Review

Adam J. Liska
Andrej Shevchenko

Max Planck Institute
of Molecular Cell Biology
and Genetics,
Dresden, Germany

Expanding the organismal scope of proteomics: Cross-species protein identification by mass spectrometry and its implications

Due to the limited applicability of conventional protein identification methods to the proteomes of organisms with unsequenced genomes, researchers have developed approaches to identify proteins using mass spectrometry and sequence similarity database searches. Both the integration of mass spectrometry with bioinformatics and genomic sequencing drive the expanding organismal scope of proteomics.

Keywords: Functional genomics / Genomics / Mass spectrometry / Phylogenetics / Protein complexes / Review PRO 0329

Contents

1 Introduction 19
2 Cross-species protein identification by MS 19
3 Efforts towards the identification of proteins by MS/MS and sequence similarity searches 20
4 Organismal diversity in functional proteomics: orthologous protein complexes and protein interaction networks 22
5 Developments in genomic sequencing and the study of proteomes by MS 24
6 Perspective: Proteomics and phylogenetic considerations for future genomic sequencing . . . 24
7 References 27

1 Introduction

Proteomics is the endeavor to understand gene function and to characterize the molecular processes of the living cell through the large-scale study of proteins found in specific biological contexts. In proteomics, mass spectrometry (MS) has become a powerful analytical technology to identify proteins by the analysis of peptides and the correlation of resultant spectra with available database sequences (reviewed in [1–3]). Genomic sequencing projects, which supply a significant number of sequences for databases, are a relatively new phenomenon in the biological sciences and thus have only a few representative complete genomes to show for their efforts, however amazing the development of these efforts may be [4]. Using established MS techniques, a limited database resource has not been conducive for facile protein identification from organisms with unsequenced genomes. Yet despite the relative deficiency of genomic sequences compared to a whole biosphere of organisms, the emerging interplay of MS and bioinformatics is significantly expanding the organismal scope of proteomics.

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

2 Cross-species protein identification by MS

Irrespective of whether the genome of an organism is sequenced or not, the identification of proteins by MS consists primarily of two analyses of peptides produced by its enzymatic digestion. MALDI-TOF MS produces spectra by the measurement of masses of intact peptides and identifies proteins by the correlation of these masses with theoretically calculated peptides from database entries (see [5] for tutorial). The second type of analysis, tandem mass spectrometry (MS/MS), produces patterns of peptide fragments that can be interpreted and/or correlated to database entries in a number of ways (for tutorial on nanoelectropsray MS/MS see [6], for nanoLC MS/MS [7], and [8] for generic review).

Using MS and available protein sequences, cross-species protein identifications are accomplished by partially aligning an analyzed protein from an organism with an unsequenced genome to a database sequence from a related organism. After DNA sequencing projects began, it became apparent that phylogenetically related organ-

 0173-0835/03/0101–19 $17.50+.50/0

isms have significant genomic sequence colinearity and their proteins have a high degree of homology. However, gene sequences are rarely identical from one species to another and genes are normally riddled with nucleotide substitutions, resulting in amino acid substitutions in proteins. As organisms become more phylogenetically distant from one another, or as certain genes become altered at higher rates, homologous genes and their corresponding proteins retain a lower percent identity.

Peptide mass mapping allows for cross-species protein identification in some cases because only a subset of all peptides from a protein digest need to be recognized [9]. Those peptides that have amino acid substitutions and corresponding shifts in mass are unrecognized and do not contribute to the identification. The theoretical predictions by Wilkins and Williams [9] proposed that proteins could be identified using peptide mass mapping if the analyzed protein and reference database entry have $>80\%$ sequence identity, although they added that these cases would need to be supported by further evidence for validation. The high mass accuracy of modern TOF and FT-MS instruments increases confidence in cross-species peptide mass mapping and loosen the sequence identity requirement, as less peptide masses would be required to produce a confident hit [10].

For proteins with a lower sequence identity compared to available sequences, the more specific MS/MS analysis of peptides provides confident cross-species identifications with a few peptide sequences, depending on the length and significance of their amino acid composition. As amino acid substitutions alter many of the peptides (and their corresponding MS/MS fragmentation patterns) but not all peptides from a protein digest, some can still be non-error-tolerantly correlated to database entries by software tools (reviewed in [11]). For proteins that have yet a lower percent identity compared to database sequences, amino acid substitutions alter every peptide's mass and MS/MS fragmentation pattern. Where peptide mass mapping and non-error-tolerant MS/MS methods fail, the identification of proteins relies primarily on predicting amino acid sequences from MS/MS spectra and using the predicted sequences to identify proteins by their similarity to existing databases entries (Fig. 1).

