinformatic limitations, this is an efficient and generic approach that underpins the majority of today's proteomics efforts.

We underscore that a typical database search does not really interpret the spectrum. Instead, it tentatively annotates the observed fragment ions and ultimately relies upon the scored similarity between the observed and predicted patterns. Not surprisingly, a very high specificity of spectrum-to-sequence correlation requires obeying stringent matching constraints. Any discrepancy between the actual and database sequences — regardless, if it originates from an amino acid substitution, unexpected post-translational modification or peculiar fragmentation pathway, might deny the peptide identification completely. To produce a hit, the database search would then require more permissive (and, consequently, much less specific) error-tolerant searches. Therefore, conventional proteomics approaches favor the organisms for whom large and accurate database sequence resources or fully sequenced genome are available.

These organisms, however, are not even remotely representing the full biosphere complexity, while much can be learned by exploring a *proteomosphere* in its entirety. Indeed, classic biochemical and enzymological techniques help to purify and assay interesting activities regardless if the exact sequences of corresponding protein factors are known. The prospective lines of research include (yet, by far, are not limited to) the adaptation mechanisms in animals and plants; infection of plants by viruses and pests; exploring venoms and natural biofluids as a source of pharmacologically important enzymes and proteinous bioregulators, among others.

Challenges and expected benefits of expanding the organismal scope of proteomics have been debated [8–10]. Rapid increase in size and organismal representation of sequence databases suggested that a sizable number of proteins from any free living species should be identifiable on the basis of their similarity to already known protein sequences. Although much has been accomplished in both technology developments and applications, new challenges have emerged. In this review, we would like to re-access the achievements and challenges of the homology- driven proteomics and discuss possible instrumentation and bioinformatic solutions.

2. Technical aspects of homology — driven protein identifications

A typical scenario in homology-driven proteomics considers the identification of unknown (i.e. not present in a database) proteins, assuming that a database resource contains distantly homologous protein sequences from phylogenetically related species. Therefore, the analysis more commonly hits a family of homologous proteins in several species, rather than a unique protein.

Homologous proteins are identified by tandem mass spectrometry via several data mining strategies (reviewed in [11]) (Fig. 1).

If the analyzed protein belongs to a conserved protein family, it is likely to comprise several peptides that are fully identical to reference sequences in a database. Its cross-species identification does not differ from the identification of known (i.e. present in a database) proteins by conventional (stringent) searching means, albeit the achieved sequence coverage might be compromised. The more peptides are sequenced in MS/MS experiments, the larger dissimilarity to the reference proteins can be tolerated, while protein assignments would be gradually deviating from the specific protein hit to a larger protein class sharing only a few identical peptide sequences.

If specifically requested, several conventional algorithms can also tolerate a single amino acid mismatch within the compared peptide sequences or disregard the enzyme cleavage specificity (Fig. 1). Protein identification still proceeds with uninterpreted MS/MS spectra. However, because of alleviated matching specificity constraints, the search space dramatically increases and lengthens database searches. Moreover, maintaining the same false discovery rate requires to employ higher confidence thresholds, which concomitantly increases the false negative rate.

The concept of sequence tag searches (Fig. 1) assumes that a single mismatched piece of peptide sequence (regardless of its length) is adjacent to the N- or C-terminus of the analyzed peptide, while the rest of its sequence is identical to the peptide in a database. A peptide sequence tag – a sequence stretch of, typically, 2 to 4 amino acid residues – could be deduced from the MS/MS spectrum and, together with *m*/*z* of the corresponding fragment ions, employed in error-tolerant searches [12]. One part of the tag – either the sequence itself or one of the two flanking *m*/*z* values – is allowed to mismatch.

