Homology-Based Functional Proteomics By Mass Spectrometry and Advanced Informatic Methods

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium

(Dr. rer. nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften

der Technische Universität Dresden

von

B.S. Biochem. Biol. Adam J. Liska

geboren am 11. November 1976 in Lincoln, Nebraska, USA

Gutachter: Prof. Dr. Michael Göttfert, Technische Universität Dresden

Dr. Andrej Shevchenko, MPI of Mol. Cell Biology and Genetics

Prof. Dr. Michael O. Glocker, Universität Rostock

Eingereicht am: Oktober 2003

Tag der Verteidigung:

“all science is philosophy, whether it knows and wills it or not.”

Martin Heidegger

Acknowledgements

I would like to thank the exceptional staff of the *Max Planck Institute of Molecular Cell Biology and Genetics* in Dresden for addressing some of my domestic concerns as well as providing a great library resource. I thank my wife for daily supporting me through this research. I would also like to thank my collaborators for insight, suggestions, criticisms, and attention to detail, as well as those who provided technical assistance through the course of experiments: Professor Shamil Sunyaev (*Genetics Division*, Harvard Medical School, USA) for mathematics and bioinformatics collaborations concerning MS BLAST and MultiTag; Dr. Andrei Popov and Prof. Eric Karsenti (*Cell Biology*, EMBL, Heidelberg, Germany) for collaborations concerning *Xenopus laevis* cell biology; Prof. Peer Bork (*Computational Biology* EMBL, Heidelberg, Germany) for assistance in manuscript preparation and bioinformatics; Dr. Alexander Golod (Russia) for programming the MultiTag software; Dr. Adriana Katz and Prof. Uri Pick (*Biological Chemistry*, The Weizmann Institute of Science, Israel) for collaborations concerning *Dunaliella salina* biochemistry; Drs. Ignat N. Shilov, Subodh Nimkar, and Dan A. Schaeffer, (Applied Biosystems, Foster City, USA) for developing advanced mass spectrometry software to support our DB searching methods; Dr. Igor Chernushevich (MDS Sciex, Toronto, Canada) for assistance concerning quadrupole time-of-flight mass spectrometry; Dr. Peg Coughlin and Prof. Timothy Mitchison for *Xenopus laevis* spindle electron microscopy (*Department of Cell Biology*, Harvard Medical School, USA); Prof. Michael Brand and Prof. Michael Göttfert (Technische Universität Dresden, Germany) for being helpful members of my Thesis Advisory Committee; Dr. Bianca Habermann (*Bioinformatics*, MPI-CBG) for collaborations concerning MS BLAST database searching and phylogenetics; Dr. Jan Havlis (currently at the *Department of Analytical Chemistry*, Masaryk University, Czech Republic), Dr. Henrik Thomas, and Anna Shevchenko for assistance with mass spectrometry (MPI-CBG); and Dr. Andrej Shevchenko for mentoring me through my dissertation research.

Summary

Functional characterization of biochemically-isolated proteins is a central task in the biochemical and genetic description of the biology of cells and tissues. Protein identification by mass spectrometry consists of associating an isolated protein with a specific gene or protein sequence *in silico*, thus inferring its specific biochemical function based upon previous characterizations of that protein or a similar protein having that sequence identity. By performing this analysis on a large scale in conjunction with biochemical experiments, novel biological knowledge can be developed. The study presented here focuses on mass spectrometry-based proteomics of organisms with unsequenced genomes and corresponding developments in biological sequence database searching with mass spectrometry data. Conventional methods to identify proteins by mass spectrometry analysis have employed proteolytic digestion, fragmentation of resultant peptides, and the correlation of acquired tandem mass spectra with database sequences, relying upon exact matching algorithms; i.e. the analyzed peptide had to previously exist in a database *in silico* to be identified. One existing sequence-similarity protein identification method was applied (MS BLAST, Shevchenko 2001) and one alternative novel method was developed (MultiTag), for searching protein and EST databases, to enable the recognition of proteins that are generally unrecognizable by conventional softwares but share significant sequence similarity with database entries (~60-90%). These techniques and available database sequences enabled the characterization of the *Xenopus laevis* microtubule-associated proteome and the *Dunaliella salina* soluble salt-induced proteome, both organisms with unsequenced genomes and minimal database sequence resources. These sequence-similarity methods extended protein identification capabilities by more than two-fold compared to conventional methods, making existing methods virtually superfluous. The proteomics of *Dunaliella salina* demonstrated the utility of MS BLAST as an indispensable method for characterization of proteins in organisms with unsequenced genomes, and produced insight into *Dunaliella*'s inherent resilience to high salinity. The *Xenopus* study was the first proteomics project to simultaneously use all three central methods of representation for peptide tandem mass spectra for protein identification: sequence tags, amino acids sequences, and mass lists; and it is the largest proteomics

study in *Xenopus laevis* yet completed, which indicated a potential relationship between the mitotic spindle of dividing cells and the protein synthesis machinery. At the beginning of these experiments, the identification of proteins was conceptualized as using “conventional” versus “sequence-similarity” techniques, but through the course of experiments, a conceptual shift in understanding occurred along with the techniques developed and employed to encompass variations in mass spectrometry instrumentation, alternative mass spectrum representation forms, and the complexities of database resources, producing a more systematic description and utilization of available resources for the characterization of proteomes by mass spectrometry and advanced informatic approaches. The experiments demonstrated that proteomics technologies are only as powerful in the field of biology as the biochemical experiments are precise and meaningful.

Table of Contents

Editorial Note

The introduction of this dissertation (sections **1**) contain material previously published in the review articles "Expanding the Organismal Scope of Proteomics: Cross-Species Protein Identification by Mass Spectrometry and its Implications" (Liska A.J. and A. Shevchenko, *Proteomics* 3, 19-28, 2003) and "Combining Mass Spectrometry with Database Interrogation Strategies in Proteomics" (Liska A.J. and A. Shevchenko, *Trends in Analytical Chemistry* 22, 291-298, 2003). Section **2.1**, "Development of the MultiTag Sequence-Similarity Protein Identification Method", contains material, previously published and submitted for publication, in the articles "MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry" (Sunyaev S., A.J. Liska, A. Golod, A. Shevchenko, and A. Shevchenko, *Analytical Chemistry*, 75, 1307-1315, 2003) and "Enhanced Error-Tolerant EST Database Searching by Tandem Mass Spectrometry and MultiTag Software" (Liska A.J., S. Sunyaev, I.N. Shilov, D.A. Schaeffer, and A. Shevchenko, *Analytical Chemistry*, submitted), respectively. Section **2.2**, "*Xenopus laevis* Functional Proteomics", contains material, previously published and submitted for publication, in the articles "Nanoelectrospray Tandem Mass Spectrometry and Sequence Similarity Searching for Identification of Proteins from Organisms with Unknown Genomes" (Shevchenko A., S. Sunyaev, A.J. Liska, P. Bork, and A. Shevchenko, *Methods in Molecular Biology,* vol. 211, 221-234, 2003) and "Homology-Based Functional Proteomics by Mass Spectrometry: Application to the *Xenopus* Microtubule-Associated Proteome" (Liska A.J., A.V. Popov, S. Sunyaev, P. Coughlin, B. Habermann, A. Shevchenko, P. Bork, E. Karsenti, and A. Shevchenko, *Molecular and Cellular Proteomics*, submitted), respectively. Section **2.3**, "*Dunaliella salina* Functional Proteomics", contains material submitted for publication in the article, "Homology-Based Proteomics By Mass Spectrometry Reveals Aspects of Salinity Adaptation in Dunaliella " (Liska A.J., A. Katz, A. Shevchenko, and U. Pick, *Plant Physiology*, submitted). Section **2.4** contains material previously published and submitted for publication in "Expanding the Organismal Scope…" and "Homology-Based Functional Proteomics…", respectively. Section **2.5** contains material previously published in "Combining Mass Spectrometry with Database Interrogation..."

Index of Figures

Index of Tables

Abbreviations

2-D, two-dimensional
ARS, aminoacyl-tRNA synthetase
BLAST, basic local alignment search tool
DNA, deoxyribose nucleic acid
cDNA, complementary DNA
CID, collision-induced dissociation
DB, database
E value, expectation value
EF, elongation factor
ESI, electrospray ionization
EST, expressed sequence tag
FTMS, Fourier transform ion-cyclotron resonance mass spectrometry
FWHM, full width at half maximum
HSP, heat shock protein
LC, liquid chromatography
MALDI, matrix-assisted laser desorption/ionization
MAP, microtubule-associated protein
MS, mass spectrometry
MS BLAST, mass spectrometry driven BLAST database searching
MS/MS, tandem mass spectrometry
MSP, mass spectrometry platform
MT, MultiTag
nanoESI, nanoelectrospray
nanoLC, nanoflow liquid chromatography
PAGE, polyacrylamide gel electrophoresis
PCR, polymerase chain reaction
PMF, peptide mass fingerprinting
PSD, post-source decay
Q(q)TOF, quadrupole time-of-flight
RNA, ribose nucleic acid
TOF, time-of-flight
TQ, triple quadruple

1 Introduction

1.1 Mass Spectrometry-Based Proteomics and Biological Sequence Database Searching

1.1.1 Technological Developments in Biology and the Emergence of Proteomics

A number of technologies are revolutionizing biological research. Molecular biology has mastered the detection and manipulation of genes by specific nucleases, PCR, Northern & Southern blot, microarray technology, RNA interference, and cell transformation by homologous recombination, among other techniques. At the same time, high-throughput DNA sequencing has paved the way for 'shotgun' whole genome sequencing. Developments in protein biochemistry now allow proteins to be isolated specifically from cells along with selective isolation of interacting protein partners. Technical and computational advances now enable mass spectrometry (MS) to ionize proteins and peptides into the gas phase with high yields, and determine their masses with high accuracy, creating a direct tie between analyzed protein fragments and database (DB) sequences (genes) in a matter of minutes. This series of tools allows unprecedented levels of molecular analysis of proteins and genes in living cells. Now proteins and the genes that give cells their unique properties can be examined rapidly and accurately, thus advancing the development of theoretical biological knowledge, applied biotechnology and the biomedical sciences to a high degree. Developments in the related fields above have enabled proteomics to arise in the biological sciences.

Proteomics is the characterization of groups of proteins that are found in specific cells or tissues[1]; the proteome is defined as the protein complement of the genome; the genome being the cells complete set of DNA. Proteomics research is carried out in order to characterize cellular protein complexes or organelles in cell biology, analyze gene expression patterns[2], and interrogate genes and genomes. The first problem in conducting proteomics is the development of accurate and versatile protein identification strategies. Although it has been possible to purify proteins using established methods of biochemistry, the crucial step in the characterization of any proteome is high-throughput protein identification. Protein identification by MS consists of associating a biochemically-isolated protein with a specific gene or protein sequence *in silico*, thus inferring its specific biochemical function based upon previous characterizations of that protein or a similar protein having that sequence identity. By performing this on a large scale in conjunction with biochemical experiments, novel biological knowledge can be developed.

In proteomics, MS has become a powerful analytical technology to identify proteins by the analysis of peptides and the correlation of resultant mass spectra with available DB sequences (reviewed in[3,4]). Genomic sequencing projects, which supply the majority 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 significant the development of these efforts may be[5]. However recently many important organisms have had their genomes sequenced, such as human[6,7], mouse[8], rice[9,10], Arabidopsis[11], and the pufferfish[12]. Using established MS techniques, a limited DB resource has not been conducive for facile protein identification from species with unsequenced genomes. Yet despite the relative deficiency of genomic sequences compared to a whole biosphere of living species, the emerging interplay of MS and bioinformatics is significantly expanding the organismal scope of proteomics.

1.1.2 Cross-Species Protein Identification by Mass Spectrometry

Irrespective of whether the genome of a species is sequenced or not, the identification of proteins by MS consists primarily of two analyses of peptides produced by proteolytic digestion of purified whole proteins. Matrix-assisted laser desorption/ionization time-of-flight (MALDI TOF) MS produces spectra by the resolution of intact peptides according to their masses, and identifies proteins by the correlation of these masses with theoretically calculated masses of peptides from DB entries; a method defined as peptide mass fingerprinting (PMF). The second type of analysis, tandem mass spectrometry (MS/MS), produces patterns of peptide fragments that can be correlated to DB entries in a number of ways (see section **2.5.2**).

Using MS and available protein DB sequences, cross-species protein identifications are accomplished by partially aligning analyzed peptides from a protein from a species with an unsequenced genome to a DB sequence from a related species. After DNA sequencing projects began, it became apparent that phylogenetically related species have significant genomic sequence co-linearity and their proteins have a high degree of homology[5]. 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 percentage of identity.

PMF enables cross-species protein identification in some cases because only a subset of all peptides from a protein digest needs to be recognized[13,14]. Those peptides that have amino acid substitutions and corresponding shifts in mass are not recognized and don't contribute to the identification. The theoretical predictions by Wilkins and Williams proposed that proteins can be identified using PMF if the analyzed protein and reference DB entry have >80% sequence identity, although the authors added that these cases would need to be supported by further evidence for validation. The high mass accuracy of modern TOF and Fourier transform ion-cyclotron resonance mass spectrometry (FTMS) instruments increases confidence in cross-species PMF and loosens the sequence identity requirement, as less peptide masses would be required to produce a confident hit[15].

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. In this method, masses of precursor ions and fragment ions (from MS/MS) are submitted for DB searching using specialized software (reviewed in[16]). Regardless of differences in DB searching algorithms and MS platforms, the conventional softwares correlate the observed masses with theoretically predicted masses derived from peptide sequences produced by *in silico* digestion of protein DB entries, and calculates the statistical significance of matches. Importantly, these softwares do not require a full representation of the fragment ions in the tandem mass spectrum and can positively identify the peptide even if only some of the fragment ions are matched. The significance of hits increases if more fragment ions are detected and if more than one peptide sequence originating from the same DB entry is recognized. Thus conventional DB mining software is inherently biased towards exact matching of spectra (and corresponding peptides) to catalogued sequences, and in practice it is mostly applied to the identification of proteins already residing in available databases. It is therefore not surprising that proteomics is largely limited to organisms with sequenced genomes, despite the fact that phylogenetically related species share significant molecular homology and that extensive protein sequence information may be available from related species.

Where PMF and non-error-tolerant MS/MS methods fail, the identification of proteins in the past has relied 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 (Figure 1).

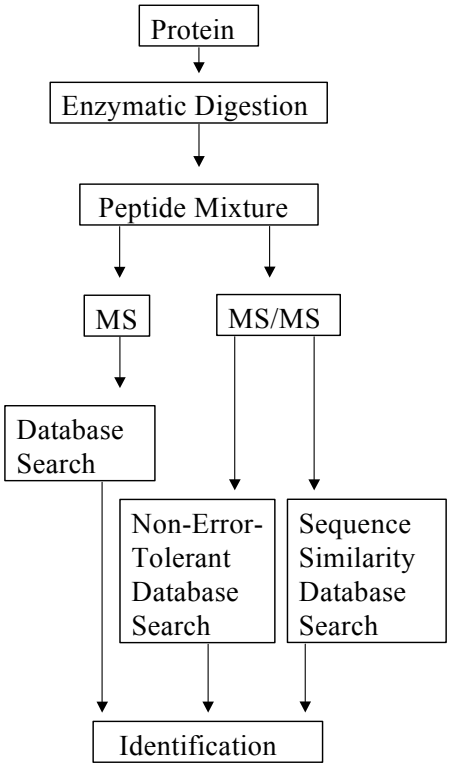

Figure 1 Strategy for cross-species protein identifications by mass spectrometry.

Proteins are identified by the analysis of peptides by either MS or MS/MS. A DB 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.

1.1.3 Mass Spectrometry Platforms

Historically, the first mass spectrometry platform (MSP) for protein identification included 2-D PAGE and MALDI TOF MS for PMF (reviewed in[17]). MALDI-TOF is most commonly used in proteomics in ion reflection mode because of its low femtomole (even attomole) sensitivity and high resolution (>10,000 full width at half maximum, FWHM). Peptide mass fingerprints are routinely acquired with better than 50 ppm mass accuracy with external calibration, and recently reported automated re-calibration methods lower the error of mass measurement below 10 ppm[18,19]. MALDI spectra can be acquired very rapidly, and the entire routine, starting from digestion of proteins, preparation of the

MALDI probes and acquiring the spectra, is now automated and optimized for a very high throughput[20,21].

MALDI identification capabilities can be additionally strengthened by the acquisition of post source decay (PSD) spectra for a few selected peptide ions[22]. However, acquisition of PSD spectra is rather slow and it is much less sensitive compared to peptide fingerprinting, fragmentation is poorly controlled, and the spectra suffer from low resolution and mass accuracy, and therefore the technology has not had a significant impact in proteomics. However the recently introduced LIFT method[23] speeds up acquisition of MALDI PSD spectra considerably.

MALDI MS/MS capabilities are mainly explored through the development of instruments with combined mass analyzers, such as MALDI-quadrupole TOF (Q(q)TOF) mass spectrometers[24,25] and MALDI-TOF/TOF[26]. Both instruments can acquire high mass accuracy peptide fingerprints, and enable the control of collision energy in MS/MS mode. However, because of the orthogonal configuration of the ion path, MALDI-Q(q)TOF machines can acquire MS/MS spectra at relatively low collision energy, with the same resolution (>12,000) and mass accuracy (<20 ppm) as in the MS mode[27].

Regardless of the employed mass analyzers, MALDI sources predominantly ionize tryptic peptides as singly charged ions. To fragment singly charged ions, higher collision energy is required and therefore cleavage of amide bonds in the peptide backbone occurs less consistently. Usually MALDI MS/MS spectra do not contain continuous ion series that facilitate the confident determination of long peptide sequences. A number of peptide derivatization methods, localizing the charged groups at the N- or C-terminus of the molecule, have been developed to improve peptide fragment patterns[28,29]. MALDI sources have also been coupled with an ion trap, allowing the acquisition of MS/MS spectra very rapidly, albeit mass accuracy of the ion trap is much lower compared to TOF analyzers[30]. However, a large number of acquired MS/MS spectra increases the specificity of DB searching, and compensates the lack of specificity of DB searching with peptide mass fingerprints, which is heavily dependant on mass accuracy.

Electrospray ionization (ESI) methods form another major cluster of MS platforms. In ESI MS, tryptic peptides are typically ionized as doubly or triply charged ions. Multiply charged ions can be efficiently fragmented at lower collision energy, and their MS/MS spectra are usually dominated by intense y- and b-ions (see[31] for the nomenclature), which facilitates DB searching and also makes the spectra more amenable for *de novo* interpretation[32].

Peptides can be either separated by liquid chromatography on-line with the mass spectrometer (LC-MS/MS), or alternatively, unseparated peptide mixtures may be directly analyzed by nanoelectrospray mass spectrometry (NanoESI)[33]. The absence of separation in NanoESI is compensated by much longer spraying time per spectrum, which is made possible by low flow injection rates (20 – 30 μL/ min), and many precursor ions can be fragmented successively[34]. However, NanoESI-MS/MS experiments are limited by the ability of the operator to recognize low abundance precursors masked by chemical noise. The specificity of precursor ion detection can be increased *via* precursor ion scanning for abundant immonium ions of amino acid residues (typically, of Leu and Ile)[35]. Precursor ion scanning is a routine operation mode of triple quadrupole (TQ) mass spectrometers and recently it has been set up also on Q(q)TOF machines[36,37]. Although fewer peptide precursor ions are typically fragmented in the course of NanoESI-MS/MS analysis, the quality and signal-to-noise ratio in their fragment spectra are routinely better than by LC-MS/MS, because the data accumulation time and collision energy can be precisely tuned by an operator during the acquisition process. However, NanoESI-MS/MS is relatively difficult to automate[38] and it has a limited ability to identify proteins in complex mixtures.

By pre-separating peptides in front of the on-line mass spectrometer, analytical methods gain higher dynamic range and ability to identify proteins in very complex mixtures[39]. By applying this method a substantial part of the proteomes of prokaryotic and low eukaryotic organisms can be characterized[39-41]. Further coupling of multidimensional LC-MS/MS analysis enables relative quantification that utilizes peptides enriched with stable isotopes as internal standards, and promises global survey of quantitative changes in the proteomes[41-43]. However, there is less control in the process of spectra acquisition, and information content of MS/MS spectra might be compromised. This might not be particularly important for protein identification by pattern searches (see later section) because with high resolution of instruments the ion statistics in the peak does not strongly affect mass accuracy, and full representation of fragments in the spectrum is not required[27]. However, poor ion statistics affects the accuracy of *de novo* sequencing, which benefits from recognizing complementary pairs of fragment ions and the full representation of low molecular weight peaks is often critical.

Because of differences in ionization mechanisms, MALDI and ESI produce different data sets when the same protein digest is analyzed[44,45]. Parallel analysis of digests by two methods increases the sequence coverage of peptide maps, but usually

requires the employment of different instrumentation. Rapidly switchable combined MALDI/ESI sources[46,47] allow changing between ionization modes within minutes without venting the mass spectrometer, and might provide an effective alternative to expanding costly instrumentation.

Other MSPs such as MALDI- and ESI-Fourier transform ion cyclotron mass spectrometry (FTMS)[48,49], linear ion trap[50], and ion trap-TOF[51] mass spectrometers are employed in proteomics, but are currently less prevalent. Depending on the type of mass spectra, specific types of databases (protein, EST, or genomic) can be interrogated with more or less efficiency.

1.1.4 Efforts Towards the Identification of Proteins by Tandem Mass Spectrometry and Sequence-Similarity Searches

The new instrument configurations described above have greatly contributed to large-scale protein identification. However, the increasing analytical precision and sensitivity of mass spectrometers does not necessarily lead to improved success in the identification of proteins from species 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 (Figure 2). Usually these spectra can only be partially interpreted due to the natural under-representation of peptide fragment ions and because of the presence of chemical noise, which may obscure peptide fragments of low intensity and misguide spectrum interpretation. To overcome this difficulty, methods for the chemical derivatization of peptides, and alternate methods of interpreting spectra and DB searching have been developed.

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 (ions containing N-terminus) or y-series (ions containing C-terminus), which are more common in tryptic peptides ionized by electrospray, resulting in the prediction of an amino acid sequence (see[31] for the nomenclature). Upon collision-induced dissociation (CID), tryptic peptides tend to break at the amide bonds between consecutive amino acid residues producing a continuous y-ion series of fragments; an amino acid sequence can be determined by measuring the mass difference between consecutive y-ions; this relies upon the varying mass values of different amino acid residues. 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[28,29]. 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[52]) or by CD_3OH[53]. ^{18}O-labeled y-ions can be recognized by a one or two Thomson shift (depending on the peptides fragments charge) and allow confident readout of a peptide sequence[54]. These methods have enabled the cloning of a few proteins via oligonucleotide primers and PCR[55]. 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.

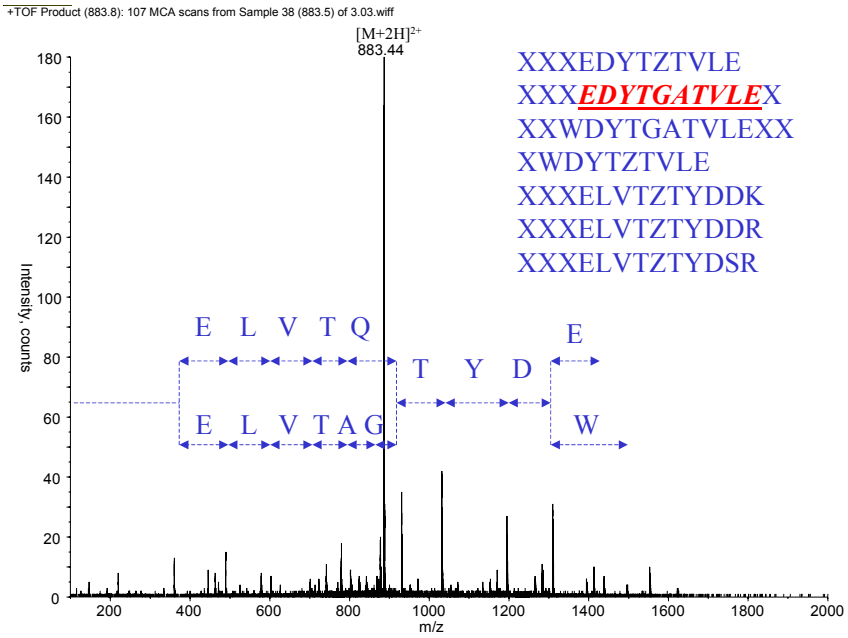

Figure 2 Interpretation of a peptide tandem mass spectrum.

The MS/MS spectrum of a doubly charged precursor ion with *m/z* 883.44 was acquired by fragmentation on a quadrupole time-of-flight 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 (delta 0.036 Daltons). The symbol X represents an unknown amino acid.

A second possibility is to interpret tandem mass spectra of peptides using specialized software that creates amino acid sequences *de novo*[56,57]. 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 according to their scored 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 DB using various sequence-similarity search algorithms.

The use of BLAST[58] or FASTA[59] DB searching engines to analyze peptide sequences produced by interpretation of tandem mass spectra is not straightforward because both algorithms have been optimized for comparing long and accurate protein sequences, whereas the interpretation of tandem mass spectra yields sets of short inherently-redundant and error-prone sequence candidates. Furthermore, it is not known in what order the peptide sequences should be aligned on the backbone of the polypeptide, if those sequences belong to a single protein or originate from a few proteins co-migrating within a single chromatographic gel band (or spot), nor what isobaric amino acids are present (such as, Leu and Ile, Lys and Gln, or Phe, or Met–sulphoxide).

To address these difficulties, common DB search engines have been manipulated to allow the input of sequences produced by MS. Modified FASTA-based software is available as stand-alone applications[57,60,61], whereas MS BLAST (Mass Spectrometry driven BLAST DB searching[62]) is accessible over the internet (http://dove.embl-heidelberg.de/Blast2/msblast.html). The limitations of FASTA-based algorithms are 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 candidate peptide sequences submitted in a query (the significance of all DB searches decreases with the increasing size of the query; i.e. the number of fragmented peptides). 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, as well as the difficulty to create one sequence prediction by manual interpretation. If DB searching with the predicted sequences makes no alignment, researchers are unable to ascertain whether the spectra were misinterpreted or no corresponding sequence exists in a DB. However, FASTA-based engines are flexible, may engage optional gapped alignment, and the statistical apparatus is specifically tailored for matching short peptide sequences.