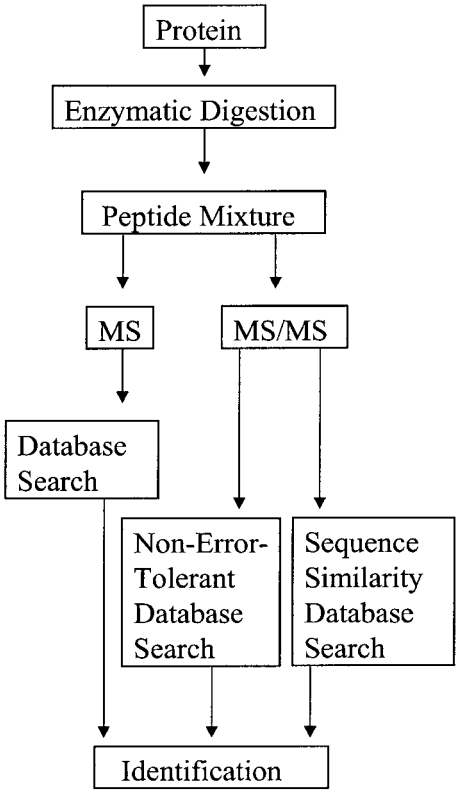

Figure 1. The strategy of cross-species protein identifications by mass spectrometry. Proteins are identified by the analysis of peptides by either MS or MS/MS. A database search follows each analysis. From MS/MS spectra, the less sensitive non-error-tolerant route or the more sensitive sequence similarity search route are used for protein identification depending on the sequence of the analyzed protein and available database resources.

3 Efforts towards the identification of proteins by MS/MS and sequence similarity searches

New instrument configurations like hybrid quadrupole TOF mass spectrometers featured with electrospray and MALDI ion sources (reviewed in [12]), MALDI TOF/TOF [13], and advanced LC MS/MS technology as MudPIT [14] have greatly contributed to protein identification on a large scale. However, the increasing analytical precision of mass spectrometers does not necessarily lead to improved success in the identification of proteins from organisms with unsequenced genomes. A central analytical consideration is the inability to always reconstruct a complete and accurate amino acid sequence from tandem mass spectra of peptides (Fig. 2). Usually these spectra can only be partially interpreted due to the natural under-representation of peptide fragments and because of the presence of chemical noise, which may obscure peptide fragments of low intensity and misguide the interpretation. To overcome this difficulty, researchers have recently developed methods for the chemical derivatization of peptides, and alternate methods of interpreting spectra and database searching.

De novo interpretation of tandem mass spectra relies on measuring the mass differences between adjacent fragment ion peaks of one of the major ion series, *i.e.* b-series

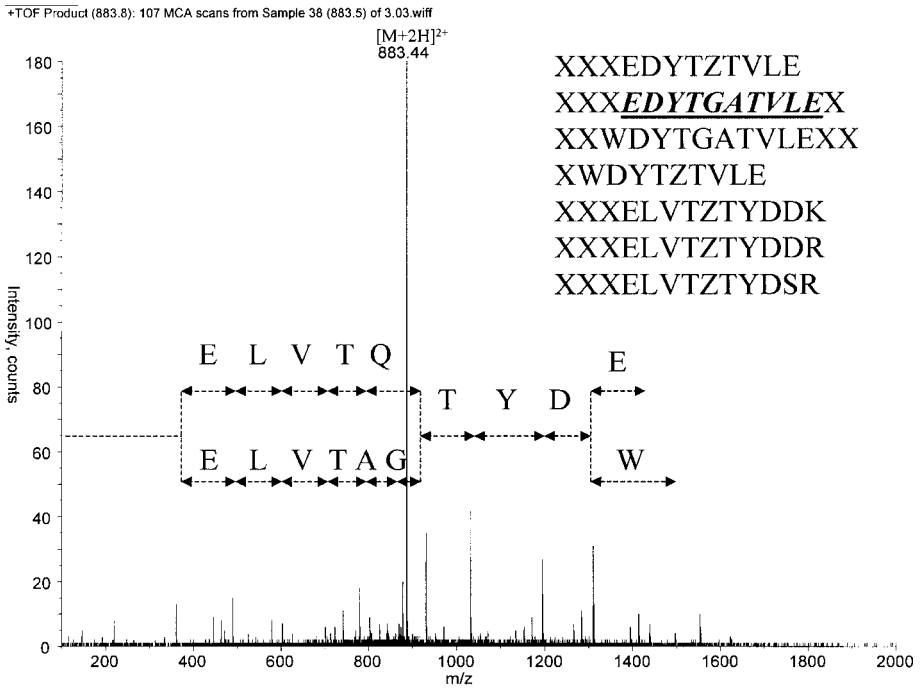

Figure 2. Interpretation of a tandem mass spectrum from a doubly charged precursor ion with *m/z* 883.44 acquired on a quadrupole TOF mass spectrometer. Manual interpretation of spectra considered precise mass difference between adjacent y-ions starting from the *m/z* segment above the precursor ion (corresponding peaks and amino acid residues are designated by arrows). Automated interpretation resulted in a few partially redundant sequences covering the *C*-terminus of the peptide (inset). The underlined sequence was matched to bovine DNA polymerase. The symbol Z represents the amino acid Q or K. The symbol X represents an unknown amino acid.

(ions containing *N*-terminus) or y-series (ions containing *C*-terminus), which are more common in tryptic peptides, resulting in the prediction of the amino acid sequence (see [15] for the nomenclature). Upon their collisional fragmentation, tryptic peptides break at their amide bonds between consecutive amino acid residues producing a continuous y-series of fragments from which the amino acid sequence can be read. One method to facilitate this interpretation is to enrich a series of fragments by attaching a strongly positively or strongly negatively charged group to the *N*-terminus of peptides [16, 17]. Another method is to introduce an isotopic label to the *C*-terminus of peptides by digesting proteins in a buffer containing $H_2{}^{18}O$ (protocols reviewed in [6]) or by CD_3OH [18]. ^{18}O-labeled y-ions can be recognized by a one or two Thomson shift (depending on the peptide's charge) and allow confident readout of the peptide sequence [19, 20]. The quality and the number of produced sequences enabled cloning of a few proteins v*ia* oligonucleotide primers and PCR. However, usually abundant amounts of protein are required, spectra interpretation remains laborious and time consuming, and therefore these approaches have never been applied in large-scale projects.