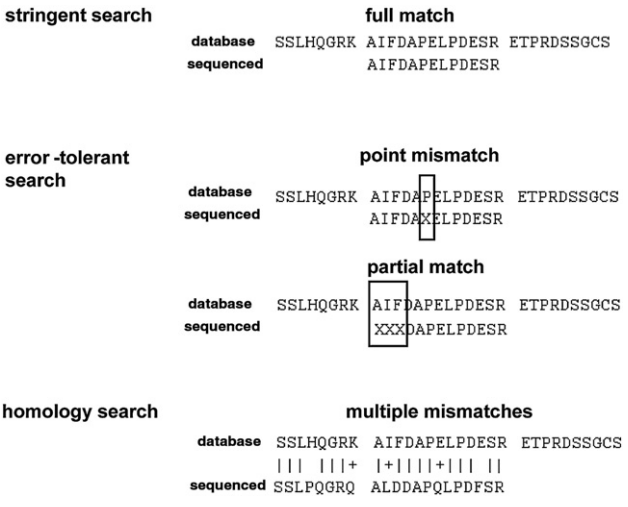

Fig. 1 – Matching peptide sequences in homology-driven proteomics. The choice of optimal identification algorithm depends on the similarity between analyzed and reference proteins, as well as the quality of available MS/MS spectra. X symbol stands for any amino acid residue. The local dissimilarity regions are boxed.

Therefore, error-tolerant searches typically produce large hit lists requiring mandatory manual inspection. The search specificity could be improved and manual inspection alleviated, if hit lists of multiple error-tolerant searches with several sequence tags produced from each of the fragmented peptides are combined and evaluated together using a dedicated statistical model [13]. It is also possible to automatically deduce tags and submit them to database searches [14,15]. Compared to a single amino acid residue mismatch, sequence tag searches are far more permissive, although they do require that matched peptides share a sizable piece of identical sequence. Importantly, it is usually straightforward to retrieve short sequence stretches even from marginal quality MS/MS spectra, which are not amenable to accurate de novo interpretation. This improves the sensitivity and sequence coverage of the analysis because it encompasses more spectra of diverse quality. However, sequence tag searches are not much used in homology-driven proteomics, probably because no integrated software is available that could cover all aspects of the protein identification pipeline starting from generating tags till estimating the confidence of database searching hits despite algorithmic solutions have been developed for each of its individual elements.

Alignments of full length sequences of analyzed peptides with reference sequences could tolerate multiple mismatches and sequence stretches identical between the two compared peptides can be short and scattered along the peptide backbone [16] (Fig. 1). Yet, it is not straightforward to obtain sequences amenable for such comparison. Despite high mass accuracy and mass resolution of instruments [17] that advance spectra interpretation algorithms [18–20], de novo sequencing still represents a considerable challenge. Tandem mass spectra are reflecting sequences of fragmented peptides, rather than not representing them accurately and completely and certain *m/z* regions are devoid of useful fragments. It is often possible to retrieve a few pieces of peptide sequence out of any MS/MS spectrum (reviewed in [10,21]), yet they are usually ambiguous and composed of intermingling correct and false stretches. Therefore, de novo interpretation typically yields a number of sequence candidates (rather than a single unique sequence), which are usually short (10–15 amino acid residues), redundant, degenerate and partially incorrect [22,23]. Their relative location at the backbone of analyzed protein is not known and, furthermore, peptides sequenced in the same experiment might originate from different proteins. Although algorithms for aligning protein sequences and registering their homology have been perfected by bionformatics over decades [24,25], peptide sequence candidate produced by automated de novo interpretation of MS/MS spectra cannot be fed to BLAST or FASTA engines directly without compromising search specificity. Several database searching tools, which handle large numbers of de novo sequence candidates have been developed [26–29]. Some of them are web-accessible, like (MS BLAST) [16,30] or FASTS [29], or are now available as a part of commercial software packages (ProBLAST from Applied Biosystems [31]; MS BLAST option in BioTools from Bruker Daltonics).