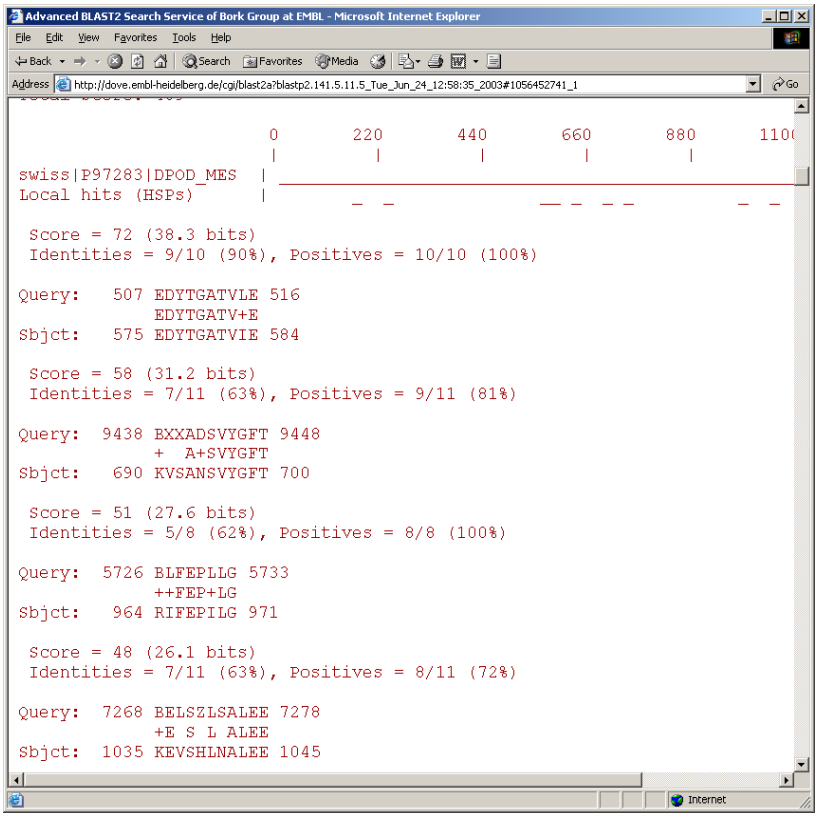

Figure 3 MS BLAST sequence alignment of an analyzed unknown protein from *Xenopus laevis*.

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; matching 10 peptides. Multiple hits from different organisms were retrieved from the MS BLAST search and many were able to make high confidence matches (see Table 4.)

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 (Figure 3)[62]. 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. This quality allows MS BLAST to be coupled with high-throughput sequencing techniques such as MALDI-TOF/TOF, MALDI-QqTOF and LC/MS/MS through a simple scripting interface[63]. Importantly, both MS BLAST and FASTS methods provide independent means of evaluating the statistical significance of alignments, and therefore it is not necessary to compare retrospectively the matched peptide sequences with actual tandem mass spectra to rule out false positive hits.

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 lower

homology to DB sequences can be identified with the limit being around 50% identity (this is dependent on which software is used)[61]. Since mass spectrometric analysis rarely reconstructs a complete sequence, the percent identity of the analyzed protein compared to the matching DB sequence is unable to be determined without extended analysis.

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 DB 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 because of the decrease in the number of homologous genes *in silico*.

1.1.5 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[64,65]. As protein complexes are often regarded as functional units of the molecular machinery of the cell[66], 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[5]. Comparison of three native protein complexes, isolated from budding yeast cells and from human cells by immunoaffinity chromatography, supports this notion[64]. Thus, it is conceivable that conserved protein complexes may be initially characterized in a model organism and then the obtained knowledge can be projected on orthologous complexes in other organisms, including humans. There are several lines of evidence why such a strategy will benefit from wider representation of model organisms, which might have uncharacterized 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 species 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 species 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 species, 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 non-aminoacyl-tRNA synthetase component Arc1p[67]. 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 non-aminoacyl-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[67]. Thus, characterization of the complex only in lower organisms or only in higher organisms provides limited knowledge of its architecture and function in general.

Most importantly, even if orthologous complexes are very similar in composition, they might be regulated via interaction with different, non-orthologous proteins or protein complexes. Orthologous cell cycle regulating ubiquitin ligases in yeast and human serve as a good example. The SCF complex (termed for Skp1–Cdc53–F-box protein) is built from conserved core subunits: Skp1, cullin homologue Cdc53, and RING H2 subunit Hrt1 (reviewed in[68]). 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*[69] and *cdc53*[70], respectively, as baits. Comparison of the patterns of co-immunoprecipitated proteins revealed orthologous core proteins, along with a pool of F-box adaptors. However, eight subunits of the signalosome complex (CSN), a conserved 500 kDa protein assembly originally discovered in *Arabidopsis*[71], 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[69]. Interestingly, the budding yeast genome only encodes for the apparent ortholog of a single subunit of CSN-CSN5, which is called Rri1[72]. However neither Rri1, nor its interaction partners

suggested by two-hybrid screening[73] or by systematic analysis of protein complexes[64,65], 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 analysis of complexes in multiple species, rather than via expanding the pattern of interactors identified in a single model organism.

1.1.6 Developments in Genomic Sequencing, Biological Sequence Databases, and Proteomics by Mass Spectrometry

The development of automated high-throughput DNA sequencing in the early 1990's made necessary technical advances 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 *S. cerevisiae*, *E. coli*, *C. elegans*, *D. melanogaster*, *A. thaliana*, and the crowning achievement of the first draft assembly of the human genome[6,7]. Genomic sequencing continues at a very high rate with the completion of a new organism every few months, if not weeks.

Protein sequence databases are continually updated with submissions produced from the cloning of genes, from which amino acid sequences are generated by translation of nucleotide sequences in their correct reading frames (www.ncbi.nlm.nih.gov, www.expasy.org/sprot/, www.ebi.ac.uk/). Whole-genome shotgun sequencing also produces large sets of nucleotide sequences that are assembled into contiguous sequences (i.e. whole chromosomes). As these genomic sequences are evaluated by gene prediction methods and open reading frames are designated, protein sequence is generated on the basis of nucleotide sequence and contributed to growing protein sequence databases (see[74] for review). Besides the sequencing of individual genes or genomic DNA, messenger RNA (mRNA) is isolated to generate complementary DNA (cDNA) libraries. cDNAs are then partially sequenced to produce expressed sequence tag (EST) nucleotide sequence databases[75]. Often cDNAs are translated to protein sequences and submitted to databases. In the hands of the mass spectrometrist, all three types of DB (protein, EST, and genomic) may be interrogated with mass spectra.

With this first completed genome, biologists began to identify large sets of proteins from *Haemophilus influenzae* using 2-D gels and MS via PMF[76]. With the completion of the sequencing of genomes 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 can be cited 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 DB interrogation, it is becoming increasingly possible to study the proteomes of species with unsequenced genomes. Cross-species identifications have been made in these species and others not cited: *Zea mays*[77,78], *Pisum sativum*[79,80], *Papaver somniferum*[81], *Spinacia oleracea*[82], *Arabidopsis thaliana*[82], *Bos taurus*[83], *Xenopus laevis*[84,85], *Pichia pastoris*[62], and *Trypanosoma brucei*[60,61]. Many earlier studies utilizing cross-species identification of unknown proteins have relied on high mass accuracy MALDI-TOF PMF, and therefore may have identified highly abundant proteins or enzymes conserved across the biosphere. As more sequence-similarity-based methods are being developed and applied, the proteomics of organisms with unsequenced genomes can be envisioned to become more productive and insightful by being able to identify a wider breadth of proteins, i.e. less conserved proteins in closely related species and conserved proteins in distantly related species.

1.2 Questions and Aims of the Thesis

Researchers desire to apply proteomics methods to a breadth of species with unsequenced genomes in an attempt to solve many practical problems and characterize species more thoroughly at the molecular level. The focus of this thesis was to extend the high-throughput capabilities of mass spectrometric protein identification to these organisms. Three goals were to be met during the research:

1. Extend the capabilities of the MS BLAST method to nanoelectrospray Q(q)TOF mass spectrometry for high-throughput analysis; and establish a standard pipeline for protein analysis.

2. Develop a method based on error-tolerant sequence tags for sequence-similarity protein identification to complement the capabilities of MS BLAST. This method should be applicable for high-throughput analysis.

3. Apply the sequence-similarity methods above to problems in cell biology of species with unsequenced genomes.

2 Results and Discussion

2.1 Development of the MultiTag Sequence-Similarity Protein Identification Method

2.1.1 The Sequence Tag for Peptide Mass Spectrum Interpretation

A major limitation to sequence-similarity protein identification rests in the quality of *de novo* interpretation of tandem mass spectra, rather than in DB searching. Tandem mass spectra imperfectly represent the structure of any peptide because upon CID only some peptide fragments are detected which could indicate the amino acid sequence of the peptide. At the same time, spectra often display unpredictable ions that originate from fragmentation of the side chains of amino acid residues, or other fragmentations (or chemical noise), and are not accounted for by typical scoring schemes applied by software for spectra interpretation. It is common in peptide sequencing at femtomole concentrations that low peptide content and high chemical noise allows only a few informative fragment ions to be detected in MS/MS spectra, from which software-assisted interpretation can not produce credible full-length peptide sequence proposals, and subsequent sequence-similarity identification will likely be ineffective for protein identification.

The peptide sequence tag approach for error-tolerant database searching developed by Matthias Mann and Matthias Wilm in 1994 helps to overcome those limitations[86]. The sequence tag utilizes a short (2-4 amino acid residue) sequence stretch, (which can be easily determined from low energy CID spectra acquired from multiply charged precursors) and a pair of masses that lock the determined stretch in the full length peptide sequence; namely the combined mass of all amino acids between the N-terminus of the tryptic peptide and the identified regions, and the mass of all amino acids between the identified region and the tryptic peptide's C-terminus (Figure 4). In stringent DB searches both masses and the sequence are required to match. Currently, sequence tags are employed in protein[87], EST[88], and genomic sequence[82] DB searching. However no statistical evaluation of the significance of matches is provided in these searches. Therefore even if a single hit was retrieved upon DB searching, the match between corresponding peptide sequence from a DB entry and the tandem mass spectrum has to be verified retrospectively by manual inspection; i.e. the predicted fragment ions corresponding to those produced by the theoretical CID of the respective peptide sequence must be overlaid on the spectrum, taking note particularly of coincidence of y-ions in the *m/z* region above the multiply-charged precursor.

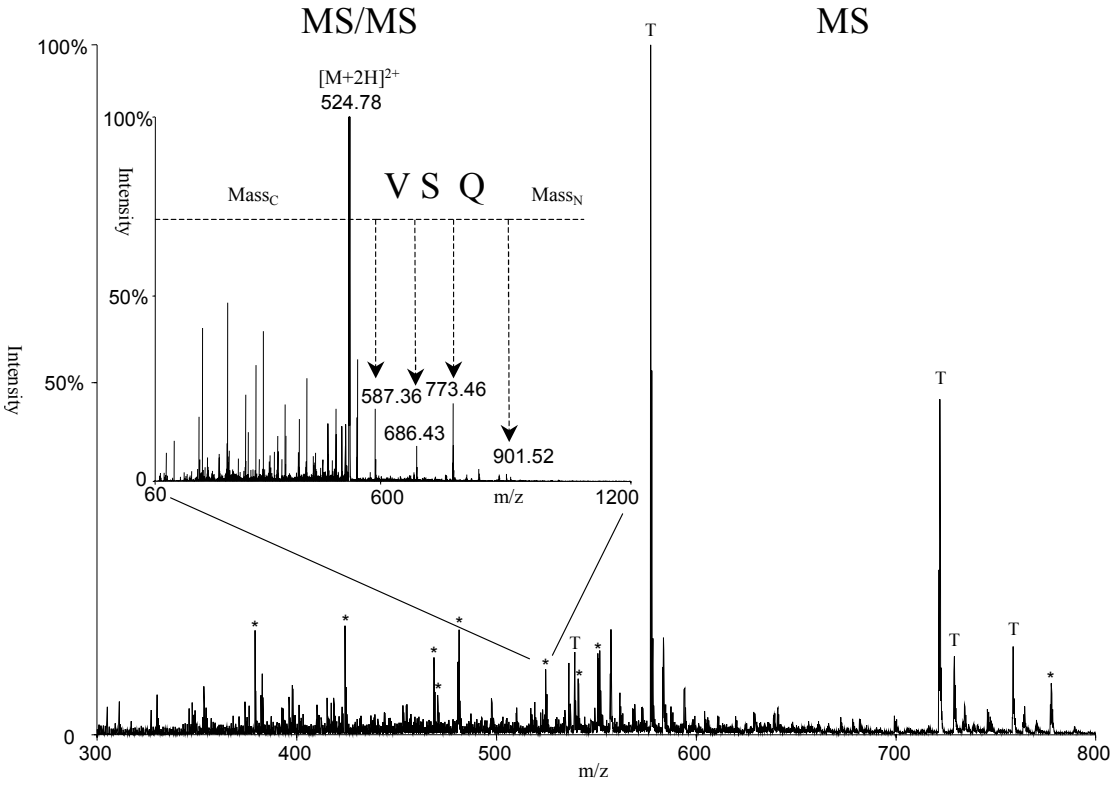

Figure 4 Analysis of a *Xenopus* protein by MS and construction of a sequence tag.

Xenopus proteins were in-gel digested and analyzed by nanoelectrospray tandem mass spectrometry. MS spectrum peaks labeled with a * were fragmented and peptide sequence tags were constructed from MS/MS spectra (inset). Abundant y-ions above the multiply charged precursor in MS/MS spectra allow direct determination of the partial amino acid sequence of a peptide and corresponding sequence tag construction. Peaks in the MS spectrum labeled with a T belong to trypsin. The resulting sequence tag from the MS/MS spectrum shown is (587.36)VSQ(901.52), parent mass 1047.55. All of the determined sequence tags from the analysis of this sample are found in Table 1. The protein was identified as Isoleucyl-tRNA synthetase.

Sequence tags can be used for error-tolerant searching allowing one of the regions of the sequence tag (and, consequently, the intact mass) to mismatch. The approach enables cross-species identifications in protein sequence databases[54]. However, loose matching requirements result in a dramatic loss of search specificity so that typically many hundreds of hits are retrieved, and manual inspection of all of them is tedious.

In the experiments below, the capability of the sequence tag search has been extended with the implementation of a statistical evaluation for the matching of multiple partial sequence tags in the identification of proteins from organisms with unsequenced genomes. Here the MultiTag (MT) approach is demonstrated to enable the identification of

distantly related proteins by sequence-similarity searching using only very short stretches of peptide sequence retrieved from tandem mass spectra, and is therefore a vastly simplified and sensitive method of exploring the proteomes of organisms with unknown genomes.

2.1.2 MultiTag Protein Identification Strategy

MT is a sequence-similarity searching approach for identifying unknown proteins via their homology to known proteins available in sequence databases. Comparison of homologous sequences of proteins from different species often shows that varying amino acid residues are distributed randomly along a polypeptide backbone, with some regions having more conservation than others. Although a single tryptic peptide may not be completely identical between the two protein sequences, their partial identity frequently occurs (Figure 5) Error-tolerant searching with sequence tags can reveal regions of partial identity without determining complete peptide sequences. Although those regions are rather short to claim positive identification of a protein homologue, typically many peptides are sequenced from a protein digest. The MT software reveals proteins to which multiple fragmented peptides are matched in an error-tolerant fashion and computes the statistical significance of the hits to discriminate true hits from false positives.

The first step of the analysis is the construction of peptide sequence tags based on raw mass spectra (Figure 6). Sequence tags are typically called from the high *m/z* region of tandem mass spectra of tryptic peptides, which are dominated by abundant y-ions and partial interpretation of the spectrum is straightforward. Sequence tags were assembled for as many fragmented tryptic peptides as possible and were used for searching a DB in a stringent fashion (matching regions 1, 2 and 3) and error-tolerant fashion; a search tolerating a mismatch of the C-terminal mass (matching regions 1 and 2); a search tolerating a mismatch of the N-terminal mass (matching regions 2 and 3); and searches tolerating one mismatch in the amino acid sequence (matching regions 1 and 3); the hits were additionally encoded by the mass of the precursor ion and by the abbreviated matching region (NC, N, C, or E, respectively) in the sequence tag. Importantly, matches of retrieved sequences to corresponding tandem mass spectra were not further inspected, and the redundant hits (matching the same peptide sequence in another DB entry, or in another search) were not removed. If stringent searches (i.e. with regions 1, 2 and 3 matched) retrieved many candidate sequences no additional verification of hits was performed. The full list of hits was then submitted to the MT program. The software identified multiple hits originating from the same protein entry, eliminated redundant hits to the same peptide in the same entry

and assigned the significance to all matches by computing an estimate of the probability that such a combination of tags may hit a protein entry at random.

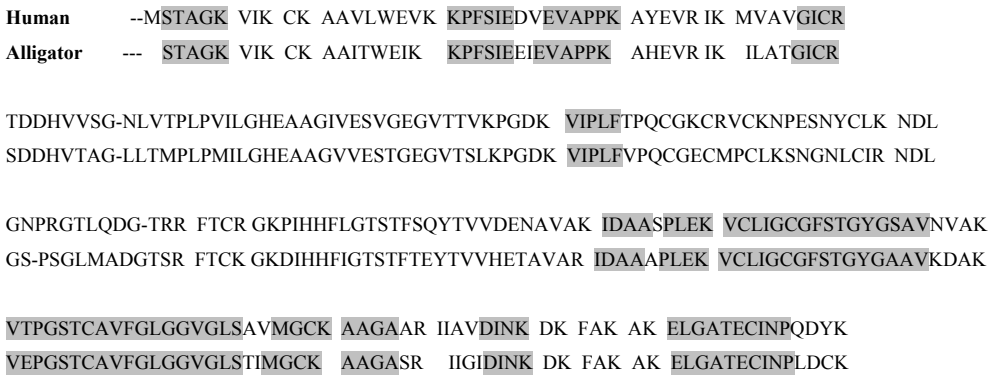

Figure 5 Sequence alignment of the human and alligator ADH protein.

Partial protein amino acids sequences for alcohol dehydrogenase are aligned above from human and alligator (75% identity). Regions alignable by error-tolerant sequence tags between the two sequences are highlighted in gray. These regions are theoretical tryptic peptides over six amino acids in length with ≥ three conserved amino acids from the N-terminus or ≥ four conserved amino acids from the C-terminus. Tryptic cleavage sites designated above are shared between both sequences. Tryptic cleavage sites not at the same point on the sequences are not designated by spaces; sites do not occur in the gray regions. Accession numbers: human, P00325; alligator, AAB28120. The sequences were aligned using the Clustal X program.

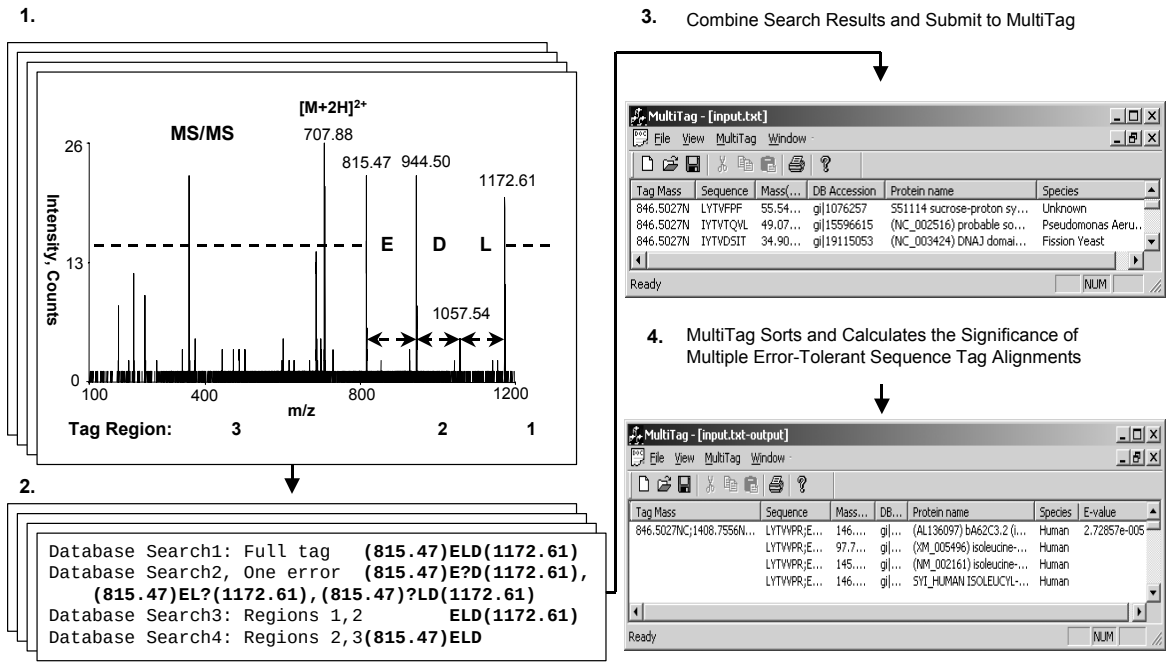

Figure 6 MultiTag method schematic.

The MT approach consists of constructing sequence tags from peptide tandem mass spectra, error-tolerant database searches, and sorting and calculation of the significance of multiple error-tolerant sequence tag alignments by the MT software. Panel 1. shows a tandem mass spectrum of a low abundance peptide with an overlaid sequence tag. Panel 2. shows one complete and three error-tolerant sequence tag database searches, which is done for each MS/MS spectrum and corresponding sequence tag. Panel 3. shows the combined list of search results (most of the 8000 entry list is not shown) from all spectra and all searches in the analysis of a single sample; "Tag Mass" column indicating the tag's parent mass followed by an "NC" for search results with complete tags, an "N" for searches with tag regions 1 and 2, an "E" for searches with tags with one amino acid error, or a "C" for searches with tag regions 2 and 3; "Sequence" column is the retrieved sequence found from the database search; "Mass" column indicates the protein's total mass in kDa from which the peptide originated; "DB Accession" the proteins accession number; "Protein name"; "Species". Panel 4. shows the MT output; "Tag Mass" column lists the tag-search code for the tags aligned; "Sequence" lists all of the full peptide sequences error-tolerantly aligned; "Mass"—"Species" same as Panel 3; "E-values" for the probability of the alignment of the group of sorted sequence tags.

2.1.3 Calculation of E-values

(in collaboration with Professor Shamil Sunyaev)

The major problem of DB searching with multiple sequence tags is the need to identify hits corresponding to truly homologous proteins in the large pool of randomly matching proteins produced of multiple degenerate searches, and therefore the evaluation of statistical significance of hits is ultimately required. The classic way to interpret the results of a DB search in the statistical framework is to assign an E-value to each hit resulting from a

search. E-values corresponding to a hit represent the expected number of better or equally good matches found in a DB at random. In the case of the MT search, a DB search hit is a protein sequence, which matches some sequence tags in a degenerate or a non-degenerate manner. E-values here give the expected number of sequences from a random DB, which would match the same combination of tags in the same way or even more specific (less likely) combination of tags in a more specific way. In order to compute E-values, the probability that a given tag with a given type of degeneracy would match a random amino acid sequence had to be determined. The probability that a given combination of tags would match a random sequence can then be computed as a product of the probabilities corresponding to individual matches. Further, the probability that any possible more specific (less likely) combination of tags than a given combination would match a random sequence has to be determined. Finally, the E-value would be given by multiplication of the latter probability to the total number of DB sequences. Below the detailed consideration of each of those steps is presented.

Let us consider a sequence tag, which is represented by an N-terminal mass m_N, three amino acids a_1, a_2, a_3 and C-terminal mass m_C. The probability that a random tryptic peptide would match this tag in a non-degenerate manner would be given as a product of three following probabilities. First, the probability that the random tryptic peptide has an N-terminal fragment of any length, whose mass lies in the interval (m_N-Δm, m_N+Δm), where Δm is mass tolerance of the instrument. Second, the probability that this fragment of random peptide has amino acids a_1, a_2 and a_3. This is simply given by the product $f(a_1)f(a_2)f(a_3)$, where $f(a_i)$ denotes frequency of amino acid a_i. And third, the probability that the mass of the random peptide fragment between these amino acids and the C-terminus would be between m_C-Δm and m_C+Δm.

In order to derive probabilities corresponding to m_N and m_C one would regard the mass of a random tryptic peptide being a result of a random process. We imagine that the sequence of the random tryptic peptide was constructed by a random generator, which consequently generates amino acids, one at a time, and the probability that next coming amino acid will be a_i is given by its frequency $f(a_i)$. Obviously, at each moment of time the generator can produce a trypsin cleavage site (K or R residue) with the probability $q=f(K)+f(R)$ and thus stops the process. The mass M of the random tryptic peptide can be regarded as an accumulated sum of masses of randomly generated amino acids:

$$M = M_1 + M_2 + M_3 + \ldots \qquad (1)$$

In probability theory, random values represented as successive sums of positive identically distributed variables as in equation (1) are called a renewal process[89]. Obviously masses of randomly generated amino acids obey the probability distribution determined by amino acid frequencies, so that the probability *p(m)* that the random mass would be exactly *m* is given by combined frequency of amino acids of mass *m*. Then, the distribution of the mass accumulated after *n+1* step, i.e. the probability that the peptide fragment of length *n+1* would have its mass smaller than *t* can be computed *via* successive convolutions:

$$F_{n+1}(t) = (1-q)\sum_i F_n(t-m_i)\,p(m_i) \qquad (2)$$

Summation here is carried over all values of amino acid masses. Multiplication to *(1-q)* is needed to take into account that the process survived the *n+1*-th step, *i.e.* the tryptic peptide has more amino acids than *n+1*.

The distribution of the total mass of the tryptic peptide, i.e. the probability that the peptide's total mass would not exceed *t* is given by allowing for all possible lengths of the peptide:

$$F(t) = q\sum_{n=0}^{\infty} F_n(t) \qquad (3)$$

Which implies that the probability that the peptide's mass would be in the interval from $m-\Delta m$ to $m+\Delta m$ is

$$P(m,\Delta m) = F\left(m+\Delta m\right) - F\left(m-\Delta m\right) \qquad (4)$$

Although not intuitively obvious, this formula holds both for the whole mass of the peptide and for any of its fragments between a fixed amino acid position and the cleavage site. Indeed, if we further consider our analogy with the renewal process, it will retain its properties regardless of the point we consider the process started (the process has no memory). Therefore, after the position of the sequence tag on the peptide sequence is fixed

through matching one mass and the short sequence stretch, the probability that the second mass would also match is given by equation (4).

Now consider matching of the sequence tag as a sequence of three consecutive independent events, namely match of the first mass, match of the short sequence stretch and the following match of the second mass. Although, the consideration is obviously symmetric with regard to N- and C-termini of the peptide, without loss of generality, we would assume that the N-terminal mass is the first mass to match. The probability that the mass of any N-terminal fragment of the peptide would be in the interval (m_N-Δm, m_N+Δm) is given by:

$$Q(m_N, \Delta m) = \frac{1}{q}\left[F(m_N + \Delta m) - F(m_N - \Delta m)\right] \qquad (5)$$

or

$$Q(m_N, \Delta m) = \frac{1}{q} P(m_N, \Delta m) \qquad (5)$$

A multiplier *1/q* was introduced because in this case the process survives the step with this mass, i.e. all peptides with arbitrary lengths with N-terminal parts matching the mass would satisfy the condition. We note that the equation 5 holds only if mass tolerance of the instrument is lower than any of amino acid masses, otherwise it corresponds to the expectation and not to the probability.