A second possibility is to interpret tandem mass spectra of peptides using specialized software that creates amino acid sequences *de novo* [21–24]. Although the software utilizes different computational principles, sequences of short peptides can be produced rapidly and accurately. However, less confident sequences and/or incomplete sequences are usually deduced from spectra of large and/or triply charged ions. For each spectrum, the software produces a list of candidate peptide sequences that are ranked in the order of their statistical confidence. However, the absence of a rigorous scoring system may lead to erroneous identifications as the correct sequence may be present in the list, but may not be ranked among the top hits. Even though it is difficult to use these sequences for cloning (where the requirement is that the sequences should be long, 100% accurate, and encode for low degeneracy primers), they can be successfully used for identifying proteins in a sequence database using various sequence similarity search algorithms [23–25].

The use of BLAST [26] or FASTA [27] to analyze peptide sequences produced by interpreting tandem mass spectra is difficult because both algorithms have usually been optimized for comparing long and accurate protein sequences, whereas the interpretation of tandem mass spectra yields sets of inherently redundant and error-prone sequence candidates. Furthermore, it is not known in what order the peptide sequences should be aligned on the backbone, if those sequences belong to a single protein or originate from a few proteins comigrating within a single band (spot), nor what isobaric amino acids (Leu and Ile, Lys and Gln, or Phe and Met-sulfoxide) are present.

To address these difficulties, common database search engines were manipulated to allow the input of sequences produced by MS. Modified FASTA-based software is

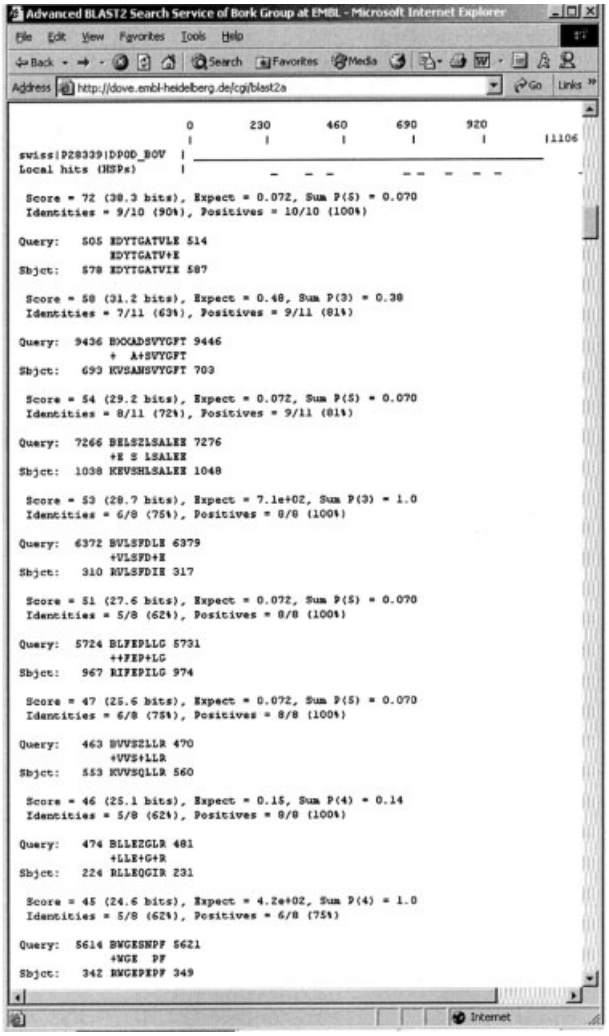

Figure 3. MS BLAST sequence alignment of an analyzed unknown protein from *Xenopus laevis*. Manual and automated *de novo* sequence prediction of 21 tandem mass spectra from fragmented tryptic peptides resulted in 792 putative peptide sequences submitted to a search string. Bovine DNA polymerase was the top hit. Multiple hits from different organisms were retrieved from the MS BLAST search and many were able to make high confidence matches (see Table 1, a))

available as stand-alone applications [23, 28, 29], whereas MS BLAST (Mass Spectrometry driven BLAST) [25] is accessible over the web [30]. The limitations of FASTA-based algorithms is that they are slow search engines and the final score of hits depends not only on the number of matched peptides, but decreases with the number of peptide sequences submitted in a query. This aspect of the software means that spectra must be represented by as few putative amino acid sequences as possible, which is difficult to do because of the inherent ambiguity of automated interpretation of tandem mass spectra. If no hit is found by a predicted sequence used for the search, researchers are unable to ascertain whether the spectra were misinterpreted or no corresponding sequence exists in a database. However, FASTA-based engines are flexible, may engage optional gapped alignment, and the statistical apparatus is specifically tailored for matching short peptide sequences. The MS BLAST software has a particular advantage of being very fast in searching and not penalizing the score of hits for submissions of numerous redundant putative sequence candidates (Fig. 3) [25]. This allows direct submission of the entire output of the sequence prediction software for all fragmented peptides, without intermediate inspection of data and arbitrary selection of the most reliable hits. These qualities allow MS BLAST to be coupled with high-throughput sequencing techniques such as MALDI-QqTOF and LC-MS/MS.