Sequence similarity searches employ peptide sequence candidates (rather than raw MS/MS spectra) and are not influenced directly by the instrument type. It is only important that sequence predictions are reasonably accurate (although several sequence variants per each interpreted spectrum are allowed) and de novo sequencing output is formatted into a query string according to a few simple conventions. Large search queries are allowed: while they could contain candidate sequences deduced from an unlimited number of interpreted spectra, their full sizes are limited by 150,000 amino acid resides in total, which is roughly equivalent to a BLAST search with 16.5 MDa protein chimera [30]. Therefore, it has become technically possible to combine automated sequence similarity searches with sensitive LC-MS/MS analysis, which used to produce thousands of MS/MS spectra per each run. Now it might be also practical to subject all acquired spectra to de novo interpretation: besides sequence similarity searches, de novo sequencing also assists in validating "conventional" peptide hits having marginal statistical confidence [32,33].

However, we see several reasons why "full de novo sequencing" approach, if taken carelessly, might be misleading.

One of the most common concerns is that poor quality MS/MS spectra might yield fully incorrect interpretations and, to control the false positive rate, such spectra should be removed prior to database searches by some quality filter [34,35]. This, however, appears to be less of a danger. Any de novo sequencing software offers an arbitrary quality score, which reflects the confidence of produced sequence candidates. Using computational experiments with simulated spectra datasets, it is relatively straightforward to determine arbitrary cut-off scores, below which the produced sequences have almost no chance of hitting correct target proteins in error-tolerant alignments (although even very high de novo quality scores do not guarantee that this peptide(s) will produce a hit). Hence, unreliable sequence candidates could be sorted out based on their arbitrary de novo confidence and, if necessary, added back to the search query later to increase the sequence coverage of already made confident hits or to consider a few more marginal alignments.

Second, by far more serious, concern relates to the reduced matching specificity of sequence similarity searches. A large fraction of MS/MS spectra acquired in a typical LC-MS/MS run under automated data-dependent control could be directly attributed to peptides originating from background proteins — trypsin, human keratins, cell medium components, etc and even more of them could be matched via error-tolerant searches [36,37]. While peptides from background proteins do not harm conventional protein identifications, they strongly affect sequence similarity searches by hitting a very large number of totally unrelated proteins to which they might bear some local similarity. For example, trypsin autolysis products used to hit a multitude of serine proteases of diverse functional specificity and species of origin. Because of low complexity sequence stretches, keratin peptides could match almost any protein in a database with varying degree of statistical confidence. We note that conventional database searches, performed prior to de novo interpretation of spectra only remove a relatively small fraction of background spectra [36] with little effect on sequence similarity hit lists.

As a practical solution, it was proposed to subtract such MS/MS spectra by comparing them to a library of non-annotated background spectra, compiled from control and

blank LC-MS/MS runs [36,38]. Spectra screening software is available as a stand-alone application, or at the public web server [30]. Depending on the relative abundances of background and target precursors, filtering removes up to a half of the total number of acquired MS/MS spectra, while no high quality spectra originating from bona fide target proteins are lost. Importantly, filtering does not rely on sequence database resources and exactly the same algorithm equally applies for processing spectral datasets acquired from known or unknown proteins [36].

We underscore that sequence similarity searches tolerate the compromised accuracy of candidate sequences and therefore no chemical modification of peptides that support de novo interpretations [39–44] is required in most cases. Therefore, the same data acquisition and sample processing routines could be employed at the uncompromised detection sensitivity, regardless if the analyzed sample might contain known or unknown proteins. For most efficient analysis, multiple datamining routines could be combined into a "layered" proteomics pipeline, which supports flexible manipulation with individual MS/MS spectra [28,30].

Taken together, it is reasonable to assume that major technical issues in homology-driven protein identifications have now been addressed. Convenient workable software solutions for data pre-processing, de novo sequencing and sequence-similarity searches are available and could be integrated into automated proteomics pipelines supporting any instrument type with adequate MS/MS capabilities.