Since the probability of the non-degenerate match of the sequence tag would be a product of probabilities of the N-terminal mass match (which importantly fixes the position of the tag along the peptide), sequence stretch match and the C-terminal mass match it will be expressed as:

$$P_{non-\deg enerate} = \frac{1}{q(1-q)} P(m_N, \Delta m) f(a_1) f(a_2) f(a_3) P(m_C, \Delta m) \qquad (6)$$

Additional multiplier *1/(1-q)* simply reflects the fact that we do not consider zero length tryptic peptides, allowed by the model if the cleavage site comes at the first step. Therefore, we work only with *1/(1-q)* fraction of realistic peptides.

Examples of probabilities for degenerate matches are given by:

$$P_{N-ter\,\min al} = \frac{1}{q(1-q)}\left[1 - P(m_N, \Delta m)\right] f(a_1) f(a_2) f(a_3) P(m_C, \Delta m)$$

and

$$P_{\sec ond\_residue} = \frac{1}{q(1-q)} P(m_N, \Delta m) f(a_1)\left[1 - f(a_2)\right] f(a_3) P(m_C, \Delta m) \qquad (7)$$

As the next step, we compute the probability that a random protein sequence containing *K* tryptic peptides would match multiple sequence tags, taking into account the tags being matched and the type of degeneracy of the match. For instance, if we had three sequence tags and the random sequence matched simultaneously sequence tag 1 with an error in the N-terminal mass, sequence tag 2 with an error in the C-terminal mass and sequence tag 3 with a mismatch at the second identified amino acid, the probability of the event would be given by:

(8)

$$P = \left(1 - e^{-K \cdot P_1(m_N) \cdot f(a_{11}) \cdot f(a_{12}) \cdot f(a_{13})}\right) \cdot \left(1 - e^{-K \cdot f(a_{21}) \cdot f(a_{22}) \cdot f(a_{23}) \cdot P_2(m_C)}\right) \cdot \left(1 - e^{-K \cdot P_3(m_N) \cdot f(a_{31}) \cdot f(a_{33}) \cdot P_3(m_C)}\right)$$

This example shows how to compute the probability that a random amino acid protein sequence would match an arbitrary combination of sequence tags.

In order to calculate E-values, we should first compute the probability that any combination of tags would match a random amino acid sequence, which is equally or more specific than the combination observed. In other words, we will need to sum up probabilities (eq. 8) of all possible matches, which do not exceed the probability of the actual DB hit. It is definitely too demanding computationally to directly enumerate all less likely combinations of tags. However, it appears to be much easier to enumerate all combinations, which are, in opposite, more likely to happen because they mostly involve matches with a very few tags. Therefore, we compute the probability that a random sequence would produce a less specific match than the actual hit (taking care of possible statistical dependence of various combinations of tags) and subtract the result from 1. E-value is then computed by multiplying the result to the DB size.

A series of computational simulations have been carried out to validate the computation of E-values described above.

Software implementation of MT uses pre-computed distribution function *F(t)*. The software imports sequence tags in the conventional format $(m_c)a_1...a_n(m_n)$[86] and peptide mass, and computes probabilities for each tag to match a random tryptic peptide. Further, the software imports a full list of hits produced by multiple degenerate and non-degenerate sequence tag searches and identifies hits corresponding to the same protein. For each hit MT first computes the probability of the match (similarly to equation 8). Then, it identifies all tag combinations giving the same or higher probability, and based on this information, assigns E-value to the hit. At the final step, MT sorts all hits according to E-values.

2.1.4 Specificity, Performance, and Limitations of Error-Tolerant MultiTag Searching

MT aligns multiple partial and/or complete sequence tags to increase the coverage of a DB sequence from available MS/MS data to raise the significance and lower the E-value of identifications. Sequence tags were used from the identification of DNA polymerase (Table 1) to perform alignments with MT to demonstrate factors that contribute to final E-values (Table 2). High E-values are given for "poor" quality tags that have short mass lengths for tag regions 1 and 3, and designate common amino acids with a high frequency in proteins, i.e. Leucine, "L." Lower E-values are given for uncommon amino acids such as Tryptophan, "W," or for tags with more amino acids in the sequence stretch. The probability that a combination of partial sequence tags will match a single DB entry is lower than if an individual tag is matched, with an increasing significance as more partial sequence tags are aligned. Two partial sequence tags were found to be not significant enough for a confident identification in some cases, depending on the character of the tags. However, the alignment of three or more partial sequence tags lowers E-values to the range of 1E-6-1E-9 when mass accuracies are sufficiently high, enabling confident protein identification. Sequence tags assembled with narrower mass tolerance increase the specificity of DB searching and lower the E-values of hits. As in case of conventional DB sequence-similarity searches, specificity of the MT search decreases with the growing size of the DB.

An intrinsic problem to all statistical approaches to homology searches relying on average amino acid frequencies is posed by low complexity regions and other proteins and/or protein regions with amino acid frequencies, which strongly deviate from the DB average[90]. If a MT identification results in a peptide from a low complexity region or in a peptide of obviously special amino acid composition, these identifications have to be

interpreted with caution, since the underlying statistical model does not account for bias in amino acid composition.

An advantage of MT over MS BLAST, besides its ability to represent noisy and low intensity spectra, is that peptide sequences retrieved by sequence tag searches can be overlaid on fragment ion spectra allowing one to determine whether the retrieved sequence is the correct sequence; this is less direct with MS BLAST. Even though relatively weak matches can be evaluated in this way, the MT approach is suited for a high-throughput setting without the need to go back to the original spectra to evaluate the identifications. The determination of the significance of sequence tags by the MT statistics takes the place of retrospective manual data evaluation.

MT, as well as any sequence-similarity searching method, is prone to errors if the analyzed protein contains low complexity sequence regions, i.e. collagen, Glycine-rich cell wall proteins, and silk proteins. Because MT only recognizes a few central amino acids accurately, it would be possible that multiple sequence tags to different regions of the same proteins could not be distinguished, thus diminishing the overall score.

2.1.5 Identification of Proteins from *Xenopus laevis* by MultiTag Searching

The MT approach was applied to the identification of proteins isolated from the African clawed frog *Xenopus laevis*. In-gel digests of *Xenopus* proteins were analyzed by PMF and NanoESI-MS/MS. Sequence-similarity searching methods were applied for protein identification because Mascot DB searching with peptide mass fingerprints and with lists of fragment masses derived from uninterpreted tandem mass spectra were unable to identify proteins by stringent matching. Two methods of sequence-similarity searching were applied in parallel to the same set of MS/MS data. Peptide sequence proposals obtained by automated *de novo* interpretation of tandem mass spectra were submitted to MS BLAST searching. In parallel, peptide sequence tags were assembled *via* partial manual interpretation of spectra (Figure 4), followed by error-tolerant DB searching and sorting and evaluating the results by MT, as described above (Table 1). From five attempted unknown proteins, MS BLAST identified three, however all five were identified by MT. Importantly, in three cases both MT and MS BLAST identified homologous sequences from the same organism or from different species, providing an independent validation of the MT approach.

Table 1 Sequence Tags used in the Identification of *Xenopus* Proteins by MultiTag.

MultiTag Protein Identifications	Tags Submitted	Mass	Matching Tags	MS BLAST Identifications	Alignments
Isoleucyl-tRNA Synthetase	(371.24)VTY(734.42)	846.5	(371.24)VTY(734.42)	No identification	None
Human P41252	(637.34)VL(849.49)	935.51			
E-value: **2.73E-5**	(587.36)VSQ(901.52)	1047.55	(587.36)VSQ		
	(488.34)LEL(843.55)	1099.56			
E-value: First False Positive: 0.15	(602.42)EQ(859.53)	1134.64			
	(979.62)DVS(1280.74)	1408.76	DVS(1280.74)		
	(1166.60)DLL(1507.80)	1619.81			
	(1363.59)VV(1561.73)	1918.98			
	(985.50)NT(1200.59)	2329.25			
Glutamyl-Prolyl-tRNA Synthetase	(492.31)TY(756.42)	902.49	(492.31)TY	No identification	None
Human XP_001958	(559.31)QAS(845.44)	942.03			
E-value: **7.14E-7**	(559.33)QVS(873.49)	971.57			
	(545.27)LWT(945.48)	1058.58			
E-value: First False Positive: 0.52	(705.35)LE(947.47)	1192.62			
	(626.40)LLDE(1096.64)	1357.67			
	(1177.60)LLA(1474.81)	1544.87	(1177.60)LLA(1474.81)		
	(926.54)FSLTDT(1590.84)	1688.94			
	(561.34)AVEP(957.54)	1711.92	(561.34)AVEP		
DNA Polymerase Delta	(401.30)PVP(694.48)	750.5		DNA Polymerase Delta	
Human S35455	(486.38)SE(702.45)	772.49		Human P28340	
E-value: **8.14E-7**	(441.21)PF(685.33)	797.42			
	(456.31)FT(704.42)	816.48	(456.31)?T(704.42)		LTFALPR
E-value: First False Positive: 0.35	(345.25)QEL(715.43)	827.49	(345.25)QEL(715.43)		
	(385.28)LY(661.42)	846.49	(385.28)LY(661.42)		DAYLPLR
	(421.29)EAW(807.44)	877.48			
	(583.32)LGG(810.44)	908.5	LGG(810.44)		VGGLFAFAK
	(543.30)LNL(883.51)	995.6			
	(470.29)FVL(829.51)	1071.58			
	(533.32)LPE(872.50)	1131.65	LPE(872.50)		
	(889.45)QS(1104.54)	1190.56			
Hsp70/Hsp90 Organizing Protein	(494.26)DSLL(922.48)	992.53		Stress-Induced	
Chinese Hamster AAB94760	(408.23)FQLA(867.48)	995.51		Phosphoprotein STI1	
E-value: **7.82E-9**	(550.27)ELL(905.48)	1017.56	(550.27)E?L(905.48)	*Xenopus* AAM77586	
	(585.40)GVDF(1003.59)	1115.67	(585.40)G?DF(1003.59)		LFDVGLLALR
E-value: First False Positive: 0.59	(674.38)NGAS(1003.52)	1186.65	NGAS(1003.52)		ALSAGNLD
	(416.24)ELL(771.45)	1350.72	(416.24)ELL		
	(856.51)NLYA(1317.73)	1415.8			VAYLNPD
Heat Shock Protein 90-beta	(385.26)FLL(758.50)	828.53	(385.26)FLL(758.50)	Heat Shock Protein 90-beta	ALLFLPR
Zebrafish NP_571385	(567.28)Y(730.35)	876.43	(567.28)Y(730.35)	Salmon AF135117(Nucleotide)	FYDGFTK
E-value: **4.35E-9**	(401.29)ES(617.37)	729.45	(401.29)ES(617.37)		LSELLR
	(708.37)NAV(992.52)	1234.64	(708.37)N?V(992.52)		
E-value: First False Positive: 0.0051	(716.34)NLL(1056.55)	1241.69	NLL(1056.55)		
					LTPDQPVV

For each sample, sequence tags were constructed from multiple MS/MS spectra from the analysis of a single in-gel digest ("Tags Submitted") and error-tolerantly searched against a protein database (resulting list of entries not shown). The "Mass" column contains the corresponding complete mass for each sequence tag. Results were sorted by MT. Groups of matching partial sequence tags resulted ("Matching Tags"). E-values for the group of partial sequence tags were calculated by the MT software (first column in Bold). The final MultiTag report gave a list of database entries with diminishing E-values (data not shown). E-values are cited (column 1, "First False Positive") for the first database entry in the list to not correspond by annotated function (i.e. HSP 90) to the most significant hit. Protein identifications made by MS BLAST are found in the column "MS BLAST Identifications" and peptide sequences aligned are in "Alignments."

This data demonstrates that MT can outperform the more generic sequence-similarity searching tool—MS BLAST—when *de novo* sequence prediction is unable to produce meaningful peptide sequences from noisy or low intensity spectra. On the other hand, MT successfully identified the proteins because sequence tags are easily assembled from tandem mass spectra where complete amino acid sequence prediction is impossible (Figure 4). The results suggest that three and more error-tolerantly matching sequence tags may unequivocally identify a homologous protein (Table 2), despite none of the sequenced peptides exactly matched the corresponding sequence from a DB entry and sequence stretches of less than four amino acid residues were determined. Both MT and MS BLAST were able to identify the proteins not identified by Mascot because they could tolerate amino acid substitutions, resulting in an offset of the peptide's total mass.

2.1.6 Homologue Identification Specificity of MultiTag Searching

By its algorithm, MT is a less generic sequence-similarity searching tool, compared to MS BLAST and FASTS since it requires identical (although short) stretches of peptide sequence for protein identification. We roughly estimated the scope of MT identification from the bottom, assuming the most unfavorable model when identical amino acids between proteins are distributed uniformly along the sequence. According to our experience and table 2, three partial matches normally give a statistically significant match. We have estimated the chance to obtain three partial matches and its dependence on the overall identity of the complete query sequence and the DB sequence. A very simple calculation assumes that the probability that a single amino acid would match between the query and the DB sequence is equal to the overall sequence identity and is independent of the sequence region and amino acid type. Assuming further a query of 10 identical tags we estimated that the MT method is able to identify almost all homologues at the level of 80% sequence identity, 75% of homologues at the level of 75% sequence identity, but only about 45% of homologues at the level of 70% sequence identity. Obviously, MT cannot achieve the specificity of the methods using the knowledge of longer sequence parts. According to simulations results, sequence based methods like MS BLAST and FASTS are able to detect about 50% of homologous sequences at the sequence identity level of ~50%[61]. According to our lower limit estimate, MT would require 71% sequence identity (in reality less) to reach the same efficiency of identifications. However, simulations with MS BLAST and FASTS were performed assuming all sequence predictions are correct, which is rarely the case. Therefore, the advantage of using MT is the ability to identify sequence similarities at

the reasonable level with high robustness with respect to the quality of the raw data and independently of the quality of automated or manual *de novo* sequence prediction techniques.

Table 2 MultiTag E-values are Dependent on Amino Acids in Tag, Number of Tags, Mass Accuracy, and Database Size.

				E-values	*PredCount*	E-values	
	Mass	Sequence Tags in the identification of DNA Polymerase	1.0 Da*	0.5 Da*	0.1 Da*	0.1 Da*	1,600,000 DB Entries**	200,000 DB Entries**
1	816.48	(456.31)?T(704.42)	6.06E+03	2.73E+03	2.56E+03	*1.76E+02*	5.11E+03	6.39E+02
2	827.49	(345.25)QEL(715.43)	1.96E+02	8.83E+01	8.34E+01	*1.23*	1.67E+02	2.09E+01
3	846.49	(385.28)LY(661.42)	2.61E+02	1.30E+02	1.29E+02	*2.66*	2.58E+02	3.22E+01
4	908.5	LGG(810.44)	8.36E+03	6.53E+03	6.92E+03	*9.81E+02*	1.38E+04	1.73E+03
5	1131.7	LPE(872.50)	1.73E+03	1.01E+03	9.52E+02	*7.06E+01*	1.90E+03	2.38E+02
6		LGG(810.44) + LPE(872.50)	5.08E+01	2.71E+01	2.51E+01	*9.23E-02*	5.16E+01	6.45E+00
7		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42)	1.24E-01	3.09E-02	2.86E-02	*2.17E-05*	5.72E-02	7.15E-03
8		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42) + (385.28)LY(661.42)	1.97E-05	3.23E-06	3.01E-06	*7.68E-11*	6.02E-06	7.53E-07
9		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42) + (385.28)LY(661.42) + (345.25)QEL(715.43)	1.49E-06	8.60E-07	8.14E-07	*1.26E-16*	1.63E-06	2.03E-07
10		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42) + (385.28)LY(661.42) + (345.25)QEL(715.43)	3.64E-09	4.17E-09	3.55E-09	*1.31E-16*	7.11E-09	8.88E-10

The E-value in Bold is shown in Table 1. In the calculation ofE-values for row 9, all tags submitted were included from Table 1. In row 10, only the tags that matched the database entry were included in the list of tags submitted for MT calculations (reduced query length. 800,000 and 200,000 database entries correspond approximately to the NCBI Nonredundant (nrdb) and SwissProt protein databases, respectively. “Mass” in column 2 indicates the full length of the peptide corresponding to the sequence tag in column 3. *800,000 database entries, **Mass Accuracy of 0.1 Da.

2.1.7 Enhanced Error-tolerant EST Database Searching by Tandem Mass Spectrometry and MultiTag Software

In general, there are a number of difficulties when searching EST databases with MS data. First of all, ESTs represent nucleotide sequences, whereas amino acid polymers are analyzed by MS. This however can easily be overcome by these sequences inherent colinearity, which allows the translation of either EST sequences, or amino acid queries, in six frames into theoretical amino acid sequence, or nucleotide sequence, respectively, for DB searching. Secondly, ESTs are generally short sequences, translating into ~150 amino acids of polypeptide sequence each. Therefore, many large proteins (i.e. 80-300kD) will only be represented by a short sequence stretch, and we would expect many analyzed peptides to be left unaligned. This may be overcome by assembling multiple EST sequences into cDNA clones of the expressed genes, potentially covering the length of the expressed protein. Thirdly, EST sequences are generated by single-pass sequencing of cDNA clones (generated from mRNAs), which would likely result in multiple errors. Thus, even in searching sequences from the organism of origin (not cross-species), an error-tolerant method such as MT would be expected to be more sensitive than a method that demanded exact matching because more partial peptide sequences could be aligned to produce a higher coverage and more significant alignment.

Protein identification by the interpretation of tandem mass spectra with sequence tags and EST DB searching has relied upon the searching of databases one sequence tag (and thus one spectrum) at a time. Since the sequence tag is only a partial interpretation of a spectrum, multiple degenerate DB sequences are recognized in most searches, matching entries from many different species and proteins with varying molecular weights (when a tag is searched error-tolerantly, even more entries are retrieved). Therefore, these retrieved sequences must be manually inspected and the false positives must be discriminated against in a time-consuming operation. The MT method overcomes this problem by correlating search results from multiple spectra to determine the most probable protein identification(s). MT sorts DB search results (including combined results from full and partial tags) by their statistical significance and assigns an E value for every set of alignments, thus indicating alignments otherwise missed and greatly facilitating interpretation of sequence tag-driven DB search results. Since there is often an abundance of EST sequence data *in silico*, the utilization of these sequences, either as independent resources or by applying alternate DB searching strategies simultaneously (reviewed in section **2.5.3**) could greatly facilitate protein identification.

To test the specificity of MT versus Mascot in EST DB searching, a model dataset was used that was generated in a screen of microtubule-associated proteins from *Xenopus laevis* (see section **2.2**). Gel separated *Xenopus* proteins were analyzed by nanoelectrospray M/MS and identified by protein DB searching using multiple techniques, which gave significant matches for a single protein often with Mascot, MS BLAST, and MT (and corresponding greater sequence coverage with the later methods generally).

To facilitate the DB searching process with sequence tags, a script has been developed (in collaboration with Applied Biosystems, Foster City, CA, USA) for automated error-tolerant searching to generate unsorted search results for submission to the previously described MT software[91]. This script "MTSearch" was developed specifically for BioAnalyst QS (Applied Biosystems, CA) to automatically search a list of complete and error-tolerant sequence tags against a DB and compile the results in an unsorted list. The previously described MT software subsequently sorts search results by the statistical significance of combinations of multiple tags and individual tags. The results of MTSearch can be directly submitted to the MT statistics software (Figure 7).

A modification was made to MT for EST DB searching. MT relies upon an expected number of peptides per protein sequence, which was previously averaged at 41 peptides per protein for protein DB searching. The average protein length in a non-redundant DB was previously determined to be 492 amino acids (corresponding to ~60kD). The average length of a tryptic peptides was designated at 12 amino acids, setting the average number of tryptic peptides per DB entry at 41. Since EST DB sequences are shorter, we would expect fewer peptides possible from each entry; therefore the parameter designating the number of peptides expected per DB entry was made adjustable to account for differences in length. Secondly, MT relies upon a designated size of the DB searched for producing a probability of a random match. This is straightforward with protein DB searches because this is a designated number of DB entries. For EST DB searching, all nucleotide sequences or the query must be translated in six frames, generating additional erroneous hypothetical sequence; only one frame is the correct translation; the number of entries were multiplied by 6 to account for this degeneracy.

A set of DB searches was performed using Mascot, relying upon its own statistics, and MT, relying upon E values for the determination of true versus false positives, as a trial to judge the sensitivity of MT. Peak lists with corresponding intensities were generated automatically from tandem mass spectra of protein digests and were submitted to Mascot for DB searching. Sequence tags were constructed by manual interpretation from the same

set of tandem mass spectra; averaging ~9 sequence tags per protein digest, which required ~4 minutes of spectrum interpretation per tag. Sequence tags were used for DB searching and the results were analyzed by MT. EST sequences from *Xenopus laevis*[92] were used as reference by both methods for protein identification. In general, MT can make direct identifications or cross-species identifications.

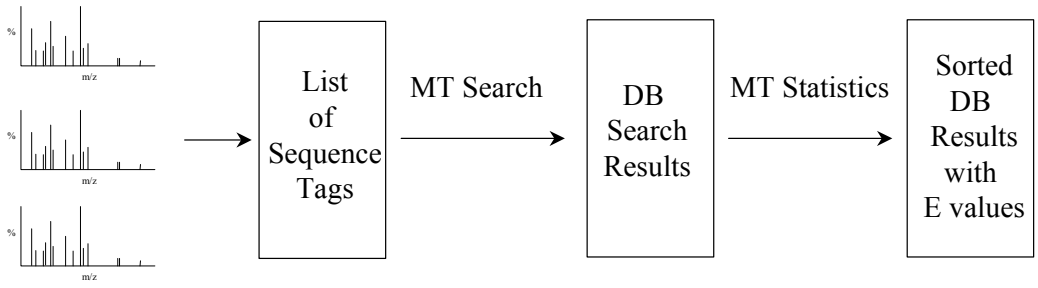

Figure 7 Integrated MultiTag database searching scheme

To interpret peptide tandem mass spectra with MT for DB searching: 1. Construct tags and list in text file; 2. Run DB search script; 3. Submit search results to MT software for sorting by probabilities. MT search script is written for Applied Biosystems Bioanalyst QS software. MT software was modified for EST DB searching as described above.

The MT software reports a "Predicted Count" (PredCount) value and an E value for every alignment. The employed statistical model implies that the first species-specific false positive should be detected at an E value of approximately 1. The second false positive should have an E value of 2, and the third false positive an E value of 3, etc. Because of the imperfections of the statistical model, E values less than 0.1 generally indicate true matches, with more significant matches having lower E values. PredCount values reflect the specificity of matches; however, PredCount does not reflect the expected number of false-positives when the entire query is searched against a DB. Contrary to E values, PredCount values very weakly depend on the number of tags in a query, and low PredCount values serve as a further statistical indicator of true matches when E values are high due to large queries where few sequence are aligned.

From the model data set, Mascot was able to recognize 49 peptides with optimized settings, making 20 identifications (Table 3). From the same dataset, MT was able to recognize 87 peptides and produced 31 identifications, which included all of the Mascot

identifications. Whereas many identifications are statistically at the borderline by Mascot, MT was able to increase the coverage of these alignments by error-tolerant matching of partial peptides to provide more evidence for a positive identification in 12 cases. Furthermore, MT was able to make significant alignments in 11 cases where Mascot could not.

Where MT was able to make a significant alignment and Mascot produced no significant matches, the top five Mascot hits were inspected to see if the same protein had been nonconfidently detected. Since it is possible that MT may recognize one specific EST with a query and Mascot may recognize a different EST corresponding to the same cDNA sequence, these top hits were carefully inspected to find alternate ESTs matching the same protein sequence. Where MT was unable to make a significant alignment, the top five MT hits were manually inspected by overlaying the retrieved peptide sequence "on the spectrum" using BioAnalyst QS, and comparing the observed fragment ions with theoretically calculated fragment ions (at a precision of 0.001 m/z), taking into consideration abundant a, b and y series ions, and immonium ions. In this manner, MT was able to detect 6 additional matches; 3 of these 6 were not in the Mascot top 5 hits (data not shown). This suggests that the MT method can also retrieve true matches that are not statistically significant; single hits below threshold should be manually inspected if no other alignments are made.

From this data, MT proves to be a sensitive method for EST DB searching, both because more identifications were made than the conventional software and more peptides were identified in total, resulting in a higher coverage of proteins.

The MT approach for enhanced EST DB searching balances the specificity and sensitivity of mass spectra interpretation using sequence tags with the sequence tags' inherent degeneracy in DB searching. In cases where EST sequences would be assembled as cDNA clones, we would expect even higher coverage because often MT hit multiple tags for different EST sequences of the same cDNA sequence (data not shown). Furthermore, MT also would be expected to have fewer false positives because there is no cutoff threshold (like conventional softwares), and all borderline hits must be discriminated against by manual inspection. However, MT does require manual spectrum and data interpretation that is not required by conventional softwares, and because of the effectiveness of the conventional softwares, MT would be applied most efficiently in cases where other methods fail to make an identification, or where they have recognized certain proteins only on the borderline of their scoring thresholds.