When trying to identify proteins by sequence similarity searches, the number of peptides recognized from a digested protein determines the success of the identification. It has been calculated that as more peptides are analyzed and matched, proteins of less similarity to database sequences can be identified with the limit being around 50% identity (this is dependent on which software is used) [29].

Besides these statistical considerations, when investigating the proteome of an organism with an unsequenced genome, the ability to identify proteins is dependent on the content of available databases. Where an abundance of database sequences exist from closely related organisms, with respect to the organism under inquiry, more homologous genes exist *in silico* to make cross-species identifications possible. If the organism being studied is very distantly related to any organism with a sequenced genome, the likelihood of protein identification decreases.

4 Organismal diversity in functional proteomics: Orthologous protein complexes and protein interaction networks

The availability of genomic sequences and progress in gene manipulation technologies has shifted the focus of functional proteomics from the identification of individual proteins towards deciphering of protein complexes and their place in a global protein interaction network (reviewed in [31]). As protein complexes are often regarded as functional units of the molecular machinery of the cell [32], their characterization provides mechanistic insight into key regulatory processes and facilitates functional interpretation of genomic sequences.

As many cellular functions are conserved throughout a variety of species, it has been inferred that orthologous protein complexes might also share similar composition and architecture [4]. Comparison of three native protein complexes, isolated from budding yeast cells and from human cells by immunoaffinity chromatography, supports this notion [33]. Thus, it is conceivable that conserved protein complexes may be initially characterized in a model organism and then the knowledge obtained be projected on orthologous complexes in other organisms, including humans. We see several lines of evidence why such a strategy will benefit from wider representation of model organisms, which might have unknown or partially sequenced genomes.

A combination of biochemical isolation of protein complexes and mass spectrometric identification of their subunits provides the most detailed characterization of their composition and organization. However, the abundance of orthologous complexes varies greatly between different organisms and cell types (and hence their ability to be identified by MS and so does the completeness of their biochemical characterization. The study of complexes is also facilitated by a multitude of investigative methods that differentially suit distinct specimens. Some organisms are more amenable to genetic manipulation than others, while some are more easily studied under a microscope. The characterization of complexes is best accomplished through the study of more than one organism, applying a set of different investigative methods, with MS being a major participant.

Orthologous protein complexes are seldom identical, even if they comprise subunits with a high degree of homology. For example, a complex of aminoacyl-tRNA synthetases in budding yeast contains three subunits: Glu-tRNA synthetase, Met-tRNA synthetase, and the nonaminoacyl-tRNA synthetase component Arc1p [34]. Despite the fact that orthologous yeast and human aminoacyl-tRNA synthetases share substantial sequence identity, the orthologous complex in higher eukaryotes comprises nine aminoacyl-tRNA synthetases and three nonaminoacyl-tRNA synthetase components: Arg-tRNA synthetase, Asp-tRNA synthetase, Gln-tRNA synthetase, Ile-tRNA synthetase, Leu-tRNA synthetase, Lys-tRNA synthetase, Met-tRNA synthetase, bifunctional Glu-Pro-tRNA synthetase, and p18, p38, p43 [34]. Thus, characterization of the complex only in lower organisms or only in higher organisms provides limited knowledge of its architecture and function.

Most importantly, even if orthologous complexes are very similar in composition, they might be regulated *via* interaction with different, nonorthologous proteins or protein complexes. Orthologous cell cycle regulating ubiquitin ligases in yeast and humans serve as a good example. The SCF complex (term for Skp1 – Cdc53 – F-box protein) is built from conserved core subunits: Skp1, cullin homologue Cdc53, and RING H2 subunit Hrt1 (reviewed in [35]). The recruitment of various adaptor proteins, which share the F-box sequence motif, forms an array of distinct ubiquitin ligases with different substrate specificity. SCF complex was immunoaffinity isolated from human and yeast cells using the epitope-tagged cullin subunits *cul1* [36] and *cdc53* [37], respectively, as baits. Comparison of the patterns of coimmunoprecipitated proteins revealed orthologous core proteins, along with a pool of F-box adaptors (Fig. 4). However, eight subunits of the signalosome complex (CSN), a conserved 500 kDa protein assembly originally discovered in *Arabidopsis* [38], were found in association with *cul1* from human and not yeast. Subsequent experiments suggested a possible role of the CSN in regulating of ligase activity [36]. Interestingly, the budding yeast genome only encodes for the ortholog of a single subunit of CSN-CSN5, which is called Rri1. However neither Rri1, nor its interaction partners suggested by two-hybrid screening [39] or by systematic analysis of protein complexes-[33, 40], were detected in the immunoprecipitate of tagged *cdc53.* Thus critical insight into the regulation of the conserved ubiquitin ligase complex SCF by another conserved complex CSN came from the isolation and comparative