3. Scope and applications of homology driven proteomics

The success rate of cross-species identifications depends on three major factors: first, on the phylogenetic distance between the analyzed and the reference organism(s); on the proteome coverage of the reference organism in a database and on the number of peptides sequenced from the analyzed protein. Computational modeling suggested that, using a dataset of 10 to 15 de novo sequenced peptides, sequence similarity searches should be able to match 50 to 60% of proteins having at least 50% of the sequence identity with the reference proteins [45]. Furthermore, because of relatively small phylogenetic divergence between individual species and availability of several completed genomes, it should be possible to identify >75% of proteins within mammalian and plant kingdoms on the proteome -wide scale [45,46]. LC-MS/MS sequencing of 10 to 20 precursors is usually unproblematic if the target protein is present at the low picomole level and therefore the above estimates of proteome coverage scope are certainly within the practical reach.

We note, however, that in phylogeneticaly distant species the similarity of protein sequences does not necessarily imply their functional similarity. Therefore, using homology- driven protein identifications for ontology annotations of corresponding protein products should be performed with great caution [47,48].

There are numerous examples of successful application of sequence-similarity identifications in plant and animal biology, microbiology, toxicology, environmental studies, to mention just a few areas (reviewed in [49–51]). Whereas most abundant proteins typically come from conserved protein families and are readily identified by conventional searches, in our experience sequence similarity searches usually add more than 25% of new identifications [46,52–58]. By matching more peptides, sequence-similarity searches increase the sequence coverage and identification confidence of hits made by conventional searches, which is helpful in distinguishing protein isoforms and/or individual members of protein families.

Among wild-bred species pronounced polymorphism of protein sequences occurs [59]. Even in a digest of the same protein spot, it is not uncommon to detect several sequence variants of the same tryptic peptide, which differ by a single amino acid between themselves and between the database reference sequence (see, for example [55]). Although polymorphism effectively precludes their identification by conventional database searching means, registering sequence commonalities by error-tolerant searches is usually unproblematic.

In conclusion, we note that the progress in EST and genomic sequencing enhances the scope and practical usability of homology-driven proteomics, rather than making it redundant and obsolete. First, since novel reference genomes appear from previously underrepresented phylogenetic kingdoms, more and more species and classes become amenable to proteomic characterization [9]. Second, homology-driven proteomics effectively bridges reference genome sequences with the natural sequence variability of wild living species.

4. Protein quantification: obligatory element of a homology-driven proteomic pipeline

Identification confidence is not directly affected by the available protein quantity and several database search engines disregard the actual abundance of precursor and/or fragment ions, solely operating with normalized peak intensities. While analyzing protein mixtures, we also presume that expert consideration of physico-chemical properties of proteins and fine tuning data acquisition and database searching settings will result in adequate protein representation in the hit lists. If necessary, relative quantities of individual protein components could be estimated using the abundances of matched peptide precursors or by MS/MS spectral counts or related indices (reviewed in [1,60–62]).

This, however, is only true if complete sequences of analyzed proteins are present in a database. In a mixture of known and unknown proteins conventional searches will only identify the former, despite the latter protein might constitute the major component. In the case of stringent searches, higher abundance of precursor peaks typically translates into a better quality of corresponding MS/MS spectra and they are more often matched, rather than falsely mismatched. However, if the analysis identifies a protein with unknown sequence by a few matched peptides, it is quite common that several high quality spectra still remain unmatched. It is not possible to tell if they originate from an uncovered piece of sequence of the already identified protein, or they indicate that major protein

Table 1 – Proteins identified in a spot from *Fagus sylvatica* by MASCOT and MS BLAST searches

	Protein	Accession no.	MASCOT search*				MS BLAST search		
			Total Score	Matched spectra	Unique peptides	Sequence coverage, %	Total Score	Matched spectra	Unique peptides
GTM	γ tocopherol methyltransferase	AAX63740	160	4	4	13			
Pdx	Pyridoxine biosynthesis protein	NP_195761	110	2	2	8	82	4	1
LHCP	Light harvesting chlorophyll a/b binding protein	CAA48410	73	2	2	8			
USP	Universal stress protein (USP)	BAA97516	68	1	1	4			
RBP	Putative RNA binding protein	NP_181084					269	18	4

*Searches were performed against NCBI; 1 missed cleavage; mass tolerance: precursor ions: 10 ppm; fragment ions: 0.6 Da; fixed modifications: carbamidomethyl (C); variable modifications: acetyl (N-terminal), oxidation (M), phosphorylation (S,T,Y).

component(s) remain unidentified by all database searching means.