Table 3 EST database searching: Mascot vs. MultiTag

MW	Mascot EST_others	P.	MultiTag EST_others	Tags	Query	PredCount	E value
175	Glutamyl-propyl-tRNA syn., 12748494	1	Glutamyl-propyl-tRNA syn., 14989013	2	11	2.70E-13	2.10E-04
175	Glutamyl-propyl-tRNA syn., 14989013	1	Glutamyl-propyl-tRNA syn., 17398490	2 & 1	9	1.10E-12	1.10E-06
165	†		Glutamyl-propyl-tRNA syn., 24091165	2	9	1.70E-08	5.70E-06
165	†		Glutamyl-propyl-tRNA syn., 12746970	1	4	5.07E-04	3.81E-03
160	Hyaluronan mediated receptor, 13252946	1	Hyaluronan mediated receptor, 13252946	2	9	5.82E-09	1.42E-04
155	†		Isoleucyl-tRNA synthetase, 24090398	2 & 1	9	2.31E-13	7.39E-08
150	†		Leucyl-tRNA synthetase, 21870435	2	12	3.36E-06	1.57E-03
150	†		Leucyl-tRNA synthetase, 21870435	2	6	3.81E-06	3.36E-04
122	Kinesin heavy chain, 17418741	3	Kinesin heavy chain, 17418741	2	9	1.42E-08	5.37E-06
122	Kinesin heavy chain, 12480559	2	Kinesin heavy chain, 12480559	2 & 1	7	6.79E-12	1.94E-08
118	Kinesin heavy chain, 17418741	2	Kinesin heavy chain, 17418741	3	10	1.27E-16	3.13E-08
	Kinesin heavy chain, 12480559	2	Kinesin heavy chain, 12480559	3	10	8.96E-16	3.13E-08
100	*Elongation factor-2, 11787464*	*2*	Elongation factor-2, 21875348	2	6	3.73E-06	1.27E-02
90	Heat shock protein, 10065828	2	Heat shock protein 90-beta, 21873865	3 & 1	13	1.12E-17	8.96E-07
	Glutaminyl-tRNA synthetase, 7699102	*2*	Glutaminyl-tRNA synthetase, 7393733	3	13	1.57E-14	8.96E-07
85	†		Cytoplasmic dynein inter. chain, 24082627	3	9	1.12E-17	4.48E-06
70	Heat shock cognate-70, 21384290	5	Heat shock cognate, 24087000	2 & 1	7	5.00E-15	3.43E-08
68	Lysyl-tRNA synthetase, 17580417	2	Lysyl-tRNA synthetase, 24097853	2	4	5.00E-06	3.21E-04
68	Lysyl-tRNA synthetase, 12473885	3	Lysyl-tRNA synthetase, 12473885	3	9	4.03E-14	1.12E-07
68	HSP70/HSP90 org. protein, 17395146	2	HSP70/HSP90 org. protein, 21874237	2 & 1	7	3.73E-15	6.19E-09
52	Alpha-tublin, 12471404	2	Alpha-tublin, 21863612	2	16	7.46E-09	3.88E-03
	Formiminotrans. cyclode., 17413939	*1*	Formiminotrans. cyclode., 12471624	2	16	5.67E-10	3.88E-03
50	Alpha-tublin, 12471404	4	Alpha-tublin, 24097682	3 & 1	18	9.70E-20	2.01E-06
	Beta-tublin, 17425087	2	Beta-tublin, 24093819	3	18	3.66E-15	2.01E-06
50	Elongation factor-1 gamma, 17414578	6	Elongation Factor-1 gamma, 17527452	5	18	3.58E-25	3.58E-06
	Elongation factor-1 alpha, 10063988	4	Elongation factor-1 alpha, 21071694	3	18	3.28E-12	3.58E-06
36	Elongation factor-1 delta, 17397886	2	Elongation factor-1 delta, 24082682	4	5	2.24E-23	1.27E-09
34	*60S Ribosomal Protein L5B, 14181865*	*1*	60S Ribosomal Protein L5B, 14181865	1 & 1	5	1.12E-09	3.36E-08
30	40S Ribosomal protein, 14185581	1	40S Ribosomal protein S3, 24085095	2	6	2.54E-08	4.33E-06
28	Elongation factor 1-beta, 17398022	2	Elongation factor 1-beta, 24091513	3 & 2	7	5.67E-31	2.24E-08
28	†		Elongation Factor 1-beta, 21088085	4	5	6.57E-19	5.22E-09
Total Peptide hits =		**49**		**87**			
Identifications =		**20**		**31**			

Two similar sets of purified proteins contributed to the identifications above. Peptides = no. of peptides matched to any single DB entry; Tags = no. of complete and partial tags matching any single EST sequence, X & Y (complete tags & partial tags, respectfully); Tags in Query = no. of tags submitted in query; † not in top 5 hits; Mascot hits below threshold score in top 5 are in *italics*; apparent molecular weights (MW) in KiloDaltons for corresponding gel bands.

2.1.8 The Broader Significance of MultiTag

The MT approach addresses an issue of growing prominence among the proteomics community: universal statistical evaluation of protein identifications[93,94]. It is a goal of the proteomics community to set a threshold for protein identifications in high throughput settings, so that a protein confidently identified in one laboratory will be confidently identified by a similar method in another institution. Consequently, the statistics of MT takes a step in this direction and determines the significance of sequence tag alignments in a manner that can be adopted as a universal standard evaluation of sequence tag identifications without the need for retrospective inspection. As the MT approach could be applied to the mining of genomic databases, the statistics will require alteration due to the size and nature of these searches. The independent statistics of MT lends the method to a wide application and to high throughput settings.

With further software developments it will be possible to completely automate the MT method for high throughput proteomics of organisms with unsequenced genomes or the analysis of highly modified proteins from organisms with sequenced genomes. Currently the ability to call sequence tags automatically is available, and a scripted interfaced can be written to create lists of sequence tags for spectra acquired from a complete LC-MS/MS run. A corresponding scripted interface for DB searching has been written that can produce a complete list of encoded retrieved DB entries for submission to MT for sorting and significance calculation (see section **2.1.7**).

Following developments in automation, MT will be a good complementary method to *de novo* sequence prediction based methods like MS BLAST and FASTS for sequence-similarity protein identification in high throughput settings, thus expanding the repertoire of spectra interpretation and DB mining tools in the hands of mass spectrometrists. As sequence-similarity methods develop, the proteomes of organisms with unsequenced genomes will become more amenable for characterization, contributing to the development of medicine, agriculture, and the biological sciences in general.

2.2 *Xenopus laevis* Functional Proteomics

2.2.1 *Xenopus laevis* as a Model System

Functional proteomics couples purification of multiprotein complexes, organelles, or specific macromolecular structures with identification of protein components by MS, and provides an effective methodology for the elucidation of the molecular architecture of cells[95,96]. An important model organism in vertebrate biology that hasn't been amenable for functional proteomics is the African clawed frog *Xenopus laevis*. Research carried out on *Xenopus* oocytes and egg extracts has produced insights into the cell cycle[97], microtubule cytoskeleton regulation by associated proteins[98], and spindle formation[99, 100]. Despite the importance of *Xenopus laevis*, its large 3070-megabase pseudotetraploid genome[101] remains unsequenced, and the genome of *Xenopus tropicalis* is planned to be sequenced by the DOE Joint Genome Institute by 2005, which significantly limits the rate at which isolated proteins can be identified. Currently, sequences of less than 7000 *Xenopus* proteins are present in a publically available DB despite a public initiative in EST sequencing (less than 221,000 largely unannotated ESTs are available, August 16, 2002, both figures from http://www.ncbi.nlm.nih.gov/). Taken together current DB resources don't provide an adequate coverage of *Xenopus'* large 3070-megabase pseudotetraploid genome[101].

Below, alternative MS data interpretation approaches (see **2.5.1**) and specialized DB searching softwares were applied to characterize the *Xenopus* microtubule-associated proteome.

2.2.2 Mass Spectrometry Analysis of Microtubule-Associated Proteins

Microtubule-associated proteins (MAPs) from *Xenopus laevis* egg extracts were isolated through binding and subsequent elution from microtubules using ATP or salt. The eluted proteins were resolved by one-dimensional gel electrophoresis and identified by MS. A first screen was conducted for proteins with a high degree of sequence similarity compared to available DB sequences using PMF, and proteins not identified by PMF were subjected to MS/MS analysis. The conventional MS/MS spectra analysis and DB searching software compares lists of observed fragment masses with predicted fragment ion masses to identify peptides that are highly similar to DB entries[102,103]. If a close homologue of the analyzed protein is not in a DB, conventional protein identification methods fail. In these cases, sequence-similarity approaches were used to identify homologous proteins beyond the limits of the conventional software.

From three gel lanes, 55 protein bands were analyzed and 61 proteins were identified by MS (Figures 8 & 11). The conventional software identified 20 proteins by PMF and 19 proteins from MS/MS spectra. From the same set of MS/MS spectra, sequence-similarity DB searching by MS BLAST identified 24 proteins and MT identified 41 proteins (this included all of the proteins identified by the conventional software and MS BLAST but one, plus 17 more) (Table 6). The presence of three known microtubule associated proteins was confirmed by Western blot (Figure 9).

In this screen, unknown *Xenopus* proteins were analyzed by NanoESI-MS/MS (Figure 10) and the resulting tandem mass spectra were used to generate peptide sequences for DB searching. In one example, upon DB searching a protein was simultaneously aligned to *Bos taurus, Mus musculus*, *Rattus norvegicus*, *Homo sapiens, Saccharomyces cerevisiae*, and *Arabidopsis thaliana* DB entries, demonstrating that some proteins are widely phylogenetically conserved and that DB entries from many distantly-related species can be used for protein identification (Table 4). To identify all of the proteins in the screen above, proteins were identified using *Xenopus* DB entries, making 29 determinations, and 32 proteins were identified by cross-species reference to homologous sequences from *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Sus scrofa*, *Gallus gallus, Gillichthys mirabilis*, *Paralichthys olivaceus*, *Salmo salar*, *Danio rerio*, *Paracentrotus lividus*, *Caenorhabditis elegans*, *Drosophila auraria,* and *Thermosynechococcus elongatus*.

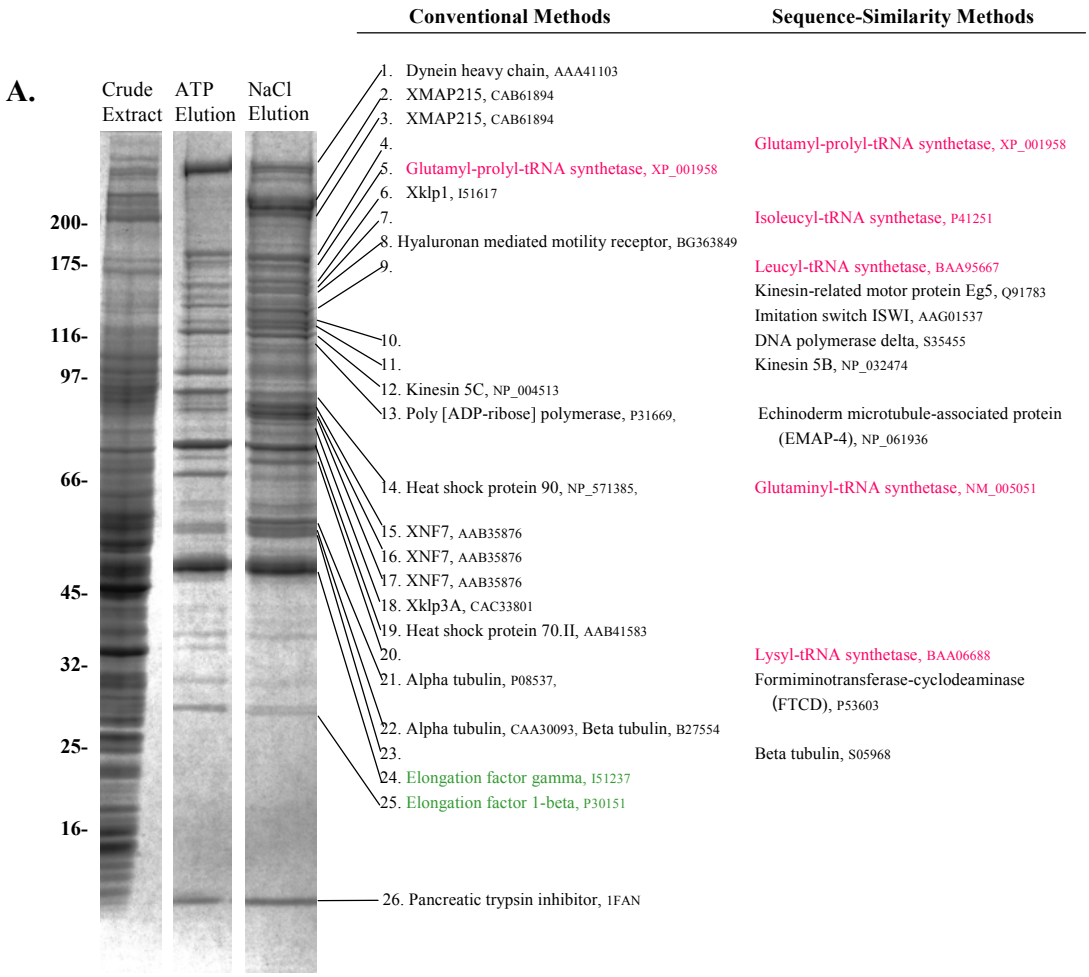

Figure 8A Identification of *Xenopus* MAPs.

The analysis of protein bands by in-gel digestion and mass spectrometry identified proteins eluted from microtubules by NaCl (A) and ATP (B). Conventional protein identification methods include: peptide mass fingerprinting, and conventional MS/MS spectra interpretation software. Sequence-similarity identification methods include: MS BLAST and MT. All proteins in the first column that were identified by MS/MS were identified by conventional and sequence-similarity methods. However only sequence-similarity methods identified proteins in the second column. Components: ARS complex (red), EF-1 complex (Green). Molecular mass markers are in kilo Daltons. *(in collaboration with Dr. Andrei Popov).*

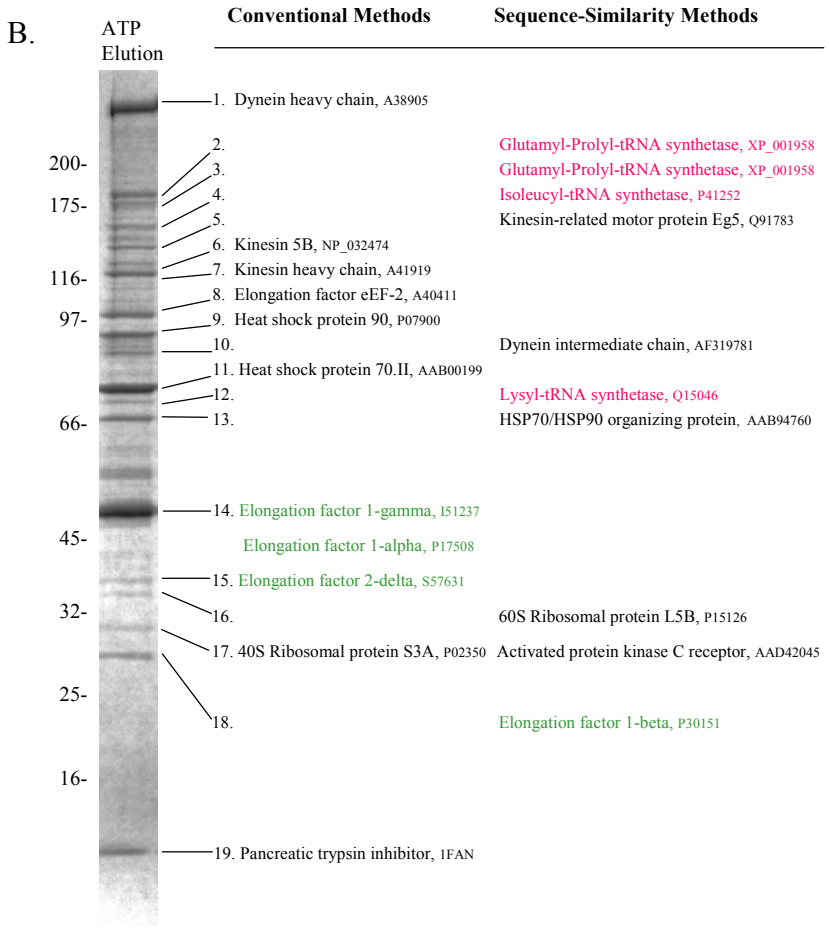

Figure 8B Identification of *Xenopus* MAPs.

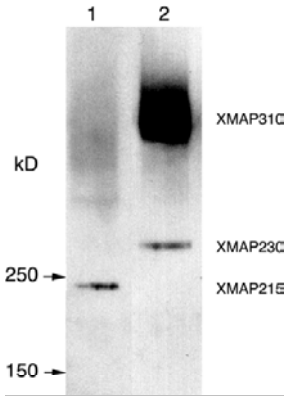

Figure 9 Immunoblot of *Xenopus* NaCl elution fractions.

NaCl-eluted proteins were resolved on a 6% polyacrylamide gel and blotted onto a nitrocellulose membrane. Blot was probed with antibodies for XMAP215 (lane 1), XMAP230 and XMAP310 (lane2). Positions of the molecular weight markers are on the left hand side, in kilo Daltons. *(in collaboration with Dr. Andrei Popov)*

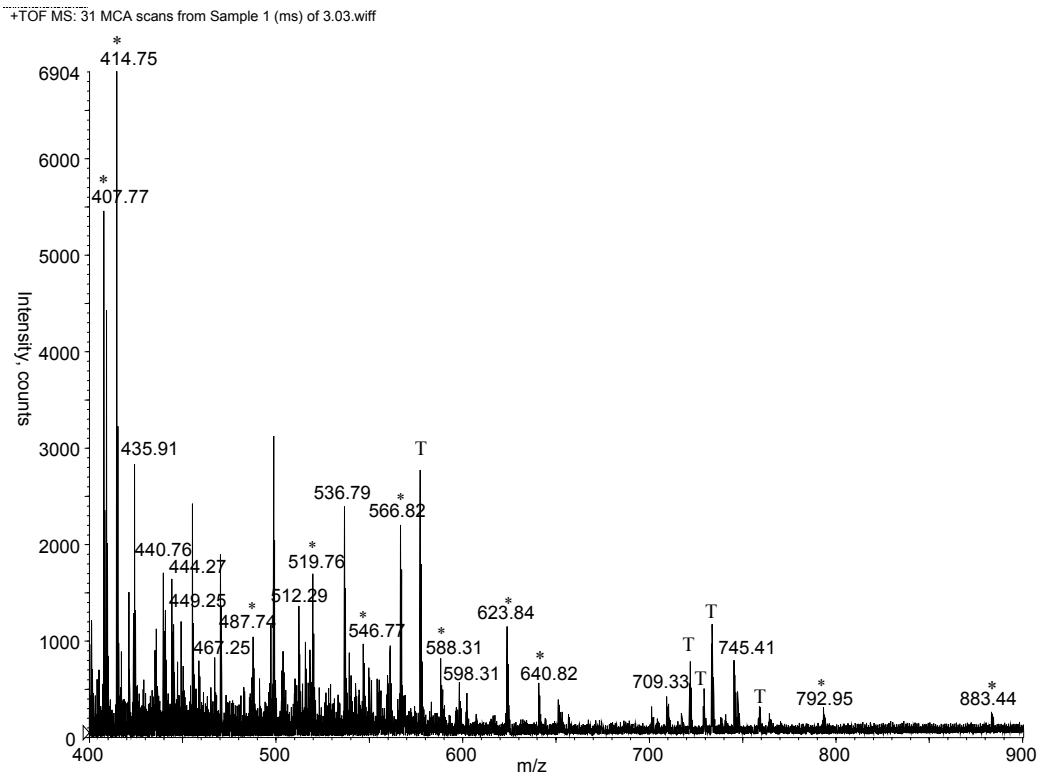

Figure 10 Time-of-Flight mass spectrum of an *in-gel* tryptic digest of a 120 kDa *Xenopus* protein.

Tandem mass spectra were acquired from peaks designated with *m/z*. Peaks originating from trypsin autolysis products are designated with T. Peaks of the peptides, which matched the sequence of bovine DNA polymerase delta are labeled with asterisks.

Table 4 Identification of a *Xenopus* Protein by MS BLAST Sequence-Similarity Searching.

M/z	z	Precursor Mass	Manual Interpretation	AUTO	BOVINE	*MOUSE*	*RAT*	*HUMAN*	*YEAST*	*ARABIDOPSIS*	Score
407.77	2	813.52	*BVVSZLLR*	/	+	+	+	+			47
414.75	2	827.49	*BLLEZGLR*	/	+	+	+	+			46
435.91	3	1304.70	BXXXTAVLZD	YES			+				
440.76	2	879.50	BLAVYD	YES							-
444.27	3	1329.79	BXXAHFNTAVLK, BXXXAHFNTAVLK	YES							-
449.25	3	1344.73	BXXXXADLL, BXXXXXADLL	NONE							-
467.25	2	932.49	BTPTPT	NONE							-
487.74	2	973.46	**BYTLDDGY**K	YES	+	+	+	+			42
512.29	2	1022.57	BVSTFPG	YES							-
519.76	2	1037.52	**BTPTGD**ZV	YES	+	+	+	+			41
536.79	2	1071.56	BLALZDPFLR	YES							-
546.77	2	1091.52	BLQDLSDFZK	YES							-
566.82	2	1131.62	*BLFEPLL*	YES	+	+	+	+			51
588.31	2	1174.60	*BVLSFDLE*	NONE	+	+		+		+	53
598.31	2	1194.62	BYGLNPEDFLK	YES		+	+				
623.84	2	1245.66	BXXSZ**LSALEE**K	YES	+	+	+	+			54
640.82	2	1279.65	**BVLSFDLE**E	/	+	+		+			54
709.33	2	1416.65	BXXEVPDZ	NONE							-
745.41	2	1488.80	BXXXTVAEA, BXXXXTVAEA	YES							-
792.95	2	1583.89	BXADSVYGFT, **BXXADSVYGFT**	YES	+	+	+	+	+	+	58
883.44	2	1764.86	BXXXEDYTZTVLE, BXXX**EDYTGATVLE**	YES	+	+	+	+	+		72

Fragmentation of peptide ions in Figure 10 enabled the production of the query above. Bovine DNA Polymerase delta was the top hit, scores of individual HSPs are presented. Sequence stretches in bold and underlined matched the bovine sequences exactly. Spectra that the software was unable to automatically predict sequences for are labeled "NONE." Spectra with automatically predicted sequences are labeled "YES". In spectra labeled "/," no automatically predicted sequences were included because high quality sequences were retrieved directly from y-ion series in the spectrum.

2.2.3 Proteins Identified in Xenopus MAP Screen

The identified proteins can be grouped in three classes: 1) previously described MAPs and motors, 2) proteins reported to be associated with the microtubule cytoskeleton, but without a known cytoskeletal function (Heat shock proteins), and 3) proteins not previously described as having microtubule localization (Table 5). In the first and second groups, several known kinesins as well as dynein heavy and intermediate chains were identified, and four previously characterized MAPs. Among the proteins of the third group components of two multiprotein complexes were detected. These are: four subunits of the 750-kD guanine nucleotide exchange EF-1βγδ complex[104], and seven aminoacyl-tRNA synthetases known to form a multicomponent complex thought to exist in all higher eukaryotes[67]. The aminoacyl-tRNA synthetase (ARS) complex has been shown to consist of eight to nine aminoacyl-tRNA synthetases and three non-synthetase components. The ARS complex is essential for aminoacylation of tRNAs prior to polypeptide synthesis

(rev. in [67]) and the EF-1 complex exchanges GTP/GDP in the binding and transportation of aminoacyl-tRNAs to the ribosome[104].

Table 5 Proteins Identified in the Microtubule-Bound Fractions

(1) MAPs and Motor Proteins

	Protein	Localization/function	ID type
1	Dynein heavy chain	ATPase domain-containing chain of the dynein complex[105]	Cr
2	XMAP310	Figure 9[106]	AB
3	XMAP230	Figure 9[107]	AB
4	XMAP215	Microtubule-associated protein, regulation of microtubule dynamics[84]	*X.l.*,AB
5	Xklp1	Chromokinesin[108]	*X.l.*
6	Eg5	Plus-end-directed microtubule motor[109],[110]	*X.l.*
7	Kinesin 5B	Kinesin heavy chain member 5B[111]	Cr
8	Kinesin 5C	Neuron-specific kinesin heavy chain member 5C[112]	Cr
9	EMAP4	a WD repeat protein, localizes to microtubules and promotes microtubule dynamics[113]	Cr
10	Xklp3	Kinesin II motor protein, Figure 12	*X.l.*
11	Xklp3A	Kinesin II motor protein[114]	*X.l.*
12	Dynein intermediate chain	Part of the dynein minus-end motor complex[115]	*X.l.*
13	Alfa tubulin	Part of the alpha-beta tubulin dimer[116]	Cr, *X.l.*
14	Beta tubulin	Part of the alpha-beta tubulin dimer[117]	Cr, *X.l.*

(2) Proteins with previously described microtubule cytoskeleton localization

1	RHAMM, (Hyaluronan mediated motility receptor)	RHAMM was reported to be associated with microtubuless in interphase and mitotic cells as well as with microtubules in vitro[118]	*X.l.*
2	ISWI (imitation switch protein)	ATP-dependent chromatin-remodeling factor[119, 120]	*X.l.*
3	Poly (ADP-ribose) polymerase (PARP)	Telomeres, mitotic centrosomes[121]	*X.l.*
4	Heat shock protein 90	Microtubules, centrosome[122, 123]	Cr
5	Heat shock protein 70.II	Microtubules[122]	*X.l.*
6	XNF7	Xenopus nuclear factor 7, protein with function in dorsal/ventral patterning of the embryo. In mitosis localizes to mitotic spindle[124]	*X.l.*
7	FTCD	Formininotransferase cyclodeaminase, microtubule-binding Golgi protein[125]	Cr

(3) Proteins not previously described as having microtubule localization

1	Ataxia telangiectasia protein	Chromatin-binding protein[126]	*X.l.*
2	Ataxia telangiectasia protein	Chromatin-binding protein[127]	*X.l.*
3	Glutamyl-prolyl-bifunctional aminoacyl tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr
4	Isoleucyl-tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr
5	Leucyl-tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr
6	DNA-polymerase delta, catalytic subunit	Part of the three-subunit DNA polymerase delta[128]	Cr
7	eEF-2	Translation Elongation factor[129]	Cr
8	Glutaminyl-tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr
9	Arginyl-tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr
10	Lysyl-tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr

Table 5 continued

11	HSP70/HSP90 organizing protein	Stress-response protein[130]	*X.l.*
12	Aspartyl-tRNA synthetase	Part of a multicomponent aminoacyl-tRNA synthetase complex[67]	Cr
13	EF 1-gamma	Beta, delta and gamma subunits of EF1 form a guanine nucleotide exchange complex (co-localize with the endoplasmic reticulum)[104]	*X.l.*
14	EF 1-alpha	Substrate of the guanine-nucleotide exchange complex[104]	*X.l.*
15	EF 1-delta-2	Homologous to the EF-delta-1, part of the guanine-nucleotide exchange complex of elongation factor-1 (EF-1)[104]	*X.l.*
16	60S Ribosomal protein L5B	60S subunit ribosome-binding protein. Was previously described in association with the ARS complex[131]	*X.l.*
17	40S Ribosomal protein S3A	40S ribosomal subunit[132]	*X.l.*
18	Activated protein kinase C receptor (RACK1)	RACK1 is a highly conserved WD protein expressed during embryogenesis[133]	*X.l.*
19	EF 1-beta	Part of the guanine nucleotide exchange complex of EF-1[104]	*X.l.*

Identified proteins are grouped into three categories (see text). Proteins were identified (ID Type) by mass spectrometry and reference to *Xenopus laevis* database sequences (*X.l.*) or cross-species referenced to sequences other than *Xenopus* (Cr). Tubulin monomers were identified with *Xenopus* and other species entries, thus we detected both *Xenopus* and pig tubulins. Additional identifications were made by immunoblot analysis using specific antibodies (AB). *(in collaboration with Dr. Andrei Popov)*

2.2.4 Association of the ARS Complex with Microtubules

A physical interaction of the ARS complex with meiotic microtubules has not been observed previously. The ARS was examined to determine if it could be a cargo complex, associated via a motor protein to microtubules. ATP eluted proteins were fractionated on a sucrose density gradient and resolved by SDS-PAGE. A fraction was identified that included seven aminoacyl-tRNA synthetases and dynein heavy chain (Figure 11). Whereas five lanes contained this pattern, one lane contained only the ARS complex but no dynein heavy chain (data not shown).