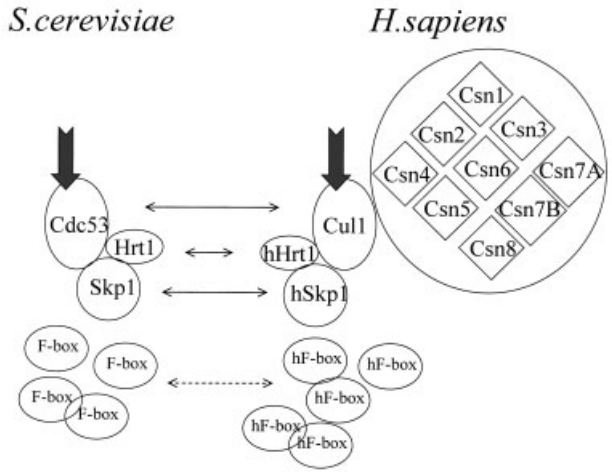

Figure 4. The orthologous SCF ubiquitin ligase complexes were purified from budding yeast and human cells by immunoaffinity chromatography (epitope tagged subunits are designated by filled arrows) and dissected by MS. Orthologous subunits are designated with pointed lines. Although complexes have similar molecular architecture, in the human cells *cull* subunits are additionally associated with the 500 kDa signalosome complex, which comprises nine subunits. No orthologs of CSN subunits were found to associate with cullin *cdc53* in the budding yeast [36, 37].

analysis of complexes in multiple species, rather than *via* expanding the pattern of interactors identified in a single model organism.

5 Developments in genomic sequencing and the study of proteomes by MS

The development of automated high-throughput DNA sequencing in the early 1990's laid grounds for genomic sequencing. The first living organism to be sequenced was *Haemophilus influenzae,* in 1995. Since the completion of the first genome, many unicellular and multicellular eukaryotic organisms' genomes have been sequenced, including *Sacharomyces cerevisiae, Escherichia coli, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana,* and the crowning achievement of the first draft assembly of the human genome [41, 42]. Genomic sequencing continues at a very high rate with the completion of a new organism every few months. Genomic sequences are annotated, translated into theoretical protein sequences, and many are made available at databases that are publicly accessible over the internet. Besides projects that endeavor to sequence the genomes of whole organisms, there has been a significant contribution to databases from the sequencing of individual genes and ESTs (mRNA) [43] from organisms whose genomes remain largely unsequenced. Both unannotated genomic sequence and EST databases can be searched with MS data [44].

With this first completed genome, biologists began to identify large sets of proteins from *H. influenzae* using 2-D gels and MS *via* peptide mass mapping [45]. With the completion of genomic sequencing of other organisms, research into the proteome of these organisms began to follow by a variety of MS techniques. These efforts have been extensive and more research has been accomplished than we can cite here.

The research community that uses MS for protein identification has made a habit of identifying proteins from only those organisms with sequenced genomes, because of the ability to easily translate those sequences and correlate them with analyzed proteins in a number of ways, as shown above. However, using MS and advanced methods of database interrogation, it is becoming increasingly possible to study the proteomes of organisms with unsequenced genomes. Cross-species identifications have been made in these organisms and others not cited: *Zea mays* [46, 47], *Pisum sativum* [48, 49], *Papaver somniferum* [50], *Spinacia oleracea* [44], *Arabidopsis thaliana* [44], *Bos taurus* [51], *Xenopus laevis* [30, 52], *Pichia pastoris* [25], and *Trypanosoma brucei* [28, 29]. Many earlier studies utilizing cross-species identification of unknown proteins relied on high mass accuracy MALDI peptide mapping, and therefore may have identified proteins or enzymes highly conserved across the biosphere. As more sequence similarity–based methods are being developed and applied, we envision that the proteomics of organisms with unsequenced genomes will become more productive and insightful by being able to identify a wider breadth of proteins, *i.e.* less conserved proteins in closely related organisms and conserved proteins in distantly related organisms.

6 Perspective: Proteomics and phylogenetic considerations for future genomic sequencing

Currently, there is a substantial debate over which organisms' genomes deserve to be sequenced next [53, 54]. A recent conference sponsored by the National Human Genome Research Institute focused on the direction of genomic sequencing and established criteria for the selection of the next organisms to be sequenced. The criteria include the ability to improve human health, the scientific utility of the new data, and technical considerations. One suggestion at the conference was to sequence the genome of an organism from each of the major branches of life to better understand the evolution of traits [53]. As already mentioned, the ability to identify proteins depends on database content, meaning this genomic sequencing proposal may also be extremely beneficial for the proteomics of the organisms with unsequenced genomes in these diverse branches, in light of cross-species protein identification by MS.

With 1.7 million known organisms, it is evident that the research community will not be able to sequence the genome of every organism. Many biological researchers investigating proteomes have already experienced the lack of genomic resources as an inability to identify proteins by MS. For example, proteomic studies in maize, an economically important organism, have been compromised due to the lack of database resources and an inability to use available database resources effectively [47]. However, plant scientists have begun to realize the limitations of non-error-tolerant methods of protein identification and now see the prospects of sequence similarity methods to contribute to proteomics [47, 55].

Yet for organisms distantly related to ones with sequenced genomes, even protein identification by sequence similarity methods will be ineffective in many cases because sequences still will not exist in databases that have significant identity to those proteins studied. For

example, whereas many proteins in mammals will have sequence similarity to humans, the more diverse classes of proteins in distantly related mammals would be unable to be identified (any protein below ~50% identity). In addition, as percent identity decreases between orthologs, it is likely that the divergent protein will take on a new function. Without the genomic sequencing of organisms in these distant phylogenetic regions, which could fill the gap between available genomic sequences and proteins from organisms with unsequenced genomes, many analyzed proteins will go unidentified because of a continuing deficiency of genomic sequence resources.