The case study presented below exemplifies the problem severity. The abundance of a silver-stained spot with apparent MW of 38.1 kDa and *pI* 5.42 visualized on a two-dimensional gel of the protein extract from leaves of the European beech *Fagus sylvatica* was decreased upon infecting the tree with root pathogen *Phytophthora citricola*. The corresponding spots isolated from control and infected plants were excised from the gels, digested with trypsin and analyzed by LC-MS/MS on a hybrid LTQ Orbitrap mass spectrometer. MASCOT searches produced four confident cross-species hits (Table 1), while homology-driven proteomics pipeline [30] additionally identified an RNA binding protein based on its homology to known *Arabidopsis thaliana* and *Oryza sativa* proteins (Fig. 2).

There is no robust methodology to determine the abundance of proteins in a mixture without internal standards [63,64]. Therefore we further followed the idea of Silva et al [65] and, solely as a ballpark estimate, inferred the relative amounts of proteins from the normalized intensities of most abundant precursors matched to correspondent sequences (Fig. 3). According to our estimate, RNA binding protein – only identified by MS BLAST searches – represented a major component of the spot, which was further supported by a large difference in MS/MS spectra count indices, which was 4.5 for peptides from RNA binding protein compared to 1 for peptides from other proteins identified in the spot digest. Its abundance dropped following the pathogen infection and is probably responsible for the apparent decrease in the spot intensity, despite the abundance of other minor components increased (Fig. 3).

	Control plant		**Infected plant**	
Protein	RNA-binding protein cp33, putative		Protein similar to RNA-binding protein	
Species/Accession	Oryza sativa / EAZ31815		Arabidopsis thaliana / NP_181084	
	Total Score: 214		Total Score: 269	
m/z	High scoring segment pairs	Score	High scoring segment pairs	Score
961.970	Query: 3826 TNEEAEAALSALDGK 3840 T EEAEAAL LDGK Sbjct: 166 TKEEAEAALTELDGK 180	77		
1253.095	Query: 4820 VSFATNEEAEAAL 4832 VSF T EEAEAAL Sbjct: 162 VSFGTKEEAEAAL 174	73	Query: 5962 VSFATNEEAEAAL 5974 VSFAT EEAE A+ Sbjct: 241 VSFATREEAENAI 253	66
1021.990	Query: 4272 TMESAEEAQAVVEK 4285 TM AEEA A VEK Sbjct: 69 TMATAEEAAAAVEK 82	66	Query: 4605 VTMESAEEAQAVLDK 4619 VTM S EEAQA +DK Sbjct: 140 VTMASGEEAQAAIDK 154	81
954.025			Query: 6198 NLPWSMSV 6205 NLPWSMSV Sbjct: 101 NLPWSMSV 108	64
821.445			Query: 1702 BHRLYVVNTAWXVR 1715 +H LYV N AW R Sbjct: 193 RHKLYVSNLAWKAR 206	58

Fig. 2 – Identification of RNA binding protein in the same spots that were excised from control and treatment gels by LC-MS/MS, automated de novo sequencing and MS BLAST search. Scores were assigned to corresponding high scoring segment pairs (HSPs) by the BLAST engine.