As p50 (dynamitin) disrupts the dynein/dynactin interaction *in vivo* and *in vitro*[134], we examined whether p50 would dissociate the ARS complex from microtubules in the presence of dynein. After p50 addition, some proteins like XNF7 (Xenopus Nuclear Factor 7) disappeared from the microtubule pellet, but we still detected by MS/MS the ARSs in the bound fraction (Figure 12). The ARS complex was eluted from microtubules with excess ATP, along with other kinesin proteins whose association is known to be ATP-dependent. Furthermore, ATP added to the egg extract also prevented the ARS complex from binding to microtubules. We therefore concluded that the binding of the ARS complex to microtubules is specific and ATP-sensitive. These experiments suggest that the ARS complex is not a “classical” dynein/dynactin cargo. Since MS analysis of

other, minor protein bands present in the ARS-containing fraction did not detect any other motor proteins, we concluded that the ARS complex binds to microtubules directly.

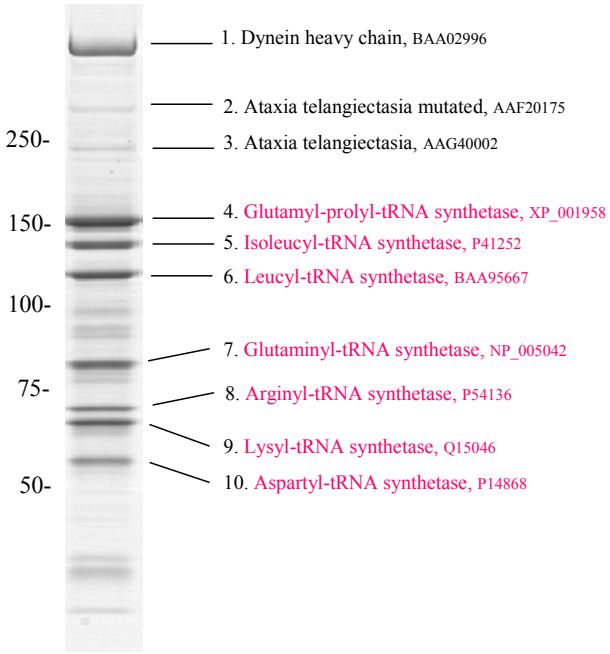

Figure 11 ***Xenopus*** **ARS complex purification.**

Microtubule-bound proteins were eluted by ATP and further fractionated on a density gradient, with one unique fraction corresponding to ca. 15S shown above. Seven aminoacyl-tRNA synthetases co-migrated on a sucrose density gradient with dynein heavy chain. Components of the ARS complex (Red). *(in collaboration with Dr. Andrei Popov)*

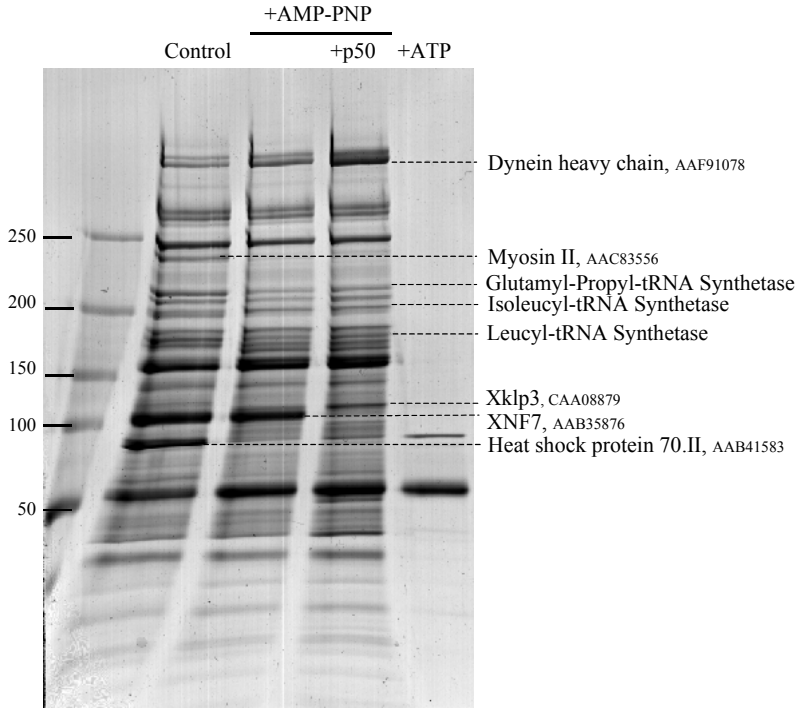

Figure 12 Isolation of the motor faction in the presence of p50.

Coomassie-Blue-stained gradient (6-20%) polyacrylamide SDS-gel showing the proteins bound to microtubules under different conditions and eluted from them with 20mM ATP. The lanes show from left to the right: 1. Molecular weight markers, in kiloDaltons; 2. Proteins bound to microtubules from native extract (control); 3. Proteins bound in the presence of AMP-PNP; 4. Proteins bound in the presence of AMP-PNP and p50; and 5. Proteins isolated from native extracts supplemented with 10 mM ATP. The major band above the 50 kD marker corresponds to alphą and beta tubulin (data not shown). *(in collaboration with Dr. Andrei Popov)*

2.2.5 *In vitro* Spindle Reconstitution and Electron Microscopy

We have demonstrated that two essential components of the protein translation machinery, EF-1 complex and the ARS complex are bound to microtubules in meiotic egg extracts. These findings suggested that protein translation may be spatially connected with the spindle. To verify this hypothesis, we assembled spindles in *Xenopus* egg extracts and analyzed their structure by electron microscopy. Remarkably, we detected ribosomes, that appeared in clusters located peripheral to the centrosomes (Figure 13a) and were distributed along the length of the spindle microtubules (Figure 13b), further suggesting that the protein translation machinery is localized on the spindle *in vivo*.

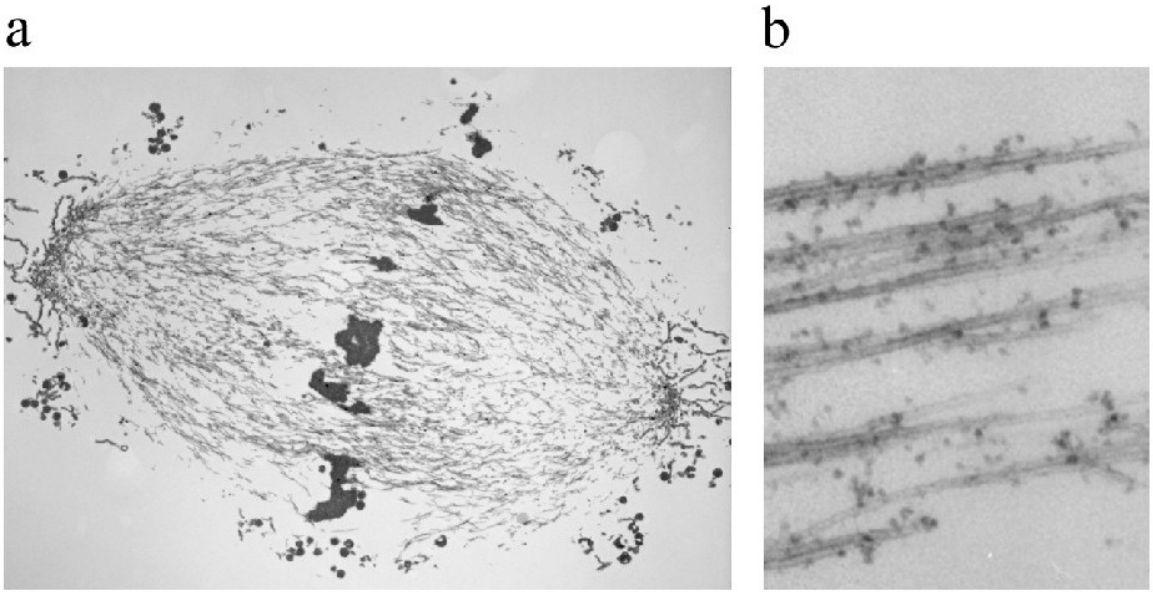

Figure 13 Electron micrographs of the *Xenopus In vitro*-reconstituted spindle.

A) Initial low magnification images show ribosomes in clusters located peripheral to the centrosomes. B) Upon higher magnification, ribosomes were found to be distributed along the length of the spindle microtubules. *(in collaboration with Dr. Peg Coughlin)*

2.2.6 Biological Implications of *Xenopus* Experiments

(in collaboration with Dr. Andrei Popov and Professor Eric Karsenti)

Due to the large size of *Xenopus* oocytes and the fact that they contain an abundance of cytoplasmic proteins for early development, these cells or extracts prepared from them have been the preferred model system for the study of the mitotic spindle and spindle-associated proteins[100]. For example, MAPs have been analyzed by purification and MS from other organisms, such as human, *Drosophila*, and yeast. However, from the sixteen microtubule-bound proteins purified by Mack and Compton in 2001 from mitotic HeLa cells, only two are on the list of proteins identified above (Eg5 and HSP70)[135]. In a systematic analysis of centrosome-associated proteins (and many MAPs are also centrosomal components, see[136]) from *Drosophila*, only one identified protein was also found in our preparation (HSP90)[123]. Of the eight proteins identified by Adams and Kilmartin from *Saccharomyces cerevisiae* spindle pole bodies, none matches those found in our preparation[137]. Comparing these studies with our results demonstrates the utility of the *Xenopus* oocyte as a model system.

The finding that the ARS complex and the EF-1 complex of the protein translation machinery are bound to microtubules in egg extracts, and electron micrographs showing ribosomes on the reconstituted spindle, prompts speculations about the potential existence of such interactions *in vivo*. The ARS and EF-1 complexes were identified primarily in the ATP-elution samples and, interestingly, there is evidence that they may interact with each other *in vitro*[138]. From our results, it could be suggested that protein translation during meiosis occurs on spindle microtubules. Indeed, ribosomes have also previously been found on microtubules isolated from sea urchin unfertilized eggs, and their attachments were mediated through another protein found in our screen (EMAP)[139]. Spatial regulation of translation could be especially important for large cells like the *Xenopus* oocyte. Although transcription of cyclin B1[140] does occur during mitosis, protein synthesis has not been directly detected in mitotic cells. On the contrary, *de novo* synthesis of several cell cycle components, including c-mos, cyclin B1, and XKID is essential during meiosis[141,142]. Furthermore, cyclin B1 mRNA was found associated with meiotic spindles and it was suggested that translation of cyclin B1 occurs locally, "on or near spindles and centrosomes"[143]. These findings lend support to this interpretation.

These findings demonstrate the power of sequence-similarity protein-identification methods combined with cell biological approaches for the functional proteomics of organisms outside the boundaries of sequenced genomes.

Table 6 *Xenopus* Protein Identifications and Statistics

No.	Peptide Mass Mapping -Mascot	MS/MS -Mascot	MS/MS -MSBLAST	MS/MS -MultiTag	BLS	BLH	MS	MT	TH	C	H	XE	EM	S	MA	E -value	PredCount
1. NaCl elution lane (Figure 8A)																	
1	Dynein heavy chain ***R. norvegicus***, AAA41103						/	/	/	/	/	/	/	/	/	/	/
2	XMAP 215kD *Xenopus*, CAB61894						/	/	/	/	/	/	/	/	/	/	/
3	XMAP 215kD, *Xenopus*, CAB61894						/	/	/	/	/	/	/	/	/	/	/
4				Glutamyl-prolyl-tRNA synthetase ***H. sapiens***, XP_001958			16	11	3	2	4	41	1	1	0.1	1.08E-05	2.84E-14
5		Bifunctional aminoacyl-tRNA Synthetase ***H. sapiens***, P07814		Glutamyl-prolyl-tRNA synthetase ***H. sapiens***, XP_001958			12	9	5	3	5	12	3	2	0.1	5.30E-08	3.09E-24
6	Xklp1, *Xenopus*, I51617						/	/	/	/	/	/	/	/	/	/	/
7	Isoleucyl-tRNA synthetase *PMF match to ATP 4*						/	/	/	/	/	/	/	/	/	/	/
8				Hyaluronan mediated motility Receptor, *Xenopus*(EST)BG363849			12	9	0	0	0	35	0	3	0.1	/	/
9				ISWI, *Xenopus*, AAG01537			14	12	2	2	7	29	5	2	0.1	1.72E-03	2.83E-08
9				Eg5 , *Xenopus*, Q91783			14	12	2	1	1	29	13	2	0.1	1.26E-02	1.94E-05
9				Leucyl-tRNA synthetase*			14	12	4*	0	4	29	0	0	/	/	/
10			DNA polymerase delta ***H. sapiens***, P28340	DNA polymerase delta ***H. sapiens***, S35455	145	3	12	12	5	2	13	31	1	1	0.1	8.14E-07	1.26E-16
11			Kinesin heavy chain *Xenopus*, AJ249840	Kinesin 5B ***M. musculus***, NP_032474	139	3	12	9	4	3	33	11	5	5	0.1	5.05E-08	6.94E-18
12		Kinesin 5C ***M. musculus***, AAC79804	Kinesin heavy chain ***M. musculus***, L27153	Kinesin 5C ***H. sapiens***, NP_004513	387	9	15	10	5	1	19	5	3	4	0.1	3.40E-08	3.54E-17
13		Poly [ADP-ribose] polymerase *Xenopus*, P31735		Poly [ADP-ribose] polymerase *Xenopus*, P31669			14	11	4	4	2	88	5	2	0.1	1.00E-04	9.95E-17
13			EMAP ***H. sapiens***, Q9HC35	EMAP, ***H. sapiens***, NP_061936	134	2	14	12	2	1	0	88	1	1	0.1	4.57	8.56E-03
14		Heat shock-like protein ***M. musculus***. CAA34748	Heat shock protein 90-beta ***S. salar***, AF135117	Heat shock protein 90-beta ***D. rerio***, NP_571385	184	4	14	5	5	3	38	190	35	3	0.1	4.35E-09	3.87E-19
14			Glutaminyl-tRNA synthetase ***M. musculus***, AK003794	Glutaminyl-tRNA synthetase ***H. sapiens***, NM_005051	104	2	14	8	5	2	7	190	14	8	0.1	6.26E-08	7.95E-16
15	XNF7 *Xenopus*, AAB35876						/	/	/	/	/	/	/	/	/	/	/
16	XNF7 *Xenopus*, AAB35876						/	/	/	/	/	/	/	/	/	/	/
17	XNF7 *Xenopus*, AAB35877						/	/	/	/	/	/	/	/	/	/	/
18	Xklp3A *Xenopus*, CAC33801						/	/	/	/	/	/	/	/	/	/	/
19	Heat shock protein 70 *Xenopus*, AAB41583						/	/	/	/	/	/	/	/	/	/	/
20				Lysyl-tRNA synthetase ***H. sapiens***, BAA06688			15	9	5	4	10	71	21	4	0.1	1.07E-07	3.80E-24
21			Formiminotransferase cyclodeaminase ***H. sapiens***, AF289023	Formiminotransferase Cyclodeaminase ***S. scrofa***, P53603	169	3	16	12	2	0	7	172	8	4	0.1	2.18E-01	3.49E-04
21		Tubulin alpha ***G. mirabilis***. AAL24509		Tubulin alpha *Xenopus*, P08537			16	4	3	3	179	172	118	3	0.1	7.11E-10	1.02E-14
22			Tubulin beta ***H. sapiens***, BC020171	Tubulin beta-5 ***G. gallus***, B27554	350	8	18	10	4	3	289	251	25	2	0.1	2.19E-07	3.19E-17
22		Tubulin alpha 1 ***H. sapiens***, AAH06468	Tubulin alpha ***P. lividus***, A60671	Tubulin alpha *Xenopus*, CAA30093	267	5	18	7	6	6	268	251	137	7	0.1	2.72E-08	3.39E-33
23	Tubulin beta-2 *Xenopus*, S05968						/	/	/	/	/	/	/	/	/	/	/
24	Elongation factor gamma *Xenopus*, I51237						/	/	/	/	/	/	/	/	/	/	/
25		Elongation factor 1-beta *Xenopus*, CAA49418	Elongation factor 1-beta *Xenopus*, P30151	Elongation factor 1-beta *Xenopus*, P30151	250	5	13	7	7	4	3	26	26	1	0.1	2.12E-08	1.06E-32
26		Pancreatic trypsin inhibitor *B. taurus*, P00974	Pancreatic trypsin inhibitor *B. taurus*, X04666	Pancreatic trypsin inhibitor ***B. taurus***, 1FAN	160	3	10	6	3	3	43	36	0	0	0.1	1.07E-08	2.00E-16
2. ATP elution lane (Figure 8B)																	
1	Dynein heavy chain ***R. norvegicus***, A38905						/	/	/	/	/	/	/	/	/	/	/
2				Glutamyl-prolyl-tRNA synthetase ***H. sapiens***, XP_001958			10	9	3	1	7	21	15	4	0.1	7.14E-07	1.09E-09
3				Glutamyl-prolyl-tRNA synthetase ***H. sapiens***, XP_001958			8	4	1	1	0	4	1	1	0.1	15.35	4.45
4				Isoleucyl-tRNA synthetase ***H. sapiens***, P41252			10	9	3	1	4	22	0	0	0.1	2.73E-05	2.89E-08
5				Eg5, *Xenopus*, Q91783			10	7	1	1	0	6	1	1	1	19.79	0.75
6		Kinesin heavy chain, ***M. musculus***, AAC06326	Kinesin heavy chain ***M. musculus***, X61435	Kinesin 5B ***M. musculus***, NP_032474	109	2	15	8	3	3	15	3	1	3	1.5	1.37E-05	1.01E-10
7	Kinesin heavy chain ***H. sapiens***, A41919						/	/	/	/	/	/	/	/	/	/	/
8		Elongation factor eEF-2 ***R. norvegicus***, CAA68805		Elongation factor eEF-2 ***C. elegans***, A40411			10	6	4	3	65	50	50	2	0.1	7.99E-07	1.02E-13
9	Heat shock protein 90 ***H. sapiens***, P07900						/	/	/	/	/	/	/	/	/	/	/
10			Cytoplasmic dynein intermediate chain *Xenopus*, AF319781	Cytoplasmic dynein intermediate chain *Xenopus*, AF319781	179	3	18	9	5	5	7	63	17	5	0.1	6.25E-28	1.80E-07
11		Heat shock protein 70.II *Xenopus*, AAB00199	Heat shock protein 68 ***D. auraria***, AF247553	Heat shock protein 70 ***P. olivaceus***, AAC33859	106	2	20	4	3	2	31	167	167	2	0.1	4.09E-09	5.78E-10
12				Lysyl-tRNA synthetase ***H. sapiens***, Q15046			15	4	1	1	6	81	25	2	0.1	1.56	6.58E-02
13			HSP70/HSP90 organizing protein ***C. griseus***, AAB94760	HSP70/HSP90 organizing protein ***C. griseus***, AAB94760	168	3	16	7	4	0	6	25	13	2	0.1	7.82E-09	6.43E-14
14		Elongation factor 1-gamma *Xenopus*, AAB29957	Elongation factor 1-gamma *Xenopus*, AAB29958	Elongation factor 1-gamma *Xenopus*, I51237	168	7	22	8	8	6	3	100	100	6	0.1	4.12E-08	5.38E-41
14		Elongation factor 1-alpha *Xenopus*, CAA37169	Elongation factor 1-alpha *Xenopus*, P17507	Elongation factor 1-alpha *Xenopus*, P17508	379	8	22	11	5	5	4	95	95	5	0.2	8.35E-07	3.03E-22
15		Elongation factor delta-2 *Xenopus*, S57631	Elongation factor 1-beta/delta ***C. elegans***, P34460	Elongation factor delta-2 *Xenopus*, S57631	106	2	9	5	4	3	2	46	46	3	0.1	2.49E-09	3.53E-18
16				60S Ribosomal protein L5B			6	5	2	1	17	66	39	2	0.1	2.21E-07	5.36E-09

Table 6 continued

				Xenopus, P15126														
17		Ribosomal protein S3 ***M. musculus***, AAH10721	40S ribosomal protein S3B *Xenopus*, P47835	40S Ribosomal protein S3A *Xenopus*, P02350	98	1	18	7	2	2	21	120	20	2	0.1	9.86E-05	1.24E-07
17		Activated protein kinase C receptor *Xenopus*, AAD42045		Activated protein kinase C receptor *Xenopus*, AAD42045			18	7	2	1	1	120	100	2	0.1	4.98E-03	5.60E-05
18			Elongation factor 1-beta *Xenopus*, P30151	Elongation factor 1-beta *Xenopus*, P30151	266	5	7	5	5	4	18	70	70	5	0.3	2.75E-09	1.16E-20
19		Pancreatic trypsin inhibitor *B. taurus*, P00974		Pancreatic trypsin inhibitor ***B. taurus***, 1FAN			10	4	2	2	40	1	0	0	0.1	3.82E-07	3.64E-08
3. Fractionation lane (Figure 11)																		
1	Dynein heavy chain ***R. norvegicus***, BAA02996						/	/	/	/	/	/	/	/	/	/	/	
2				Ataxia telangiectasia *Xenopus*, AAF20175			5	5	2	2	0	0	0	0	0.1	8.88E-11	4.07E-07
3			Ataxia telangiectasia *Xenopus*, AAG40002		150	3	/	/	/	/	/	/	/	/	/	/	/	
4	Glutamyl-prolyl-tRNA Synthetase *PMF match to ATP 2*			Glutamyl-prolyl-tRNA synthetase *Xenopus*, EST, BI443016			/	/	/	/	/	40	14	3	/	/	/	
5	Isoleucyl-tRNA synthetase *PMF match to ATP 4*						/	/	/	/	/	/	/	/	/	/	/	
6		Leucyl-tRNA synthetase ***H. sapiens***, BAA95667	Leucyl-tRNA synthetase ***H. sapiens***, BAA95667	Leucyl-tRNA synthetase ***H. sapiens***, BAA92590	416	8	14	7	3	1	2	/	/	/	/	7.73E-05	1.50E-07
7	Glutaminyl-tRNA synthetase *PMF match to NaCl 14*						/	/	/	/	/	/	/	/	/	/	/	
8			Arginyl-tRNA synthetase ***H. sapiens***, P54136	Arginyl-tRNA synthetase ***T. elongatus***, NP_681615	439	9	12	7	2	1	17	/	/	/	/	3.42E-05	4.95E-03	
9	Lysyl-tRNA synthetase *PMF match to NaCl 20*						/	/	/	/	/	/	/	/	/	/	/	
10		Aspartyl-tRNA synthetase ***H. sapiens***, AAH00629	Aspartyl-tRNA synthetase ***H. sapiens***, P14868	Aspartyl-tRNA synthetase ***H. sapiens***, P14868	361	6	10	5	5	4	2	/	/	/	/	2.98E-21	3.55E-09	

Protein bands were analyzed by peptide mass fingerprinting and Mascot database searching (Blue) from the NaCl elution lane (section 1), ATP elution lane (section 2), and Density Gradient fractionation lane (section 3). One set of tandem mass spectra was analyzed by Mascot (Orange), MS BLAST (Red), and MultiTag (Green). BLS: MS BLAST score, BLH: no. of high scoring pairs in MS BLAST identification, MS: no. of MS/MS acquired where sequence tag interpretation was attempted, MT: no. of complete tags submitted to MT, TH: no. of tags in top MT hit, C: no. of complete tags in top hit, H: hits that are homologues to the top MT hit with E-values <0.1, including the top hit, XE: no. of *Xenopus* ESTs retrieved from non-error-tolerant searches with all sequence tags from the analysis of the digest, EM: no. of *Xenopus* ESTs matching the top MT hit, S: no. of sequence tags matching *Xenopus* ESTs matching the top MT hit, MA: mass accuracy in Daltons used in MT evaluation, E: E-values calculated by MT. PredCount: Predicted Count calculated by MT. *Each individual sequence tag matched to a different database entry from four different species. Identifications in Black were identified by EST DB searching.

2.3 *Dunaliella salina* Functional Proteomics

2.3.1 Plant Proteomics

The characterization of proteomes is a precise method to identify the proteins, and their corresponding genes, which act together to produce the unique biochemistry and physiology of cells. The proteomes of plants have been characterized in multiple species, such as *Arabidopsis*[144], rice[145], maize[77], pea[79], wheat[146], and poppy[81], among others. The sequencing of the *Arabidopsis*[11] and rice genomes[9,10] has facilitated proteomic research into these organisms. Genome sequences and corresponding protein sequence databases provide a reference for the identification of proteins by the correlation of analyzed peptide fragments with *in silico* sequences by MS and DB searching. Conventional protein-identification algorithms, such as Mascot[102] and SEQUEST[147], correlate mass data from the mass spectra of protein digests (PMF) or fragment ion tandem mass spectra of peptides (produced by the proteolytic digestion of whole proteins), and are primarily capable of exact matching (reviewed in[16]), subsequently restricting proteome characterization in many plant species with unsequenced genomes (reviewed in[148]). Despite this limitation, homologous proteins in different species often have conserved amino acid sequences, enabling existing DB entries to serve as a reference for the identification of homologous proteins in other phylogenetically related species.

Here MS and multiple DB searching strategies were applied simultaneously (reviewed in section **2.5.3**) to characterize the proteome of the green alga *Dunaliella salina. Dunaliella* can adapt to the most hypersaline conditions on earth. As such it is recognized as a model photosynthetic organism for analyzing salinity tolerance. In contrast, most plants can adapt to low or moderate salinities and their growth is severely limited at salinities exceeding 200 mM NaCl[149].

Differential mRNA screens carried out in *Arabidopsis thaliana* and rice have shown that plants respond to salt stress by up-regulation of expression of a large number of genes involved in diverse physiological functions[150-153]. *Dunaliella* responds to salt stress by massive accumulation of glycerol, the internal osmotic element in *Dunaliella*, by enhanced elimination of Na^+ ions and by accumulation of distinct proteins[154]. However, no comprehensive analysis of salt up-regulated genes/proteins has been carried out in *Dunaliella*, mainly due to absence of sufficient genomic information. Sequence information for *D. salina* is limited to approximately 50 protein entries and 3000 nucleotide entries in NCBI (August, 2003), thus restricting the ability to identify proteins by MS using conventional methods.