We understand that as genomic sequences become available to the public in the form of annotated database entries, these sequences are immediately used in proteomics to identify isolated gene products. Historically, with the completion of a genomic sequencing project, those sequences were utilized to identify proteins from the organism with the newly sequenced genome. Since MS relies upon databases to make protein identifications, it is evident that as more genomic sequences are produced, it will be possible to identify more proteins. This has been the case since the inception of proteomics.

However, in addition, every sequenced genome provides a resource that enables researchers to identify homologous proteins in many organisms. Consider the impact made by the complete genomic sequencing of *A. thaliana* upon plant proteomics (Table 1). The *Arabidopsis* genomic sequences provide a resource for the identification of proteins from *Arabidopsis* itself and many different species of plants as well. With the application of MS and sequence similarity methods, a single sequenced genome will enable the identification of more proteins from the proteomes of organisms with unsequenced genomes than by the use of methods that are only able to identify proteins of high homology to database entries. For example, all cross-species identifications in the study of the maize proteome by Chang *et al.* [46] could have been accomplished using database entries from the *Arabidopsis* genome and sequence similarity searches (Table 2).

Table 1. Sequencing of the *Arabidopsis* genome and its effects in proteomics

Year	One development in genomics	Development in proteomics	Ref. organisms in the ID of proteins	MS	MS/MS	SSS	Citiation
2000	*Arabidopsis thaliana*	*Papaver somniferum*	*ARABIDOPSIS (31), Pisum sativum (6), Glycine max (6), Nicotiana tabacum (3), Solanum tuberosum (3), Oryza sativa (3), Vitis vinifera (3), Protect neriifolia (2), Lavatera thuringiaca (2), Brassica oleracea (2), Fritillaria agrestis (1), Zea mays (1), Brassica juncea (1),* *Datisca glomerata* (1), *Hordeum vulgare* (1), *Saccharomyces cerevisiae* (1), *Spinacia oleracea* (1), *Linum usitatissimum* (1), *Citris paradisi* (1), *Catharanthus roseus* (1), *Schizosaccharomyces pombe* (1), *Batis maritima* (1), *Thermotoga maritima* (1), *Alcaligenes entrophus* (1), *Amycolatopsis mediteranei* (1), *Malus domestica* (1), *Mesmryanthemum crystallinum* (1)	X		X	[50]
		Zea mays	*ARABIDOPSIS* (2), *Beta vulsaris* (2), *Brassica napus* (2), *G. max* (2), *C. roseus* (4), *O. sativa* (3), *N. tabacum* (3), *Medicago sativa* (3), *H. vulgare* (2), *Chlamydomonas remhardtii* (1)	X			[46]
		Pisum sativum	*ARABIDOPSIS* (8). *Z. mays* (3), *G. max* (3), *O. sativa* (2), *Lycopersicum esculentum* (2), *Nicotiana sylvestris* (1), *H. vulgare* (1), *Carica papaya* (1), *Helianthus annuus* (1), *Onobrychis vicifolia* (1), *Sesbania rostrata* (1), *Physcomitrella patens* (1)	X	X		[48]
		A. thaliana	*ARABIDOPSIS* (33), *Z. mays (3)*	X	X		[44]

Table 1. Continued

Year	One development in genomics	Development in proteomics	Ref. organisms in the ID of proteins	MS	MS/MS	SSS	Citiation
		Zea mays	ARABIDOPSIS (14), *O. sativa* (22), *Triticum aestivum* (15), *P. sativum* (5), *H. vulgare* (5), *S. oleracea* (5), *Cucumis sativus* (4), *N. tabacum* (3), *S. tuberosum* (2), *Picea rubens* (1), *Secale cereale* (1), *Populus nigra* (1), *Schismocarpus matudai* (1)	X			[47]
2001		*Trypanosoma brucei*	ARABIDOPSIS, *Mus musculus, D. melanogaster, Entamoeba histolytica, Caenrohaditus elegans, O. sativa, S. cerevisiae, and others (9)*		X	X	[28]
		X. laevis	ARABIDOPSIS (1)[a], Bas taurus (1)[a], *M. musculus* (1)[a], *Ratus norwegicus* (1)[a], *Homo sapiens* (1)[a]		X	X	[30]

Organisms in the column "Developments in proteomics" had proteins identified by cross-species identification. Organisms in the next column contributed reference database entries used in the identification of proteins from organisms in the previous column, with the number of proteins following reference organism.
MS designates that proteins were identified peptide mass mapping, MS/MS by tandem mass spectrometry, and SSS by sequence similarity searches.
a) Multiple alignments are made to different species from the analysis of a single protein.
b) Peptides were sequenced by Edman degradation.