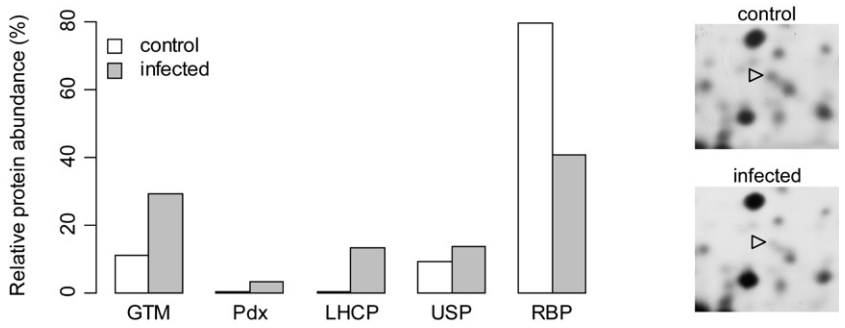

Fig. 3 – Differences in the relative abundance of five protein components identified in the silver stained spot (designated with the pointer at the inset images) in protein extracts of control and infected plants. Full protein names are provided in Table 1.

Hence, even qualitative interpretation of the LC-MS/MS analysis ultimately requires to determine if it identified the major component(s) in the sample, or only minor, yet relatively conserved proteins were hit. In our experience, this is the most serious bottleneck of the current homology-driven proteomics efforts. Effectively, it requires quantifying proteins with unknown sequences and relating most abundant, yet unmatched, precursors to identified components. Although it might be impossible to offer a generic solution, we could think of several empirical approaches partially addressing these issues. For example, quantitative comparison of peptide profiles acquired in several independent experiments might reveal coherent changes in abundances of certain precursor peaks [66], irrespectively if they have (or have not) been matched to identified proteins. Alternatively, large species-specific libraries of non-annotated reference MS/MS spectra could be composed by the systematic analysis of samples with known content and used for identifying newly acquired spectra independently of sequence database resources [36,38,67,68].

5. Conclusions

Homology-driven proteomics is a mature field that enjoys steady expansion of its application scope. By inferring identities of unknown proteins, it provides insight into the molecular mechanisms of complex biological phenomena that were previously addressed by descriptive phenomenological approaches. It has become possible to determine the representative molecular composition of proteomes of biological fluids, tissues or organs obtained from a multitude of free-living species and quantify their dynamic response to a variety of endogenous or environmental stimuli. The techniques supporting sequence similarity identifications are generic and can be incorporated into any proteomics pipeline. We anticipate that, in meantime, sequence similarity identifications should become routine and commonly used by proteomics laboratories studying organisms with unknown genomes and/or wild-bred species. The reviewing policies might request the obligatory use of both stringent and error-tolerant searches as a prerequisite of adequate representation of the reported proteome composition.

It has also become obvious that homology driven protein assignments should be supported with some quantitative evidence, since sequence-similarity interpretations are inherently biased towards matching known or most conserved proteins. On the positive side, massive efforts in developing label-free quantifications methods are on-going and corresponding software algorithms for unbiased extraction of precursor intensities will be of considerable value for validating homology-driven identifications.

Despite apparent significance of homology-driven proteomics for pharmacological, environmental, conservational, agricultural, among many other fields of biological science, there might be some less explored, yet intriguing opportunities. In particular, it seems exciting to build a tighter alliance with developmental and evolutionary biology. In our (may be, somewhat biased) opinion these fields “stagnate” within the realm of established model organisms amenable to genetic manipulations. Yet, it is conceivable that studying the cellular molecular machinery on a wider organismal scope might dramatically improve our understanding of the evolution and adaptation mechanisms.

Acknowledgements

Application field of homology-driven proteomics is large and diverse and, because of the space constraints, the review could not encompass all essential work in the field. We therefore sincerely apologize for not being able to cite and discuss many important contributions of our colleagues. We are grateful to members of Shevchenko laboratory their support and stimulating discussions. This work was, in part, supported by NIH NIGMS grant 1R01GM070986-01A1 to A.S.

REFERENCES

[1] Cox J, Mann M. Is proteomics the new genomics? Cell 2007;130:395–8.
[2] Domon B, Aebersold R. Mass spectrometry and protein analysis. Science 2006;312:212–7.
[3] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.