The conventional DB-searching algorithm, Mascot, and two sequence-similarity DB-searching algorithms, MS BLAST and MultiTag, and protein and EST DB searching, were applied for the identification of proteins from 2D gels. In an attempt to characterize *Dunalella's* unique physiology resulting in resistance to high saline conditions, we identified 61 proteins that are up-regulated in 3M salt, from three sub-cellular fractions: crude plasma membrane, chloroplast soluble proteins, and cytosol. The induced proteins included many members of the Calvin cycle, starch biosynthesis and degradation, amino acid biosynthesis, energy production, chaperones, and protein synthesis and degradation. Sequence-similarity protein identification techniques were essential for effective identification of more than half of the proteins analyzed. From these results, we expect the proteomics of many plants with unsequenced genomes to be more amenable to characterization than previously facilitated by conventional methods.

2.3.2 Analysis of Dunaliella Proteins by Mass Spectrometry

D. salina cells were cultured in 0.5M or in 3M NaCl, and fractionated into crude plasma membrane, cytoplasmic soluble, and chloroplast soluble fractions. After separation of each fraction on 2-D gel electrophoresis the spots were differentially analyzed to identify components up-regulated by high salt (Figure 14A, B, C). In general, the fractionation elevated the overall number of salt up-regulated spots (>2-fold induction) from 30 in a total cell protein extract to 75 (+ 3 references) in the combined 3 fractions after exclusion of cross-contaminations between fractions (data not shown).

Initial peptide mass fingerprinting analysis of spots up regulated in the total cell extract failed to identify any protein (data not shown). From the three cellular fractions, PMF was able to identify 9 of 78 spots using reference protein DB sequences from *Dunaliella sp.* and *Chlamydomonas sp.*, among others (Table 7). The remaining 69 proteins from the three 2-D gels were analyzed by nanoelectrospray MS/MS and Mascot protein DB searching, which identified 23 additional proteins (in one case two proteins were identified in one spot).

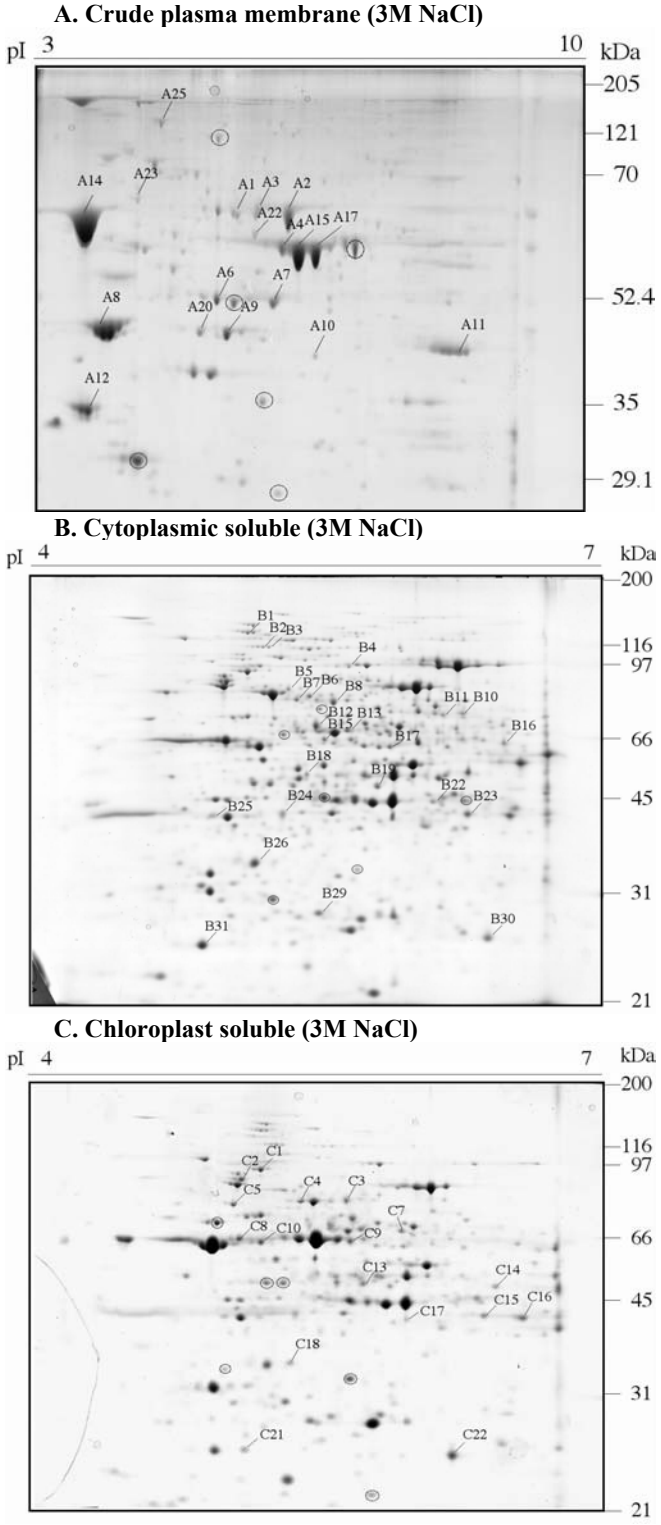

Figure 14 *Dunaliella* 2-D gel protein separation

Proteins were first resolved by isoelectric point from pH of 3-10 or 4-7, then according to mass, ~21-200kiloDaltons. Gels were stained with Coomassie. *(in collaboration with Dr. Adriana Katz)*

Table 7 *Dunaliella* Protein Identification

Spot	Protein Identifications	F.I.	OMW	TMW	Species	Accession	PF	Ma-EST/Ac	Ma	MB	MT
	Antioxidation										
B30	Iron-superoxide dismutase pre.	3	27	27	*V. unguiculata*	AAF28773				1	
B31	Thioredoxin peroxidase	2	26	22	*T. elongatus*	BAC09006			2	3	
C21	Thioredoxin peroxidase	3	25	23	*R. conorii*	AAL02989				1	
	Chaperones										
A22	Chaperonin precursor	3	70	63	*P. sativum*	AAA66365		2 / AV397884	4	8	
A24	Heat shock protein 70	#	100	70	*A. albimanus*	AAC41543		2 / BI874228	4	10	
C2	Luminal binding protein 5 pre.	4	98	74	*N. tabacum*	CAA42660		2 / AV643368	3	7	
C4	Heat shock 70 put. mit. pre.	3	87	70	*O. sativa*	AAO17017				2	
	Calvin Cycle / Carbon Acquisition										
A8	Carbonic anhydrase	#	47	64	*D. salina*	AAC49378				4	
A12	Carbonic anhydrase	#	35	64	*D. salina*	AAC49378				3	
A14	Carbonic anhydrase	#	60	64	*D. salina*	AAC49378	X				
A17	Rubisco pre.	#	55	53	*C. moewusii*	AAA84152	X				
A4	Rubisco large sub.	1	55	53	*C. reinhardtii*	AAA84449		3 / BI726309	6	9	
A15	Rubisco large sub.	1	55	13	bacteria OTI-8	BAA92486	X				
B18	Rubisco large sub.	4	60	52	*H. capensis*	AAK96893	X				
B13	Rubisco activase chl. pre.	2	75	45	*C. reinhardtii*	AAA33091			2	4	
C7	Rubisco activase chl. pre.	2	75	45	*C. reinhardtii*	AAA33091				4	
A9	Phosphoribulokinase pre.	1	45	46	*S. oleracea*	AAA34036		1 / AV392278		6	
A20	Phosphoribulokinase	3	45	42	*C. reinhardtii*	AAA33090		1 / BE453265	1	3	
A11	NADP-Glyceraldehyde-3-phosphate DH	6	42	40	*Chlamydo. sp.*	AB035312				8	
B25	Sedoheptulose-1,7-bisphosphatase	4	42	39	*Chlamydo. sp.*	BAA94305				6	
B8	Dihydroxyacetone kinase	6	76	65	*S. pombe*	AF059204				4	
C3	Dihydroxyacetone/glycerone kinase-like	4	88	64	*A. thaliana*	BAB02871				8	
B29	Triose phosphate isomerase	2	29	23	*B. belcheri*	BAA22631				4	
	Starch Biosynthesis / Pentose Phosphate Pathway										
B10	Phosphoglucomutase chl. pre.		73	69	*S. tuberosum*	AJ240053		2 / BE128973	2	10	
B17	6-Phosphogluconate DH decarboxylating	1	64	54	*M. sativa*	AAB41553		2 / BF269268	4	8	
B15	ADP-Glucose pyrophosphorylase small sub.	2	66	55	*C. reinhardtii*	AAF75832			5	10	
B16	ADP-Glucose pyrophosphorylase large sub.	3	65	57	*L. esculentum*	AAC49943				5	
B26	Inorganic pyrophosphatase pre.	2	34	31	*C. reinhardtii*	CAC42762		1 / BM498985	1	8	
C18	Inorganic pyrophosphatase pre.	3	39	31	*C. reinhardtii*	CAC42762				6	
	Energy										
A1	Glucose-6-phosphate 1-DH	6	60	66	*D. bioculata*	CAB52685	X				
A3	Glucose-6-phosphate 1-DH	5	60	66	*D. bioculata*	CAB52685	X				
C13	Plastidic NADP-dependent malate DH	2	58	47	*D. bioculata*	CAC15546			2		3
B19	Plastidic NADP-dependent malate DH	4	50	47	*D. bioculata*	CAC15546	X				
A2	Dihydrolipoamide S-acetyltransferase	4	60	44	*T. elongatus*	BAC08851				8	
A6	Pyruvate DH E1 alpha sub.	4	50	47	*A. thaliana*	AAB86803				3	
B24	Thiamin biosynthetic enzyme	2	41	37	*G. max*	BAA88227					1
C17	Adenosine kinase	3	48	38	*P. patens*	CAA75628		1 / BJ172248	1	1	
B23	Ferrodoxin NADP oxidoreductase put.	3	42	43	*A. thaliana*	AAF19753				8	
C15	Ferrodoxin NADP oxidoreductase put.	2	49	43	*A. thaliana*	AAF19753		2 / BG647868	1	4	
C16	Ferrodoxin NADP oxidoreductase put.	3	49	43	*A. thaliana*	AAM65564		1 / BG647868	1	3	
C5	ATP synthase beta chain mit. pre.	2	86	62	*C. reinhardtii*	CAA43808		1 / BE642669	1	4	
C22	ATP synthase delta chain chl. pre.	4	24	24	*C. reinhardtii*	AAB51365				3	

Table 7 Continued

	Pyrimidine and Amino Acid Biosynthesis									
A7	Glutamine synthetase	#	50	42	*C. reinhardtii*	AAB01817		5	7
A20	Glutamine synthetase	3	45	41	*C. reinhardtii*	AAB01818	2 / AV623601	2	3
B1	Carbamoyl phosphate synthetase large chain	3	127	130	*A. thaliana*	AAB67843		3	11
B4	Aspartate kinase-homoserine DH put.	3	95	100	*A. thaliana*	BAC43372			3
B11	2-Isopropylmalate synthase put.	4	72	74	*A. thaliana*	AAF26002	2 / AV631505	2	8
B12	D-3-Phosphoglycerate dehydrogenase	3	70	66	*A. thaliana*	BAA20405	2 / BF269268		4
	Protein Biosynthesis and Degradation									
B2	Zinc metalloprotease	3	112	118	*A. thaliana*	BAB02957	1 / BE249333	2	5
B3	Zinc metalloprotease	3	112	118	*A. thaliana*	BAB02957	2 / AV628512		4
A10	TGF-beta receptor interacting homolog	2	42	36	*A. thaliana*	AAC49079			5
C9	Processing peptidase put. mit.	2	72	59	*A. thaliana*	AAF14827			7
A21	26S Proteasome regulatory particle triple-A	2	58	50	*O. sativa*	AB037154	4 / AV620391	4	7
C14	Translation elongation factor Tu mit.	5	57	44	*R. americana*	AAD11872			5
C1	Translation elongation factor EF-G	4	103	78	*G. max*	X71439	1 / BI727515	2	9
	Cytoskeleton									
C8	Beta tubulin	3	72	50	*C. incerta*	AAB60936	X			
C10	Alpha tubulin	2	71	29	*Z. mays*	S39969	X			
	Glycosilation									
B22	GDP-mannose pyrophosphorylase	1	45	40	*A. thaliana*	CAC35355			4	
	Na+ transpot									
B6	NQR alpha sub.	5	79	51	*V. cholerae*	AAF95439			1	
B7	NQR alpha sub.	4	77	51	*V. cholerae*	AAF95439			1	
	Others									
B5	GTP-binding protein typA	7	92	68	*A. thaliana*	BAB08691			6	
						PMF IDs =	***9***			
						Mascot EST IDs = 20				
						Mascot Protein IDs =		23		
						MS BLAST(s) Ids=			50	
						MS BLAST(m) Ids=				2

Columns: Spot: no. corresponding to 2-D gels; Protein identification: proposed biochemical function of the protein based on MS analysis and DB searching; FI: fold induction of the spot from changing growth conditions from 1M to 3M NaCl; OMW: observed molecular weight in kiloDaltons calculated by software X; TMW: theoretical molecular weight of the identified protein based on the aligned database sequence; Species: origin of corresponding retrieved DB sequence; Accession: retrieved DB sequence; PF: Peptide mass fingerprinting, X = positive identification; Ma: no. of peptides positively identified by Mascot by protein DB searching; MB: no. of peptides positively identified by scripted or manually interpreted MS BLAST protein DB searching; MT: no. of peptides positively identified by MultiTag protein DB searching; Ma-EST/Ac: no. of peptides positively identified by Mascot EST DB searching / accession number; put. = putative; mit. = mitochondrial; chl. = chloroplast; DH = dehydrogenase; sub. = subunit; pre. = precursor.

The complete set of tandem mass spectra from the analysis was further interpreted using the MS BLAST sequence-similarity protein identification approach. Amino acid sequences were predicted *de novo* from tandem mass spectra and assembled into modified BLAST queries for DB searching, as previously described[62]. MS BLAST identified 50 proteins, which included all of the proteins identified by Mascot (except one) plus 28 more (Table 7).

Further EST DB searching of all tandem mass spectra with Mascot confirmed the protein identifications made in 20 of the cases, using primarily *Chlamydomonas reinhardtii* sequences (Table 7). All spots still unidentified were analyzed by MS BLAST EST DB searching; however, this did not contribute to the characterization of any of the unknown spots, although the method confirmed the identification of a few proteins already identified (data not shown).

The MT approach for sequence-similarity identification was used as a final attempt to identify the remaining unidentified proteins because of its demonstrated enhanced sensitivity over MS BLAST (see section **2.2**), but this technique was only able to confirm an identification made by Mascot and subsequently missed by MS BLAST, and contributed only one new identification which relied upon only one peptide alignment. Whereas MS BLAST relies upon representing a peptide along its full length, MT is more sensitive when low abundance proteins are analyzed and full amino acid sequences can not be discerned, rendering MS BLAST ineffective[91].

2.3.3 *Dunaliella* Proteins Induced in 3M NaCl

Of the 61 identified proteins, the largest categories are enzymes involved in carbon assimilation or mobilization and in production of metabolic energy. Up regulation of four major Calvin cycle enzymes (Rubisco, Phosphoribulokinase, NADP Glyceraldehyde 3-P DH, Sedoheptulose 1,7 bisphosphatase) and of Rubisco activase, suggest that high salinity enhanced CO_2 assimilation. The large accumulation of a different form of Rubisco under high salt (spot A17) may be an adaptation response for more efficient CO_2 assimilation. The pronounced induction of plasma membrane-associated carbonic anhydrases is a typical response of algae and cyanobacteria to carbon limitation. It has been reported before in *Dunaliella* and proposed to enable the algae to overcome the limitations in CO_2 availability in hypersaline solutions[155].

Up-regulation of two ADP-pyrophosphorylase subunits suggests enhanced starch biosynthesis whereas the large accumulation of glucose 6-phosphate dehydrogenase, a key

enzyme in starch mobilization and NADPH production, suggests enhanced starch degradation and generation of Redox energy for biosynthesis of carbon metabolites.

Notable enzymes in the category of metabolic energy production, in addition to glucose 6-phosphate dehydrogenase, are 2 subunits of pyruvate dehydrogenase (dihydrolipoamide S-acyl transferase, pyruvate dehydrogenase E1 alpha subunit), the key enzyme in channeling carbon into the citric acid cycle for production of NADH and the chloroplastic malate dehydrogenase, which catalyzes translocation of reducing power between the chloroplast and the cytoplasm in plant cells. Based on these results we propose that high salt induces enhanced starch mobilization, CO_2 fixation and Redox energy production in order to meet the need for massive glycerol biosynthesis at high salinity (Figure 15). This interpretation is consistent with previous indications for synthesis of glycerol in *Dunaliella* from starch and photosynthetic carbon assimilation[156].

Another group of up-regulated proteins are key enzymes in ammonia assimilation (glutamine synthetase) and in biosynthesis of different amino acids. Induction of glutamine synthetase in plants usually reflects accumulation of ammonia, either from enhanced photorespiration or from enhanced protein degradation (see below). Carbamoyl phosphate synthetase, aspartate kinase-homoserine dehydrogenase, 2-isopropyl malate sysnthase and 3-phosphoglycerate dehydrogenase are key enzymens in the biosynthesis of arginine (and pyrimidines), threonine (and methionine, isoleucine), leucine and serine (and cysteine, glycine), respectively.

Various regulatory proteins involved in protein synthesis initiation (eIF3=TGF-beta receptor interacting protein), elongation and processing of proteins are up-regulated at 3 M NaCl. Related to this is GDP-mannose pyrophosphorylase, involved in protein glycosylation. Two other categories include general stress-related proteins in plants: antioxidants (Fe superroxide dismutase, thioredoxin peroxidase), involved in oxidative stress, and chaperones (chaperonin, HSP-70) involved in protecting proteins under stress conditions. NQR alpha is a subunit of the Na^+ extrusion complex involved in salinity tolerance in bacteria, as will be discussed below.

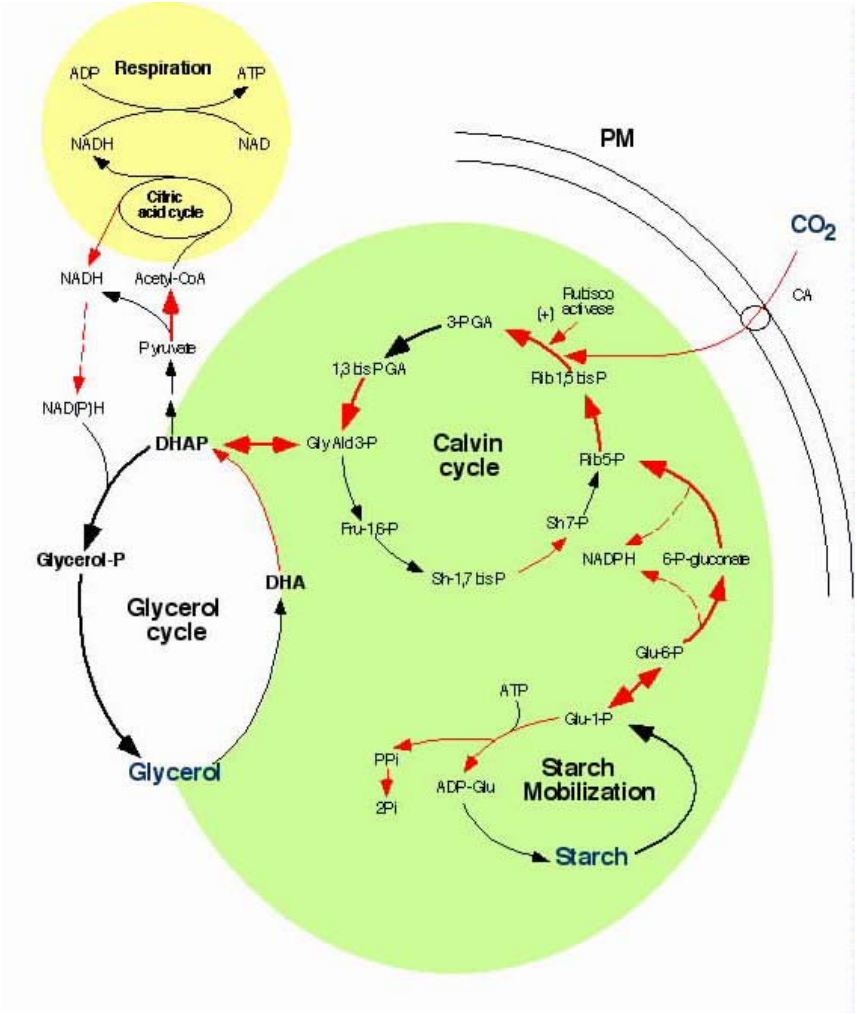

Figure 15 Salt-activated carbon flux in *Dunaliella*

Here is a metabolic network in Dunaliella that is activated by high salt. Paths in red are activated, some to produce an excess of glycerol. *(in collaboration with Dr. Adriana Katz)*

2.3.4 Salinity Tolerance in *Dunaliella*

(in collaboration with Dr. Adriana Katz and Professor Uri Pick)

Presented here is the first large-scale proteome analysis of salt up-regulated proteins in a lower plant whose genome is largely uncharacterized. Even though the analysis was limited mostly to soluble proteins that are up-regulated by no less than 2-fold and eliminated low abundance proteins, we were able to identify about 80% of the selected protein components. This success is partly due to the fractionation which increased the resolution of up-regulated proteins and mostly to the layered mass spectroscopic analysis and in particular the sequence-similarity searching algorithm MS BLAST.

Most proteins identified in the crude plasma membrane fraction were soluble proteins derived from the chloroplast or cytoplasm. The contamination of plasma membrane preparations with soluble proteins has been observed in many cases and probably results from adsorption of soluble proteins that are released during cell lysis. Integral membrane proteins were not identified in the crude plasma membrane fraction probably because they are under-represented by isoelectic focusing (IEF)[157]. Thylakoid membrane and integral plasma membrane proteins are currently being resolved and analyzed by different procedures (Pick and Katz). Considering these limitations, it may be expected that the overall number of salt up-regulated proteins in *Dunaliella* is much larger than revealed in this study.

The observation that major Calvin cycle enzymes are up-regulated by high salinity in *Dunaliella* contrast observations in plants and cyanobacteria of suppression of photosynthetic carbon assimilation and Calvin cycle enzymes under salt stress[158,159]. The typical response of plants to salt/drought stress is inhibition of photosynthesis and enhanced photorespiration, which results primarily from CO_2 limitation and is predicted to consume excess photosynthetically-produced Redox energy. The present results suggest that in *Dunaliella* photosynthesis was not inhibited by high salinity: the accumulation of plasma membrane carbonic anhydrases and of major Calvin-cycle enzymes suggests compensation for decreased CO_2 availability and enhanced CO_2 assimilation. Also the up-regulation of glucose 6-phosphate dehydrogenase, 6-phosphogluconate dehydrogenase, ferredoxin NADP oxidoreductase, pyruvate dehydrogenase and NADP malate dehydrogenase in *Dunaliella* suggest enhancement, rather than inhibition, of photosynthetic and respiratory Redox energy production and mobilization. These results may have a broader significance for identifying rate-limiting steps in carbon metabolism and energy production in plants under stress. Up-regulation of distinct key enzymes in the Calvin cycle and Redox-energy generation may relieve the general inhibition of photosynthesis under stress, which is a major limitation in the ability of plants to cope with salt stress.

This difference between the response of *Dunaliella* and higher plants to high salt may be explained by the need for massive glycerol biosynthesis in *Dunaliella* at high salinity: at 3M NaCl, the internal glycerol concentration in *Dunaliella* is close to 5M and it constitutes the major carbon pool under these conditions. As shown in Figure 15, these results suggest that glycerol is produced from enhanced CO_2 assimilation and starch degradation channeled through the Calvin cycle to dihydroxyacetonephosphate, which is reduced to yield glycerol. Plants and cyanobacteria utilize different osmotic elements

(proline, glycine-betaine) that are derived from amino acids. *Dunaliella* resembles yeast and fungi in the utilization of glycerol as an osmotic element[160]. Proteomic analysis of salt-induced proteins in *S. cerevisiae* revealed up-regulation of 3 glycerol biosynthetic/dissimilation enzymes, including dihydroxyacetone kinase[161], that were identified in this work.

The large increase in glutamine synthetase level in 3M NaCl indicates enhanced production of ammonia. The up-regulation of carbamoyl synthase can reflect enhanced synthesis of arginine, which is a storage form of assimilated ammonia in plants. Enhanced ammonia production may result either from photorespiration or from enhanced protein degradation. Although none of these possibilities can be excluded, there are indications against enhanced photorespiration. For example, the up-regulation of 3-phosphoglycerate dehydrogenase indicated that serine biosynthesis at high salinity in *Dunaliella* proceeds primarily from 3-PGA. High photorespiration activity produces glyoxalate which serves as an alternative substrate for production of serine and suppresses biosynthesis from 3-PGA. Also the indications for high CO_2 assimilation activity and the massive production of glycerol are inconsistent with enhanced photorespiration. It seems more likely that the upregulation of glutamine synthetase reflects ammonia production from enhanced protein degradation.

Conversely, the up-regulation of key enzymes in amino acid biosynthesis suggests a need for enhanced synthesis of new proteins. Another indication for enhanced biosynthesis and degradation of proteins at high salinity is the up-regulation of several regulatory factors in protein translation initiation and elongation and protein processing enzymes. Of particular interest is the dual-function eukaryotic initiation factor eIF3=TGF beta-receptor interacting protein (spot A10). Homologs of this protein in mammals and in plants interact with plasma membrane receptors and as such are part of a signal-transduction pathway in response to external stimuli[162,163]. A fission yeast homologue, of this protein Sum-1, is part of protein translation initiation complex 3 and was shown to be re-localized under salt or heat stress into distinct cytoplasmic domains[164]. Also the two mitochondrial translation elongation factors, EF-G and EF-Tu, may have dual functions: in addition to their established roles in protein biosynthesis, bacterial homologues of both proteins have chaperone properties and were proposed to protect proteins against mis-folding under stress[165,166]. The mitochondrial processing peptidase, zinc metaloproteases, and GDP-manose pyrophosphorylase are involved in processing and glycosylation of proteins. It may be noted that a major plasma membrane salt-induced proteins in *Dunaliella*, a triplicated

transferrin-like protein, is heavily glycosylated [167,168]. Conversely, the up-regulation of a 26S proteasome regulatory subunit suggests enhanced protein degradation at high salinity. The overall picture emerging from these results is of a dynamic reorganization of protein composition in different cellular compartments, by synthesis and processing of novel proteins as well as by massive degradation of other proteins.

Another particularly interesting up-regulated protein is the NQR alpha subunit homolog, a component of a bacterial Redox-driven Na+ extrusion system[169]. It has recently been reported evidence for Redox-driven Na+ extrusion system in *Dunaliella*[170]. The identification of a homologous protein to the bacterial system provides a tool for identification and cloning of this unique Na+ transport system that so far has not been identified in eukaryotes.