Table 2. Arabidopsis homologues to database references used in maize protein identification [46]

Accession No.		Accession No.		Identity
X89451	*B. napus*	AAL32658	*Arabidopsis*	89%
Z97178	*B. vulgaris*	AAK32918	*Arabidopsis*	89%
X83499	*C. roseus*	AAK82464	*Arabidopsis*	88%
U40212	*C. reinhardtii*	AAL32658	*Arabidopsis*	72%
U53418	*G. max*	BAB02581	*Arabidopsis*	89%
P37228	*G. max*	082399	*Arabidopsis*	77%
X91347	*H. vulgare*	P57751	*Arabidopsis*	77%
AF020271	*M. sativa*	BAA97065	*Arabidopsis*	80%
X77944	*N. tabacum*	AAK73989	*Arabidopsis*	88%
38199	O. sariva	T48154	*Arabidopsis*	83%
D67043	*O.sativa*	AAA79369	*Arabidopsis*	83%
Z26867	*O. sativa*	AAG10639	*Arabidopsis*	88%

Protein sequences from organisms in the left column were used to identify maize proteins by peptide mass mapping. *Arabidopsis* homologues exist to all maize proteins that were cross-species identified.

Even though Chang *et al.* identified maize proteins using database entries from many different plants, this data only underscores the fact that many proteins are highly homologous in related organisms, and sequence similarity searches will likely be successful in making proteomics a reality in a significant number of organisms with unsequenced genomes.

The expanding organismal scope of proteomics depends upon the creation of software tools for sequence similarity searching and related methods that couple MS with bioinformatics, as discussed above, and the sequencing of genomes. In this context, organisms representative of diverse phylogenetic lineages must have their genomes sequenced. More specifically, the proteomics of organisms with unsequenced genomes is probably focused to certain phylogenetic branches, which could be better represented by genomic sequencing, giving a broad resource for many independent researchers. These sequenced genomes can represent many phylogenetically related organisms depending on the nucleotide substitution rate in those lineages and the ability to annotate future genomic sequences [56, 57]. For example, research in maize and other economically important grasses, such as wheat and barley, will benefit from the complete sequence of the rice genome, as 98% of the known proteins from these grasses have homologs in rice [58].

With the complete sequencing of the pufferfish genome, we can predict that studies into the proteomes of other fishes will capitalize on these sequence resources by

using sequence similarity search methods [59]. Once a bird's or reptile's genome is sequenced, we can expect to see developments in the proteomics of related organisms. For the timely expansion of the organismal scope of proteomics, the selection of closely related organisms for genomic sequencing is not an optimized use of available resources. From a proteomics perspective, it makes no difference whether the human or the chimpanzee has its genome sequenced, because only one of the organisms needs to have its genome sequenced for the successful proteomics of both using the discussed analytical methods.

With the use of MS and emerging bioinformatic techniques, proteins could potentially be identified from any organism depending on the availability of diverse genomic sequences and the annotation of those sequences. As many biologists are without protein identification support for their research, we can directly conclude from these developments that where proteomics studies are desired, genomics should utilize its efforts on organisms phylogenetically situated to positively affect the proteomics of their phylogenetic neighbors. The future of protein identifications by mass spectrometry and the efforts of biological scientists involved in proteomics of organisms with unsequenced genomes depends to a large degree on the sequencing of the genomes from underrepresented classes and distantly related organisms in accordance with the findings of molecular systematics.

We are grateful to other members of AS group for their support and Drs. M. Brand, B. Habermann, and A. Hyman of the Max Planck Institute of Molecular Cell Biology and Genetics for critically reading the manuscript and useful discussions.

Received May 17, 2002

7 References

[1] Mann, M., Hendrickson R. C., Pandey A., *Annu. Rev. Biochem.* 2001, *70*, 437–473.

[2] Yates, J. R. III, *Trends Genet.* 2000, *16*, 5–8.

[3] Griffin, T. J., Aebersold, R., *J. Biol. Chem.* 2001, *276*, 45497–45500.

[4] Rubin, G. M., Yandell, M. D., Wortman, J. E., Gabor Milkos, G. L. *et al.*, *Science* 2000, *287*, 2204–2215.

[5] Jensen, O. N., Wilm, M., Shevchenko, A., Mann, M., *Methods Mol. Biol.* 1999, *112*, 513–530.

[6] Shevchenko, A., Chernushevich, I., Wilm, M., Mann, M., *Methods Mol. Biol.* 2000, *146*, 1–16.

[7] Yates, J. R. III, Link, A. J., Schieltz, D., *Methods Mol. Biol.* 2000, *146*, 17–26.

[8] Chalmers, M. J., Gaskell, S. J., *Curr. Opin. Biotechnol.* 2000, *11*, 384–390.

[9] Wilkins, M. R., Williams, K. L., *J. Theor. Biol.* 1997, *186*, 7–15.

[10] Clauser, K. R., Baker P., Burlingame, A. L., *Anal. Chem.* 1999, *71*, 2871–2882.

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

[12] Chernushevich, I., Loboda, A., Thomson, B., *J. Mass Spectrom.* 2001, *36*, 849–865.

[13] Medzihradszky, K. F., Campbell, J. M., Baldwin, M. A., Falick, A. M. *et al.*, *Anal. Chem.* 2000, *72*, 552–558.

[14] Washbum, M. P., Wolters D., Yates, J. R. III, *Nat. Biotechnol.* 2001, *19*, 242–247.

[15] Biemann, K., *Biomed. Environ. Mass Spectrom.* 1988, *16*, 99–111.

[16] Bartet-Jones, M., Jeffery, W. A., Hansen, H. F., Pappin, D. J. C., *Rapid Commun. Mass Spectrom.* 1994, *8*, 737–742.