[4] Shadforth I, Crowther D, Bessant C. Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. Proteomics 2005;5:4082–95.
[5] Hernandez P, Muller M, Appel RD. Automated protein identification by tandem mass spectrometry: issues and strategies. Mass Spectrom Rev 2006;25:235–54.
[6] Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol 2007;367:87–119.
[7] Palagi PM, Hernandez P, Walther D, Appel RD. Proteome informatics I: bioinformatics tools for processing experimental data. Proteomics 2006;6:5435–44.
[8] Liska AJ, Shevchenko A. Expanding organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. Proteomics 2003;3:19–28.
[9] Liska AJ. The morality of problem selection in proteomics. Proteomics 2004;4:1929–31.
[10] Standing KG. Peptide and protein de novo sequencing by mass spectrometry. Curr Opin Struct Biol 2003;13:595–601.
[11] Liska AJ, Shevchenko A. Combining mass spectrometry with database interrogation strategies in proteomics. Trends Anal Chem 2003;22:291–8.
[12] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 1994;66:4390–9.
[13] 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.
[14] Tabb DL, Saraf A, Yates 3rd JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 2003;75:6415–21.
[15] Frank A, Tanner S, Bafna V, Pevzner P. Peptide sequence tags for fast database search in mass-spectrometry. J Proteome Res 2005;4:1287–95.
[16] 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.
[17] Scigelova M, Makarov A. Orbitrap mass analyzer — overview and applications in proteomics. Proteomics 2006;6:16–21.
[18] 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.
[19] Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA. Proteomics-grade de novo sequencing approach. J Proteome Res 2005;4:2348–54.
[20] Mo L, Dutta D, Wan Y, Chen T. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. Anal Chem 2007;79:4870–8.
[21] Johnson RS, Davis MT, Taylor JA, Patterson SD. Informatics for protein identification by mass spectrometry. Methods 2005;35:223–36.
[22] Pitzer E, Masselot A, Colinge J. Assessing peptide de novo sequencing algorithms performance on large and diverse data sets. Proteomics 2007;7:3051–4.
[23] Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X. Performance evaluation of existing de novo sequencing algorithms. J Proteome Res 2006;5:3018–28.
[24] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[25] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] Junqueira M, Spirin V, Balbuena TS, Thomas H, Adzhubei I, Sunyaev S, et al. Protein identification pipeline for the homology-driven proteomics. J Proteomics 2008;71:346–56.
[31] Arif SA, Hamilton RG, Yusof F, Chew NP, Loke YH, Nimkar S, et al. Isolation and characterization of the early nodule-specific protein homologue (Hev b 13), an allergenic lipolytic esterase from *Hevea brasiliensis* latex. J Biol Chem 2004;279:23933–41.
[32] Thomas H, Shevchenko A. Simplified validation of borderline hits of database searches. Proteomics 2008;8:4173–7.
[33] 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.
[34] Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. Mol Cell Proteomics 2005;4:1180–8.
[35] Bern M, Goldberg D, McDonald WH, Yates 3rd JR. Automatic quality assessment of peptide tandem mass spectra. Bioinformatics 2004;20(Suppl 1):I49–54.
[36] Junqueira M, Spirin V, Santana Balbuena T, Waridel P, Surendranath V, Kryukov G, et al. Separating the wheat from the chaff: unbiased filtering of background tandem mass spectra improves protein identification. J Proteome Res 2008;7:3382–95.
[37] Chalkley RJ, Baker PR, Hansen KC, Medzihradszky KF, 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: I. How much of the data is theoretically interpretable by search engines? Mol Cell Proteomics 2005;4:1189–93.
[38] Yates 3rd JR, Morgan SF, Gatlin CL, Griffin PR, Eng JK. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem 1998;70:3557–65.
[39] Shevchenko A, Chernushevich I, Ens W, Standing KG, Thomson B, Wilm M, et al. Rapid 'De Novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. Rapid Commun Mass Spectrom 1997;11:1015–24.
[40] Uttenweiler-Joseph S, Neubauer G, Christoforidis S, Zerial M, Wilm M. Automated de novo sequencing of proteins using the differential scanning technique. Proteomics 2001;1:668–82.
[41] Samyn B, Sergeant K, Memmi S, Debyser G, Devreese B, Van Beeumen J. MALDI-TOF/TOF de novo sequence analysis of 2-D PAGE-separated proteins from *Halorhodospira halophila*, a bacterium with unsequenced genome. Electrophoresis 2006;27:2702–11.
[42] Keough T, Youngquist RS, Lacey MP. A method for high-sensitivity peptide sequencing using postsource decay matrix-assisted laser desorption ionization mass spectrometry. Proc Natl Acad Sci U S A 1999;96:7131–6.