In summary, these results suggest that the response of the halotolerant alga *Dunaliella* to high salinity involves up-regulation of different enzymes and metabolic pathways, some of which differing from higher plants, such as carbon assimilation and mobilization for glycerol biosynthesis and Redox-drived Na+ extrusion, others common to plants and related organisms (chaperones, antioxidative enzymes), and others whose relationship to salt stress in plants has not yet been clarified. The latter include the factors regulating protein biosynthesis, processing and degradation, which may have cardinal importance in adaptation to high salinity.

2.3.5 Cross-Species Protein Identification Specificity of Mascot and MS BLAST

The Mascot software may detect a match as being significant when one or multiple MS/MS spectra from the analyzed peptides are correlated with a DB sequence from a related species, taking into consideration certain statistical issues, such as the goodness of fit between the observed and theoretically predicted fragment ions, mass accuracy, and the size of the DB. Mascot can generally align a few (one or two) peptides from proteins from organisms with unsequenced genomes with DB sequences from related organisms; for instance, in this study of *Dunaliella salina*, the majority of Mascot identifications were made by using reference sequences from *Chlamydomonas sp*.

However, Mascot is prone to produce false positives (like every protein identification method associated with MS); this usually occurs with the correlation of one or two spectra/peptides to a DB entry. A false positive occurs when the DB sequence recognized by Mascot correlation methods and scoring rules doesn’t truly represent the peptide that was fragmented to produce the resulting spectrum. In these cases, by

substituting isobaric amino acid combinations within the length of a peptide sequence, the sequence will mistakenly indicate an unrelated protein; a false positive. For instance, Mascot recognized the peptide YPIDWFK and a related miscleavage form, YPIDWFKK (1095.575), putting the protein identification well above the scoring threshold. However, upon MS BLAST analysis of the same data set, two other peptides and the isobaric peptide sequence YPL*VS*DFKK (1095.596) were recognized as aligning to a different protein entry that shared no significant sequence similarity with the Mascot identification (Figure 16). Furthermore, the complete fragment ion series could be read for this peptide (and its fully cleaved product, YPLVSDFK) in the MS BLAST identification, the entry had the appropriate predicted molecular weight, and a homologue existed in *Chlamydomonas reinhardtii*.

Upon close manual interpretation, that is, by overlaying of the retrieved sequence on to the spectrum and examining the closeness of fit of the predicted fragment ions, the high sensitivity, resolution and mass accuracy of Q(q)TOF MS enables the discernment of inconsistent mass accuracy errors and the inspection of the absence of usually intense fragment ions (such as, the b-2 ion, the a-2 ion, or b series or y series ions, or immonium ions), indicating the likelihood of a false positive identification. By extending the sequence coverage of the DB entry with the other tandem mass spectra, ambiguous sequences can be correlated with more evidence. By inspecting the data in this way, false positives can be minimized, as it has been done in this study.

The MS BLAST approach has been demonstrated here to produce higher sequence coverage than Mascot by aligning more peptides in an error-tolerant sequence-similarity manner, increasing the confidence of protein identifications by MS. MS BLAST recognized 266 peptides in total from tandem mass spectra (averaging 5.32 peptides per identification, range 1-11) whereas Mascot was able to align only 60 peptides in total to DB entries (averaging 2.6 peptides per identification, range 1-6). In 19 cases, MS BLAST extended Mascot sequence coverage (i.e. from 1 to 8 peptides aligned); MS BLAST averaged 6 peptides per identification for this group. In 28 additional cases, MS BLAST made identifications where Mascot was unable to produce any significant alignments, averaging 4.4 peptide alignments per identification. EST DB searching with Mascot averaged 1.75 peptides per identification in 20 cases.

In this manner confident cross-species protein identifications were determined. This data further demonstrates the fact that all proteins identified in proteomics are assigned with

certain confidence levels, with some protein assignments having higher confidence levels than others.

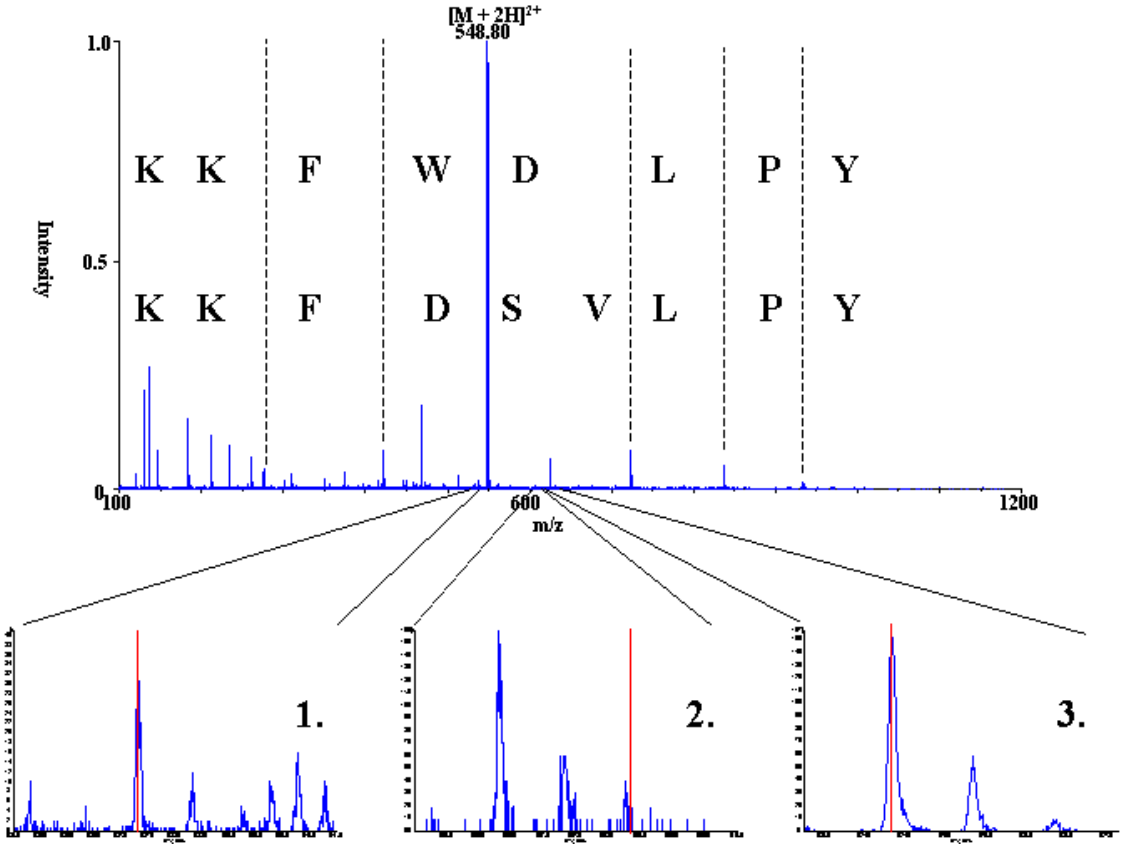

Figure 16 Mascot false positive identification.

A doubly-charged peptide ion of precursor mass *m/z* 548.80 was fragmented on a Q(q)TOF mass spectrometer; the acquired *m/z* range was from 100 to 1200 Thomsons (above). Mascot matched the amino acid sequence YPIDWFKK (1095.575 amu) from a protein DB to the tandem mass spectrum above. However, *de novo* sequencing and MS BLAST analysis of the same mass spectrum recognized the near isobaric peptide sequence YPLVSDFKK (1095.596 amu). Both sequences matched all of the same y-ions (fragments which retain the C-terminus of the peptide) except for three discrepancies (positions are shown in spectra 1-3 above, m/z 536-540, 605-610, 623-627, respectively). The sequence retrieved by Mascot required an ion match at the Red line in spectrum 2 to have a full y-ion series. The sequence retrieved by MS BLAST required an ion match at the Red lines in spectrum 1 and 3 to have a full y-ion series.

2.3.6 Assigning Biochemical Function to Proteins based on Sequence Alignments

Extending sequence coverage may help exclude false positive functional assignments based on sequence identity, but only to a certain degree. Rost has demonstrated using

bioinformatics that proteins with high sequence similarity (70-90%) can have divergent biochemical functions[171]. To overcome this problem for sequence-similarity protein identification, the analyzed proteins in proteomics need to be found in specific functional contexts. For example, here we attempted to identify soluble proteins induced under specific stress conditions. The experimental context used to recognize a set of proteins in proteomics already puts the proteins to be identified in a functional setting, thus imposing certain criteria on the character of most of those identifications; i.e. kinases, ribosomal, glycolytic, etc. For example, if we were to identify one of the *Dunaliella* proteins as a human histone (and if we did not question the integrity of the biochemical preparation), we would likely doubt the validity of such an identification, even if extensive sequence-similarity was determined. In the case demonstrated here, the multiple proteins identified were recognized to fit into coherent metabolic pathways and biochemical patterns, thus further suggesting that such sequence-similarity identifications likely correspond to the proposed biochemical function of the protein.

2.3.7 Sequence-Similarity Protein Identification in Plant Proteomics

The above study demonstrated that sequence-similarity protein identification techniques by MS could identify more than twice as many proteins recognized by the conventional software, subsequently greatly enhancing the proteome analysis in the alga *Dunaliella salina*. The DB searching sensitivity of MS BLAST was demonstrated to be a significant advantage over Mascot because of the BLAST algorithm's ability to perform true sequence-similarity alignments, and in no cases did these alignments rely upon precursor mass correlation or the exact matching of predicted peptide fragment ions with observed ions. Furthermore, this capability allows MS BLAST to extend the sequence coverage capable of conventional methods, thus utilizing a greater proportion of mass spectra for protein identification, and increasing the confidence of identifications. One disadvantage of MS BLAST is the requirement that amino acid sequences need to be predicted from peptide tandem mass spectra, either manually or automatically, thus requiring a certain abundance of protein and a certain level of spectra quality.

Even though these techniques were used with nanoelectrospray MS for the above study, the MS BLAST and MT sequence-similarity DB searching methods can be applied to data generated by different MS platforms (reviewed in section **1.1.3**). Recently, the method has been applied with MALDI TOF-TOF MS[172]. Similarly, MT may be applied to all MS data where peptide sequence tags can be determined from tandem mass spectra (see

section **2.5.2**). However, these methods are most effective in the analysis of gel separated proteins and due to statistical considerations they most likely will not be able to be utilized with shotgun proteomics methods such as MudPIT[39].

With these methods and future related developments, the proteomes of plants with unsequenced genomes will be more amenable for characterization by high-throughput MS techniques. Not only will more conserved proteins be able to be identified in distantly related plant species from those with sequenced genomes, but also more divergent homologous proteins will be able to be identified from those species that are closely related to organisms with sequenced genomes, such as maize, wheat, and barley, using the rice genomic sequences. Using these methods for protein identification, immediate, rapid and effective proteome analysis will be possible in plant biochemistry and physiology in many species, without having to wait for the completion of future genomic sequencing in many cases.

2.4 MS BLAST Specificity and Phylogenetic Considerations for Future Genomic Sequencing

2.4.1 Calculation of MS BLAST Specifity and Phylogenetic Reach of Protein Identification Using Available Resources.

To estimate the success of proteome characterization for a selection of organisms with unsequenced genomes, we calculated the specificity of MS BLAST to identify homologous proteins. The success rate of MS BLAST searches was correlated to the phylogenetic distance of the organism under study to the next closest organism with a fully sequenced genome. With the analysis of 10 peptides (each 10 amino acids in length, with two undetermined residues placed at randomly chosen position in their sequence), MS BLAST can successfully identify proteins down to a limit of 65% identity. Taking into consideration eight species with sequenced genomes, we propose groups of species where sequence-similarity methods will be effective (Figure 18). With these developments, functional proteomics in important model species with unsequenced genomes has the potential to be advanced by MS and sequence-similarity DB searching.

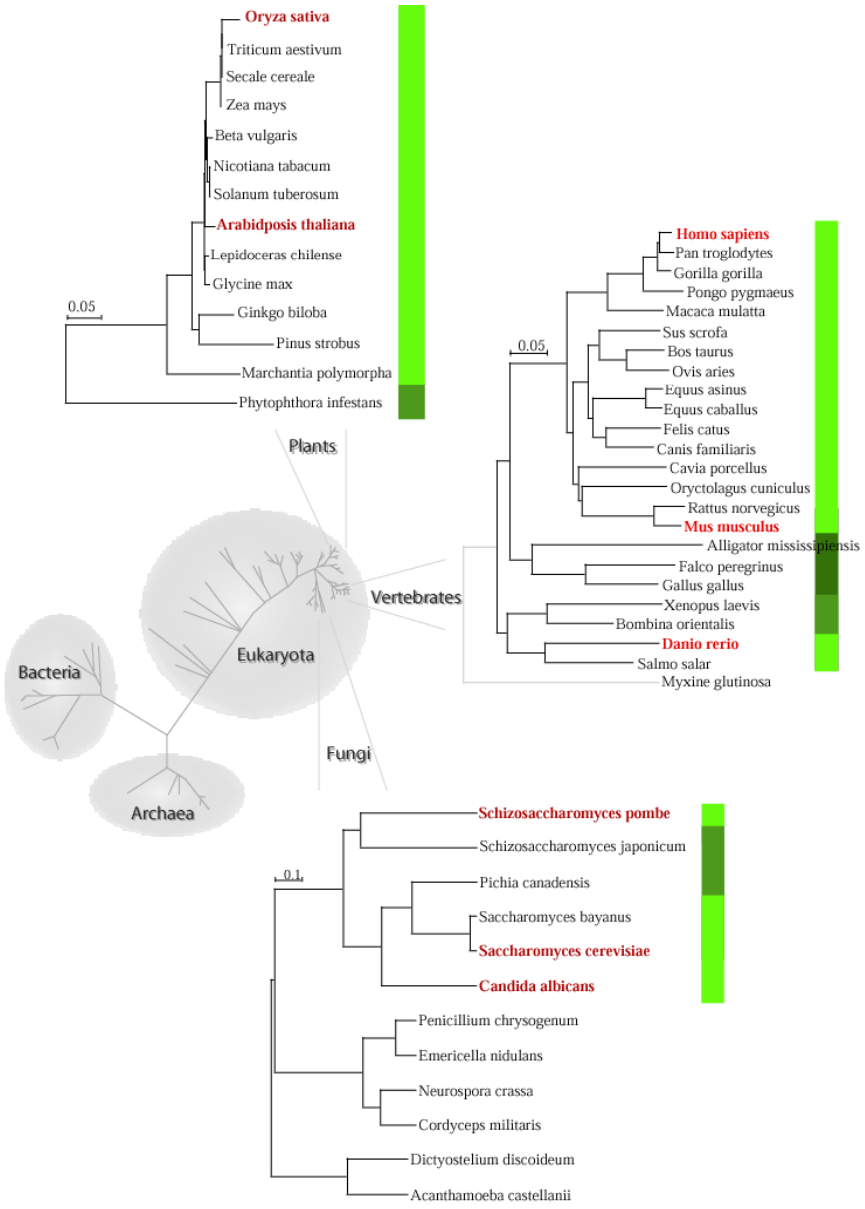

Figure 17 Predicted success of proteomics in organisms with unsequenced genomes

Three partial phylogenetic trees of major subkingdoms branch from a phylogenetic tree of all living organisms. Phylogenetic analysis was used to estimate the success of proteome characterization by mass spectrometry and sequence-similarity database searching, based on the specificity of the MS BLAST to identify homologous proteins. To estimate the success of MS BLAST searches, queries comprised of 8 peptides sequences were used, 10 amino acids in length and with 2 errors in each peptide to simulate ambiguities in spectrum interpretation. Organisms in red have sequenced genomes. Color code is based on the distance between species and corresponding protein identification coverage of the proteome by MS BLAST; light green: >90% coverage; middle green: 30-50% coverage; dark green: up to 30% coverage. *(in collaboration with Dr. Bianca Habermann)*

2.4.2 Genomic Sequencing and Proteomics

Currently, there is a substantial debate over which organisms' genomes deserve to be sequenced next[173,174]. 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[173]. As already mentioned, the ability to identify proteins depends on DB 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 over 1.7 million described species, and potentially as many as 10 million species in the biosphere, it is evident that the research community will not be able to sequence the genome of every species. Many biological researchers investigating proteomes have already experienced the lack of genomic resources as an inability to identify proteins by MS. For example, proteomics studies in maize, an economically important organism, have been compromised due to the lack of DB resources and an inability to use available DB resources effectively[78]. 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[78,148].

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 human, 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.

As genomic sequences become available to the public in the form of annotated DB 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 *Arabidopsis thaliana* upon plant proteomics (Table 8). 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 DB entries. For example, all cross-species identifications in the study of the maize proteome by Chang *et al.* could have been accomplished using DB entries from the *Arabidopsis* genome and sequence similarity searches (Table 9) Even though Chang *et al.* identified maize proteins using DB 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, species 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[175,176].

Table 8 Sequencing of the *Arabidopsis* genome and its effects in proteomics.

Year	Genomics Development	Proteomics	Ref. Organisms in the ID of Proteins	MS	MS/MS	SSS	Citation
2000	*Arabidopsis thaliana*	***Papaver somniferum***	*ARABIDOPSIS* (31), *P. sativum*(6), *Glycine max*(6), *Nicotiana tabacum*(3), *Solanum tuberosum*(3), *Oryza sativa*(3), *Vitis vinifera*(3), *Protea neriifolia*(2), *Lavatera thuringiaca*(2), *Brassica oleracea*(2), *Fritillaria agrestis*(1), *Z. mays*(1), *Brassica juncea*(1), *Datisca glomerata*(1), *Hordeum vulgare*(1), *S. cerevisiae*(1), *S. 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**	[81]
		Zea mays	*ARABIDOPSIS* (2), *B. vulgaris*(2), *B. napus*(2), *G. max*(2), *C.roseus*(4), *O. sativa*(3), *N. tabacum*(3), *M. sativa*(3), *H. vulgare*(2), *C. reinhardtii*(1)	X			[77]
		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		[79]
		A. thaliana	*ARABIDOPSIS* (33), Z. mays(3)	X	X		[82]
		Z. 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			[78]
2001		***T. bruei***	*ARABIDOPSIS* , Mouse, *D. melano*, *E. histoly*, *C. elegans*, *O. sativa, S. cerevisiae, and others*(9)		X	X	[60]
		X. laevis	**ARABIDOPSIS* (1) *B. taurus(1) , *Mouse (1), *Rat (1), *Human (1),		X	X	[85]

Organisms in the column "Developments in Proteomics" had proteins identified by cross-species identification. Organisms in the column "Ref. Organisms…" 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. *Multiple alignments are made to different species from the analysis of a single protein. **Peptides were sequenced by Edman degradation.

Table 9 *Arabidopsis* homologues to database references used in maize protein identification[77]

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*	O82399	*Arabidopsis*	77%
X91347	*H. vulgare*	P57751	*Arabidopsis*	77%
AF020271	*M. sativa*	BAA97065	*Arabidopsis*	80%
X77944	*N. tabacum*	AAK73989	*Arabidopsis*	88%
U38199	*O. sativa*	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. BLASTP searches were performed at NCBI to create the table above.

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[12]. Once a bird's genome 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 MS 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.

2.5 Analytical Strategies in Proteomics

2.5.1 Analytical Strategies

Protein identification by MS and sequence DB searching was established in 1993, and since that time the proteomics community has witnessed a proliferation of analytical strategies for protein identification. Analytical strategies are composed of three components: *mass spectrometry platforms*, *spectra-database sequence correlation methods*, and *sequence databases*. To extend protein identification capabilities, as well as to advance the efficacy of protein identification in organisms with unsequenced genomes, a number of recent developments are pointing to new analytical strategies to interrogate proteomes. Specific types of mass spectrometers produce spectra of varying quality, and alternate interpretation methods are suited for specific types of spectra. When a specific *mass spectrometry platform* is combined with a specific *correlation method* and a specific type of *database*, this combination may be more or less effective for protein identification than a different combination. Here, to discuss these relationships, we will designate specific analytical strategies by the annotation "*mass spectrometry platform*"(where "MS/MS" means any tandem MS method, unless otherwise named)—"*spectra-DB sequence correlation method*"–"*sequence database*." To enhance protein identification ability, a number of combinations must be developed, compared, and employed simultaneously in proteome analysis. Here this complexity is attempted to be systematically described and potential new strategies for protein identification are recognized. Multiple strategies are now applied simultaneously to increase sensitivity, throughout, and reliability of the characterization of proteomes. Now, by assessing the complexity of the interplay of MS, bioinformatics and sequence databases, we can begin to predict future approaches and challenges in the development of proteomics.

2.5.2 Spectra-Sequence Correlation Methods and Analytical Strategies

Mass spectra are correlated with DB sequences primarily in three ways: the *mass pattern*, the *amino acid sequence*, and the *sequence tag* (Figure 18). These three methods derive information of different qualities from peptide MS/MS, and they each suit the interpretation of spectra (more or less effectively) depending on the spectra's signal-to-noise, mass accuracy and resolution. Furthermore, each of the different methods has distinct capabilities

to identify analyzed peptides whose sequences share only partial identity with DB sequences.

Mass patterns (composed of lists of *m/z* values of detected peaks along with the corresponding peak intensities) are used in two types of MS analysis: PMF and MS/MS. In PMF, masses of intact peptides are determined and are used for DB searches. Historically, *mass patterns* derived from peptide mass fingerprints were first used to search protein sequence databases (PMF-*mass pattern*-Protein DB)[177-180] (Figure 19)(Table 10). Observed peptide masses are compared with peptide masses calculated from the *in silico* digestion of protein DB sequences with trypsin, and resulting matches are scored accordingly. In these softwares, mostly mass *values* have been used, but now there are attempts to incorporate peak intensities to improve the specificity of the identifications[102,181,182]. When analyzed peptide sequences deviate away from the identity of corresponding sequences in DB entries, either because of amino acid substitution or post-translational modifications, the probability of successful identification by this method diminishes[13], and MS/MS must be employed.

A second common analytical strategy correlates MS/MS spectra through *mass patterns* with protein DB sequences (MS/MS-*mass pattern*-Protein DB)[15,102,147,187-189]. In these cases, observed masses of peptide precursors and masses and intensities of their fragment ions are compared with theoretical peptide masses and fragments derived from sequence databases with the application of particular enzyme specificity and peptide fragmentation rules. The method of scoring of the similarity between MS/MS spectrum and DB sequence employs certain peptide fragmentation models, and those models are instrument-dependent. To this end, DB searching programs usually allow the specification of instrument type. MS/MS spectra with higher mass accuracy will be able to interrogate databases more specifically, increasing the probability of identification with fewer peptides[15]. Furthermore, MS/MS spectra with high signal-to-noise will give best results, as true peptide fragment ions won't be obscured by background peaks in spectra. *Mass pattern* methods are currently more diverse and have experienced a greater attention in the proteomics community than other protein identification methods. The correlation of *mass patterns* with DB sequences also have some "error-tolerant" capabilities that withstand amino acid substitutions between those peptides observed and sequences present in a DB[15,103]. In addition to protein DB interrogation, spectra can be correlated through *mass patterns* with EST databases (MS/MS-*mass pattern*-EST DB)[15,102,147,190] and more recently with genomic databases (MS/MS-*mass pattern*-Genomic DB)[184,190] (Figure

19). Besides using complete *mass patterns* for protein identification, the recently developed "peptide end sequencing" makes use of N- and C-terminal peptide fragment ions detected in the low *m/z* region, along with a parent mass, for the identification of low abundance proteins[191,192]. However, if multiple non-isobaric amino acids substitutions occur within individual peptides or if unknown multiple post-translational modifications exist, then the probability decreases that the protein will be identified by these methods. In these cases, a different method of spectra interpretation is employed.

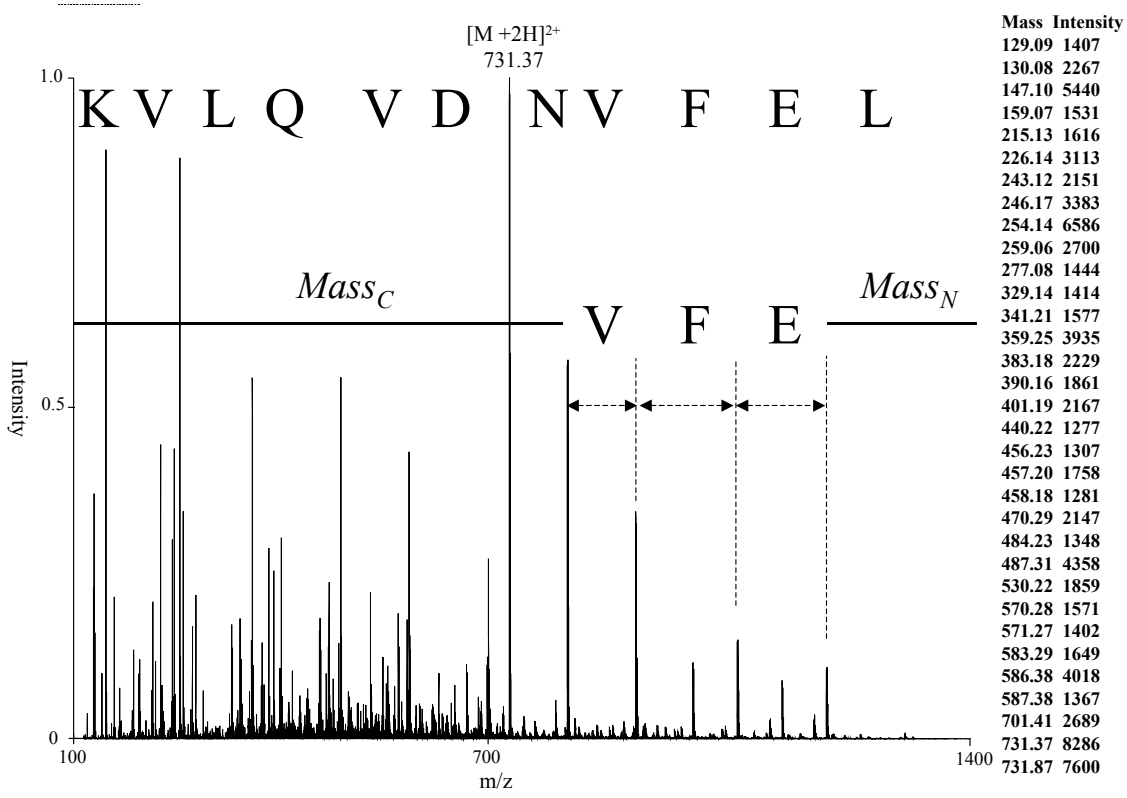

Figure 18 Representative Information from the Mass Spectrum in Proteomics.

Mass spectra can be represented by primarily three types of information for their correlation with DB sequences. *Mass patterns* are composed of lists of *m/z* values of detected peaks and corresponding peak intensities (partial list shown above/right). *Amino acid sequences* are derived from spectra considering precise mass differences in ion series, and annotated in the form of series of amino acid symbols (KVLQV…). *Sequence tags* consist of partial amino acid sequences combined with two mass values which lock the sequence within the length of a peptide, and a parent mass (the sequence tag for the above spectrum is (815.44)VFE(1190.62), peptide mass 1460.72.