[17] Keough, T., Youngquist, R. S., Lacey, M. P., *Proc. Natl. Acad. Sci. USA* 1999, *96*, 7131–7136.

[18] Goodlett, D. R., Keller, A., Watts, J. D., Newitt, R. *et al.*, *Rapid Commun. Mass Spectrum.* 2001, *15*, 1214–1221.

[19] Shevchenko, A., Chernushevich, I., Ens, W., Standing, K. G. *et al.*, *Rapid Commun. Mass Spectrom.* 1997, *11*, 1015–1024.

[20] Uttenweiler-Joseph, S., Neubauer, G., Christoforidis, S., Zerial, M. *et al.*, *Proteomics* 2001, *1*, 668–682.

[21] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E. *et al.*, *J. Comput. Biol.* 1999, *6*, 327–342.

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

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

[24] Pevzner, P. A., Mulyukov, Z., Dancik, V., Tang, C. L., *Genome Res.* 2001, *11*, 290–299.

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

[26] Altschul, S., Madden, T., Schaffer, A., Zhang, J. *et al.*, *Nucleic Acids Res.* 1997, *25*, 3389–3402.

[27] Pearson, W. R., *Methods Mol. Biol.* 2000, *132*, 185–219.

[28] Huang, L., Jacob, R. S., Pegg, S. C., Baldwin, M. A. *et al.*, *J. Biol. Chem.* 2001, *276,* 28327–28339.

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

[30] Shevchenko, A., Sunyaev, S., Liska, A., Bork, P. *et al.*, Methods Mol. Biol. 2002, *211*, 221–234.

[31] Pandey, A., Mann, M., *Nature* 2000, *405*, 837–846.

[32] Alberts, B., *Cell* 1998, *92*, 291–294.

[33] Gavin, A. C., Bosche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[34] Ibba, M., Soll, D., *Annu. Rev. Biochem.* 2000, *69,* 617–650.

[35] Deshaies, R. J., *Annu. Rev. Cell Dev. Biol.* 1999, *15*, 435–467.

[36] Lyapina, S., Cope, G., Shevchenko, A., Serino, G. *et al.*, *Science* 2001, *292*, 1382–1385.

[37] Seol, J. H., Shevchenko, A., Shevchenko A., Deshaies, R. J., *Nat. Cell Biol.* 2001, *3,* 384–391.

[38] Wei, N., Deng, X., *Trends Genet.* 1999, *15*, 98–103.

[39] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, *Nature* 2000, *403,* 623–627.

[40] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[41] Venter, J. C., *Science* 2001, *291*, 1304–1351.

[42] International Human Genome Sequencing Consortium, *Nature* 2001, *409,* 860–921.

[43] Boguski, M. S., Lowe T. M. J., Tolstoshev, C. M., *Nat. Genet*. 1993, *4*, 3327–333.

[44] Kuster, B., Mortensen, P., Andersen J. S., Mann, M., Proteomics 2001, *1*, 641 –50.

[45] Langen, H., Gray, C., Roder, D., Juranville, J. F. *et al.*, *Electrophoresis* 1997, *18*, 1184–1192.

Dr. Andrej Shevchenko received his PhD from the Leningrad Institute of Technology in 1991. From 1994–2000 he worked in the Peptide and Protein Group in the European Molecular Biology Laboratory (EMBL) in Heidelberg. Since 2001 he has been a Group Leader in the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden. His fields of research interest are: functional proteomics, bioinformatics, protein complexes and protein interaction networks.

[46] Chang, W. W., Huang, L., Shen, M., Webster, C. *et al.*, *Plant Physiol.* 2000, *122*, 295–318.

[47] Porubleva, L., Velden, K. V., Kothari, S., Oliver, D. J. *et al.*, *Electrophoresis* 2001, *22*, 1724–1738.

[48] Peltier, J. B., Friso, G., Kalume, D. E., Roepstorff, P. *et al.*, *Plant Cell* 2000, *12*, 319–341.

[49] Gomez, S. M., Nishio, J. N., Faull, K. F., Whitelegge, J. P., *Mol. Cell. Proteomics* 2002, *1*, 46–59.

[50] Decker, G., Wanner, G., Zenk, M. H., Lottspeich, F., *Electrophoresis* 2000, *21*, 3500–3516.

[51] Koc, E. C., Burkhart, W., Blackburn, K., Moseley, A. *et al.*, *J. Biol. Chem.* 2000, *275*, 32585–32591.

[52] Tournebize, R., Popov, A., Kinoshita, K., Ashford, A. J. *et al.*, *Nat. Cell Biol.* 2000, *2*, 13–19.

[53] Gewolb, J., *Science* 2001, *293*, 409–410.

[54] Gewolb, J., *Science* 2001, *293*, 585–586.

[55] van Wijk, K. I., *Plant Physiol.* 2001, *126*, 501–508.

[56] Muse, S. V., *Plant Mol. Biol.* 2000, *42*, 25–43.

[57] Iliopoulos, I., Tsoka, S., Andrade, M. A., Janssen, P. *et al.*, *Genome Biol.* 2001, *2*, INTERACTIONS001.

[58] Goff, S. A., Ricke, D., Lan, T. H., Presting, G. *et al.*, *Science* 2002, *296*, 92–100.

[59] Aparicio, S., Chapman, J., Stupka, E., Putnam, N. *et al.*, *Science* 2002, *297*, 1301–1310.