[43] Schnölzer M, Jedrzejewski P, Lehmann WD. Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. Electrophoresis 1996;17:945–53.
[44] Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, et al. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. Rapid Commun Mass Spectrom 2001;15:1214–21.
[45] 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.
[46] 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.
[47] Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
[48] Rappsilber J, Mann M. What does it mean to identify a protein in proteomics? Trends Biochem Sci 2002;27:74–8.
[49] Carpentier SC, Panis B, Vertommen A, Swennen R, Sergeant K, Renaut J, et al. Proteome analysis of non-model plants: a challenging but powerful approach. Mass Spectrom Rev 2008;27:354–77.
[50] Padliya ND, Cooper B. Mass spectrometry-based proteomics for the detection of plant pathogens. Proteomics 2006;6:4069–75.
[51] Pandhal J, Wright PC, Biggs CA. Proteomics with a pinch of salt: a cyanobacterial perspective. Saline Syst 2008;4:1.
[52] 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.
[53] 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.
[54] Guercio RA, Shevchenko A, Lopez-Lozano JL, Paba J, Sousa MV, Ricart CA. Ontogenetic variations in the venom proteome of the Amazonian snake *Bothrops atrox*. Proteome Sci 2006;4:11.
[55] Shevchenko A, Leal de Sousa MM, Waridel P, Bittencourt ST, Valle de Sousa M, Shevchenko A. Sequence similarity-based proteomics in insects: characterization of the larvae venom of the Brazilian moth *Cerodirphia speciosa*. J Proteome Res 2005;4:862–9.
[56] Liska AJ, Shevchenko A, Pick U, Katz A. Enhanced photosynthesis and redox energy production contribute to salinity tolerance in dunaliella as revealed by homology-based proteomics. Plant Physiol 2004;136:2806–17.
[57] Fullekrug J, Shevchenko A, Simons K. Identification of glycosylated marker proteins of epithelial polarity in MDCK cells by homology driven proteomics. BMC Biochem 2006;7:8.
[58] Russeth KP, Higgins L, Andrews MT. Identification of proteins from non-model organisms using mass spectrometry: application to a hibernating mammal. J Proteome Res 2006;5:829–39.
[59] Starkweather R, Barnes CS, Wyckoff GJ, Keightley JA. Virtual polymorphism: finding divergent peptide matches in mass spectrometry data. Anal Chem 2007;79:5030–9.
[60] Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 2007;389:1017–31.
[61] Mueller LN, Brusniak MY, Mani DR, Aebersold R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 2008;7:51–61.
[62] Steen H, Pandey A. Proteomics goes quantitative: measuring protein abundance. Trends Biotechnol 2002;20:361–4.
[63] Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Nat Acad Sci USA 2003;100:6940–5.
[64] Beynon RJ, Doherty MK, Pratt JM, Gaskell SJ. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. Nat Methods 2005;2:587–9.
[65] Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics 2006;5:144–56.
[66] Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. Proteomic characterization of the human centrosome by protein correlation profiling. Nature 2003;426:570–4.
[67] Craig R, Cortens JC, Fenyo D, Beavis RC. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 2006;5:1843–9.
[68] Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 2007;7:655–67.