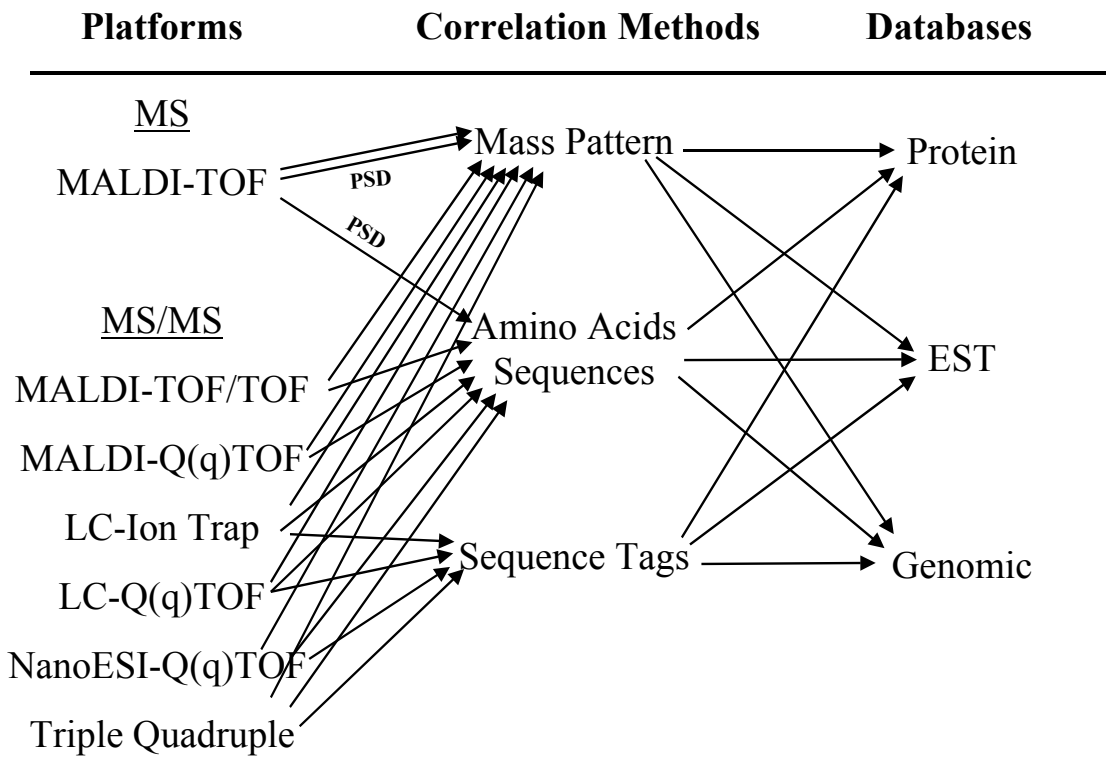

Figure 19 Strategy Network

Analytical strategies are composed of three components: mass spectrometry *platforms*, spectra-sequence *correlation methods*, and sequence *databases*, which can be read left to right, in the scheme above. Arrows show the interaction of mass spectra with representation methods (mass pattern, amino acid sequences, sequence tags)(see Figure 18) and nucleotide or amino acid databases (protein, EST, genomic). Arrows represent robust and less established interactions equally, despite practical limitations to some.

Table 10 Analytical Strategies

Analytical Strategy	Ref.
MS	
PMF-*mass pattern*-Protein DB	[177]
PMF-*mass pattern*-EST DB	[102]
PSD-*mass pattern*-Protein DB	[22]
PSD-*amino acids*-Protein DB	[183]
MS/MS	
MS/MS-*mass pattern*-Protein DB	[147]
MS/MS-*mass pattern*-EST DB	[147]
MS/MS-*mass pattern*-Genomic DB	[184]
MALDI-Q(q)TOF-*amino acids*-Protein DB	[62]
ESI-Q(q)TOF-*amino acids*-Protein DB	[85]
ESI-TQ-*amino acids*-Protein DB	[85]
LC-Q(q)TOF-*amino acids*-Protein DB	[63]
ESI-Ion Trap-*amino acids*-Protein DB	[185]
MS/MS-*amino acids*-EST DB	*
MS/MS-*amino acids*-Genomic DB	*
MS/MS-*sequence tag*-Protein DB	[86]
LC-Ion Trap-*sequence tag*-Protein DB	[186]
MS/MS-*sequence tag*-EST DB	[88]
MS/MS-*sequence tag*-Genomic DB	[82]

Representative analytical strategies are listed above. Analytical strategies are represented by the annotation "*mass spectrometry platform*"(where "MS/MS" means any tandem MS method, unless otherwise named)—"*spectra-sequence correlation method*"–"*sequence database*." Researchers who contributed to the development of these strategies are cited in the column on the right. *These strategies are currently established in house and are under development.

Amino acid sequences enable spectra interpretation for the identification of proteins that are homologous to DB sequences despite having no peptides of identical precursor mass with those theoretically predicted from DB entries (as required by *mass pattern* searches). In this analytical strategy, *amino acid sequences* produced from MS/MS spectra can be correlated with protein DB sequences (MS/MS-*amino acids*-Protein DB)[57,61,62]. *Amino acid sequences* can be produced *de novo* from MS/MS spectra of peptides primarily by two methods: via chemical modification of peptide N- or C-termini or by direct computer-assisted interpretation of spectra by sequence prediction algorithms (see section **1.1.4**). Chemical modification methods demand relatively large sample quantities, are manually laborious, and ultimately obscure spectra for parallel *mass pattern* interpretation. However, sequence prediction algorithms rapidly generate putative *amino acid sequences*, although often multiple degenerate sequences are predicted with similar statistical confidence. To utilize this information, numerous peptide sequences from multiple fragmented peptides must be compiled in a query for a sequence-similarity DB search. This has been accomplished in dedicated softwares based on FASTA[59] and BLAST[58] sequence homology searching algorithms. Mass spectrometry driven BLAST (MS BLAST) is one example of this type of strategy[62].

Amino acid sequence-based interpretation methods are also flexible for the development of alternate analytical strategies (Figure 19). In one proteome analysis, MS BLAST was employed to correlate MALDI-Q(q)TOF spectra with protein DB sequences (MALDI-Q(q)TOF-*amino acids*-Protein DB)[62]. However, this is just *one* strategy among other possible paths considering available MSPs, databases, and developments in high-throughput spectra processing and MS BLAST DB searching capabilities. NanoESI-QqTOF-*amino acids*-Protein DB and NanoESI-TQ-*amino acids*-Protein DB[85], as well as LC-QqTOF-*amino acids*-Protein DB[63], are other options for sequence-similarity identification depending on available instrumentation (Table 10). Furthermore, peptide sequence can be generated *de novo* from LC-ion trap mass spectra and used for protein DB interrogation (LC-Ion Trap-*amino acids*-Protein DB)[185,193]. In addition, MS/MS-*amino acids*-EST DB and MS/MS-*amino acids*-Genomic DB are other possible DB searching strategies that could also be employed for protein identification with *amino acid sequences* (currently both approaches are operating in house) (Table 10). All of these strategies enable new capabilities for protein identification; however, their efficiency is limited when sequence prediction fails.

Sequence tags enable spectra interpretation for the identification of proteins from low intensity spectra or where background chemical noise interferes with full-length amino acid sequence determination. *Sequence tags* consist of a few (2-4) determined amino acids, along with two mass values, which lock the sequence stretch within the length of the peptide[86] (Figure 18), effectively combining mass data with sequence data. This requires that only one section of a complete mass spectrum be interpreted correctly by manual inspection. *Sequence tags* can be employed to identify proteins by correlating MS/MS spectra (of preferably high mass accuracy[15]) with protein (MS/MS-*sequence tags*-Protein DB)[87], EST (MS/MS-*sequence tags*-EST DB)[88], or genomic DB sequences (MS/MS-*sequence tags*-Genomic DB)[82] (Figure 17). However, it is rather difficult to assemble *sequence tags* from MALDI-TOF/TOF and MALDI-Q(q)TOF spectra[194], mostly because prominent y- or b- fragment ion series often cannot be determined unambiguously[25].

Error-tolerant *sequence tags* enable the identification of proteins that are homologous to DB entries[91]. The recently developed MT software correlates multiple (often partial) *sequence tags* with individual DB entries, whereas previous identification techniques using *sequence tags* rely upon the correlation of individual spectra to DB entries alone. The MT approach is similar to approaches with *mass patterns*[15,102,147] or *amino acid sequences*[57,61,62] that rely upon information from multiple spectra to increase the confidence of protein identifications. The MT approach was also extended to EST DB searching, and in the future, a MT approach could greatly facilitate genomic DB searches with *sequence tags*.

The correlation of *mass patterns*, *amino acid sequences*, and *sequences tags* with DB sequences all have their own unique evaluation schemes to discriminate correct from false positive protein identifications. *Mass pattern* identification methods primarily score the quality of a tandem mass spectrum's fit to a predicted model spectrum, while taking into consideration other DB search parameters (*i.e.* SEQUEST[147], Mascot[102], Protein Prospector[15], Scope[187], Sonar MS/MS[188], ProbID[189]). Protein identifications made using *amino acid sequences* are evaluated by the significance of the alignment of a sequence query to a DB sequence and the probability of that alignment occurring in a DB of a specific size (*i.e*. CIDentify[57], MS BLAST[62], and FASTS[61]). In the evaluation of multiple *sequence tags* (MT), the probability that such a set of *sequence tags* align at random to a DB entry (within a DB of a particular size and at a particular mass accuracy) provides a measure of confidence of the identification.

The existence of multiple scoring schemes for the determination of the significance of spectra-sequence alignments raises the question of whether proteins identified by one method are considered positive identifications by another similar method (for instance, two *mass pattern* methods, or two *amino acid sequence* methods), which could ultimately compromise the certainty of proteome characterization. A solution to alternate scoring systems can be provided with the implementation of algorithms that calculate probabilities that the spectra acquired derive from specific known sequences, rather than their similarity occurs at random[195]. Another possibility is to develop empirical statistical models based on search results to assess the validity of protein identifications by MS and DB searches[196]. This approach suggests that in the future statistical data interpretation methods can be applied to search results acquired by different MSPs and individual softwares[196].

2.5.3 Bridging the Gap: A Network of Strategies

The character of various types of MS/MS spectra directly determines the operation of *spectra-DB sequence correlation methods* (*mass lists*, *amino acid sequences*, and *sequences tags*) and different methods enable unequal abilities to interrogate databases (protein, EST, and genomic). Alternate *MSPs*, *spectra-DB sequence correlation methods*, and *databases* can be combined yielding varying degrees of effectiveness for protein identification, and new analytical strategies are rapidly creating novel ties between different types of spectra through the three discussed interpretation mechanisms (Figure 19). Each particular proteome analysis will require a different strategy or *strategies* to successfully identify proteins at a high throughput depending on the MSPs at hand, the proteins' abundance, the quality of experimental mass spectra, and the availability of species-specific DB sequences, as well as the DB type or types employed. This applies both to the proteomics of organisms with sequence resources as well as those species with unsequenced genomes.

Simultaneously applying multiple analytical strategies has enabled the most effective approaches in the analysis of complex protein mixtures by increasing sensitivity, throughout, and reliability of the characterization of proteomes. In 1996, multiple strategies began to be employed simultaneously with the application of PMF-*mass pattern*-Protein DB and MS/MS-*mass pattern*-Protein DB together in one study (Table 11)[197]. Today, multiple strategies are employed on a regular basis for high-throughput proteomics. In the recent characterization of the human nucleolus, five strategies (PMF-*mass pattern*-Protein DB, MS/MS-*mass pattern*-Protein DB, NanoESI-Q(q)TOF-*sequence tags*-Protein DB,

NanoESI-Q(q)TOF-*sequence tags*-EST DB, and NanoESI-Q(q)TOF-*sequence tags*-Genomic DB) were all employed simultaneously[87]. In addition, in this study, multiple protein (nrdb and IPI) and genomic (phases 0-3 of the uncompleted human genome sequence) databases were interrogated, adding addition reference sequences to facilitate protein identification.

The effective characterization of proteomes from organisms with unsequenced genomes relies upon the use of multiple analytical strategies as well. The proteome of maize leaves was analyzed using two strategies, PMF-*mass pattern*-Protein DB and PMF-*mass pattern*-EST DB[78]. However, using only these two strategies enabled the identification of 216 spots out of 300 analyzed from 2-D gels. The authors recognized the importance of MS/MS methods (perhaps homology-based) for further studies to be more comprehensive. Similarly, in the proteomics of the pea symbiosome, NanoESI-Ion Trap-*mass pattern*-Protein DB and NanoESI-Ion Trap-*mass pattern*-EST DB methods were applied, but failed to identify almost one half of the proteins, with 46 identifications out of 89 spots analyzed from 2-D gels, despite the application of MS/MS methods[198]. In another example, five strategies (PMF-*mass pattern*-Protein DB, NanoESI-Q(q)TOF-*mass pattern*-Protein DB, NanoESI-Q(q)TOF-*amino acids*-Protein DB, NanoESI-Q(q)TOF-*sequence tags*-Protein DB, and NanoESI-Q(q)TOF-*sequence tags*-EST DB) were applied simultaneously to characterize the African Clawed frog *Xenopus laevis* microtubule-associated proteome, successfully identifying 62 proteins from 55 protein bands from one dimensional gels.

Table 11 Application of Analytical Strategies in Parallel

Year	Proteomics	Ref.
1993	**Bacteria Proteomics**	[177]
	PMF-*mass pattern*-Protein DB	
1996	**Yeast Proteomics**	[197]
	PMF-*mass pattern*-Protein DB	
	MS/MS-*mass pattern*-Protein DB	
2001	**Maize Proteomics**	[78]
	PMF-*mass pattern*-Protein DB	
	PMF-*mass pattern*-EST DB	
2002	**Pea Symbiosome Proteomics**	[198]
	NanoESI-Ion Trap-*mass pattern*-Protein DB	
	NanoESI-Ion Trap-*mass pattern*-EST DB	
2002	**Human Nucleolus Proteomics**	[87]
	PMF-*mass pattern*-Protein DB	
	MS/MS-*mass pattern*-Protein DB	
	MS/MS-*sequence tag*-Protein DB	
	MS/MS-*sequence tag*-EST DB	
	MS/MS-*sequence tag*-Genomic DB	
2002	**African Clawed Frog Proteomics**	[199]
	PMF-*mass pattern*-Protein DB	
	MS/MS-*mass pattern*-Protein DB	
	MS/MS-*amino acids*-Protein DB	
	MS/MS-*sequence tags*-Protein DB	
	MS/MS-*sequence tag*-EST DB	

Representative proteomic studies are shown above. The name of the organism is in bold. The analytical strategies that were employed in those particular proteome studies are listed below the name. Researchers who conducted these studies are cited in the column on the right and the year of the study in the left hand column. The legend of Table 10 describes strategy annotation.

3 Conclusion

Considering the developments in MS informatics presented here, mass spectrometrists can begin to systematically develop, apply, and evaluate the effectiveness of these strategies for the functional characterization of proteins based upon sequence identity (Figure 19). In many proteomics studies, one strategy will produce the greatest number of identifications, while alternative methods will produce diminishing returns (but the methods nonetheless allow more data to be accumulated in the analysis of mass spectra, increasing the confidence of individual protein identifications and adding to the volume of identifications). This was demonstrated in the *Dunaliella* study where MS BLAST identified twice as many proteins as Mascot, and subsequent application of MultiTag identified only one more protein (see section **2.3**). However, in the *Xenopus* study, MS BLAST identified only three more proteins than Mascot, but MultiTag identified almost twice as many as Mascot (see section **2.2**). This excentuates the fact that each proteomics study has its own qualities, such as DB availability and spectra quality, thus multiple strategies must be explored in order to develop a set of tools that can identify proteins in many different situations; each alternative strategy has the potential to perform a vital function in future proteomics studies. In the future, we can expect new MSPs, new spectra-sequence correlation methods, as well as perhaps new types of databases to contribute to the proliferation of protein identification strategies. We can also begin to predict future strategies that provide new potential for the MS community (i.e. currently MS BLAST is being developed for EST and genomic DB searching; these strategies will be significant resources where species have catalogued raw genomic or EST sequence and limited protein databases, currently this includes dog, chicken, *Xenopus laevis*, and *Chlamydomonas*, among others). However as a general trend, sequence-similarity protein identification methods are able to identify twice as many proteins as conventional softwares, producing a significant contribution to proteomics.

Sensitive and confident protein identification by MS is a never-ending problem. All of the world's species will not have their genomes sequenced, sequence databases will always be at various stages of development, and the biological sciences will find new specimens to be analyzed at the level of the proteome. However, the inherent homology of proteins in phylogenetically related species can be exploited for mass spectrometry-based proteomics. In order to thoroughly and sensitively characterize these proteomes, the application of multiple analytical strategies provides a successful approach. In the future, we can expect that the network between various types of mass spectra and different types of

DB sequences will become more and more integrated with the development of these informatic approaches.

4 Materials and Methods

4.1 Peptide Tandem Mass Spectrometry for MultiTag Development

4.1.1 Software

MT is a stand-alone application on the Microsoft Windows platform. MT code was written using C++ language with Microsoft Visual C++ and Microsoft Foundation Classes (Microsoft Inc. CA). Sorting and statistical evaluation of ~5,000 hits takes about 1 second on the Pentium IV workstation. Dr. Alexander Golod programmed MT.

4.1.2 Sample Analysis

Proteins in a purified extract from *Xenopus laevis* oocytes were separated on a one-dimensional polyacrylamide gel and visualized by staining with Coomassie (**4.2.1**). Protein bands were excised and in-gel digested with trypsin as previously described[200]. Extracted peptides were first analyzed by PMF on a Reflex IV (Bruker Daltonik, Bremen, Germany) matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometer, and obtained peptide mass fingerprints were submitted for DB searching by Mascot (Matrix Science Ltd, UK) software[102]. None of the samples were positively identified. Samples were further analyzed by nanoelectrospray tandem mass spectrometry on a QSTAR Pulsar *i* Q(q)TOF instrument (MDS Sciex, Canada).

4.1.3 Interpretation of Tandem Mass Spectra and Database Searching

Sets of uninterpreted tandem mass spectra were used to search databases first with Mascot[102] (the conventional software), and when no positive identifications were achieved, the spectra were interpreted manually. Sequence tags were determined by the interpretation of tandem mass spectra using BioAnalyst QS software (Applied Biosystems, CA). DB searching was performed using the PepSea program (a part of the BioAnalyst QS package) against a comprehensive non-redundant protein sequence DB downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). No constraints on the protein molecular weight or species of origin were imposed. The mass tolerance was set to 0.05 Da for fragment ions and 0.1 Da for precursor ions. Hits of error-tolerant searches were pooled in a spreadsheet (MS Excel) and were encoded by a peptide precursor mass and a letter code for the matched regions of the corresponding sequence tag in order to facilitate subsequent processing by the MT program. The entire pool of hits was submitted to the MT program for sorting and statistical evaluation. MS/MS spectra were further analyzed by MS BLAST sequence-

similarity DB searches at http://dove.embl-heidelberg.de/Blast2/msblast.html against the "nrdb" protein DB.

4.1.4 EST Database Searching

4.1.4.1 Software Alteraton

MT-Integrated DB Search Software. The "MTSearch" script was developed to automatically generate a list of DB search results from a list of sequence tags; this was performed by Dr. Ignat Shilov of Applied Biosystems (Foster City, CA, USA). Tags were used for searching a DB in a stringent fashion (matching regions 1, 2 and 3, see figure 5) and error-tolerant fashion: a search tolerating a mismatch of the C-terminal mass (matching regions 1 and 2); a search tolerating a mismatch of the N-terminal mass (matching regions 2 and 3); and searches tolerating one mismatch in the amino acid sequence (matching regions 1 and 3); the hits were additionally encoded by the mass of the precursor ion and by the abbreviated matching region (NC, N, C, or E, respectively) in the sequence tag and compiled in a list for submission to MT.

MT Modifications The existing MT software was modified so the average number of tryptic peptides DB entry could be specified. The average protein length in a non-redundant DB was previously determined to be 492 amino acids (corresponding to ~60kD). The average length of a tryptic peptides was designated at 12 amino acids, setting the average number of tryptic peptides per DB entry at 41. Since the average length of an EST entry codes for 166 amino acids (EST_others, Nov. 27, 2002, NCBI), this number was divided by 12 and the value for EST DB searching was set at 14.

4.1.3.2 Database Searching

Mascot queries were generated from tandem mass spectra using the processing script Mascot v.1.6b2 as an extension of BioAnalyst QS software (Applied Biosystems). Spectra were centroided and peaks were merged at 0.05 Thomsons, and peak lists contained mass values from peaks ≥2% base peak. DB searches with Mascot were performed on an internal server with a precursor mass tolerance of 0.1 Da and a fragment ion mass tolerance of 0.05 Da, default precursor charge states were set at +2 and +3, with trypsin enzyme specificity, one miscleavage allowed, variable methionine oxidation, fixed carboxyamidomethyl cysteine, instrument type set at default, and no restrictions for protein molecular weight, but restricted to DB entries from the species *Xenopus laevis*. The Mascot identifications were made using the *Peptide Summary Report* for enhanced sensitivity.

Sequence tags were generated as previously using the Bioanalyst QS software. MT searches were performed using the PepSea software as a part of BioAnalyst QS, with a precursor mass tolerance of 0.1 Da and a fragment ion mass tolerance of 0.05 Da, with trypsin enzyme specificity, and fixed carboxyamidomethyl cysteine. Search results were analyzed with the MT software described above. MT parameters were set: 1,396,530 DB entries searched (6 frames X 232,755 *Xenopus laevis* EST entries), 0.1 Da mass accuracy, and 14 for number of peptides per entry. MT has no species restriction parameter and therefore all cross-species alignments were ignored.

Both methods searched the same DB EST_others (November 27, 2002), from the National Center for Biotechnology Information, and only used the *Xenopus laevis* subset of this DB. The identity of all ESTs was verified by blastx DB searches at the NCBI internet site.

4.2 Xenopus Experiments

4.2.1 Purification of MAPs From Xenopus Egg Extract.

Mitotic MAPs were prepared by Dr. Andrei Popov in the laboratory of Professor Eric Karsenti. The complete procedure is described in[199]. Mitotic *Xenopus* egg extracts were prepared according to A. Murray[97]. Assembled microtubules were prepared with pig brain tubulin and stabilized by addition of Taxol and pelleted. To bind MAPs and motors to microtubules, the extract was incubated with prepolymerized microtubules, plus GTP and AMP-PNP (Adenosine-5'-imidotriphosphate). AMP-PNP, a non-hydrolysable analogue of ATP, was shown to stabilize motors interaction with microtubules[201]. The microtubule pellet was collected and prepared by SDS-PAGE.

4.2.2 Mass Spectrometry Analysis.

Individual protein bands were in-gel digested with trypsin as previously described[200]. Collected peptides were analyzed first by peptide mass mapping on a Bruker Reflex IV™ MALDI TOF mass spectrometer. Peptide mass maps were searched against NCBI's protein DB (MSDB) using the Mascot server ver.1.8[102] with a mass tolerance of 150 ppm. No search parameters were imposed to limit species-specificity for all searches. Proteins not identified by peptide mass mapping were subjected to MS/MS analysis by nanoelectrospray mass spectrometry. All MS/MS spectra were first used for DB searches with Mascot at a tolerance of 0.1 Da for the precursor mass and 0.05 Da for fragment ion masses. All

MS/MS spectra were then used for DB searches with MS BLAST (http://dove.embl-heidelberg.de/Blast2/msblast.html) against a protein DB (nrdb95), and then finally again analyzed by MT (Table 1). For MS BLAST, BioAnalyst software from Applied Biosystems (Foster City, CA) predicted *de novo* amino acid sequences with a tolerance of 0.1 Da for the precursor mass and 0.05 Da for fragment ion masses. The MSBlast processing script (version 1.1beta) automatically generated peptide sequence queries from MS/MS spectra for the MS BLAST DB searches. For MT interpretation, sequence tags were manually constructed, using BioAnalyst, from abundant y-ion series with m/z usually larger than the precursor ion in MS/MS spectra. All sequence tags contained 2-5 amino acid residues within the tag. Complete sequence tags (containing all three regions) were first searched against NCBI's protein nonredundent DB (March 6, 2002) and NCBI's EST_others DB (March 6, 2002) using PepSea DB searching and subsequent MT analysis as previously described[91].

4.2.3 Density Gradient Fractionation.

The motor fraction was prepared above. Proteins were then eluted by addition of ATP, concentrated by centrifugation, and resolved on 5-45% sucrose gradients. Fractions were collected manually from the top of the tube and analyzed on an SDS 6-20% polyacrylamide gradient gel. This experiment was completed by Dr. Andrei Popov in the laboratory of Professor Eric Karsenti, and the complete procedure is described in[199].

4.2.4 Immunoblot Analysis.

SDS-PAGE resolved proteins were transferred onto nitrocellulose membranes and probed with antibodies by Dr. Andrei Popov in the laboratory of Professor Eric Karsenti, and the complete procedure is described in[199].

4.2.5 Motor Fraction Isolation in the Presence of p50.

Taxol-stabilized microtubules were prepared as described above. The p50 experiment was completed by Dr. Andrei Popov in the laboratory of Professor Eric Karsenti, and the complete procedure is described in[199].

4.2.6 Spindle Assembly and Electron Microscopy.

The anaphase spindle was prepared from *Xenpous laevis* egg extracts and analyzed by electron microscopy. Electron micrographs were compared with rough endoplasmic

reticulum in the same field for particle comparison. This experiment was completed by Dr. Peg Coughlin in the laboratory of Professor Tim Mitchison, and the complete procedure is described in[199].

4.3 Dunaliella salina Experiments

4.3.1 Cellular fractionation

Algae were cultured in 0.5 M NaCl medium for control conditions and at 3M NaCl medium for induced conditions. Cells were fractionated into crude plasma membrane, cytoplasmic soluble, and chloroplast soluble as described in[202]. These experiments were completed by Dr. Adriana Katz in the laboratory of Professor Uri Pick.

4.3.2 Two-Dimensional (2D) PAGE

Isolated proteins were resolved by 2-D PAGE as desribed in[202]. The intensity of protein spots from the 3 M and 0.5 M control gels were compared. Spots up regulated more than 2-fold were selected for analysis by MS. These experiments were completed by Dr. Adriana Katz in the laboratory of Professor Uri Pick.

4.3.3 Mass Spectrometry Analysis of Protein Spots

Individual protein spots were manually excised from 2-D gels and in-gel digested with the protease trypsin as previously described[200]. Extracted protein digests were analyzed first by PMF on a Bruker Reflex IV matrix-assisted laser-desorption ionization time-of-flight (MALDI-TOF) mass spectrometer in reflectron mode and anchor-chip sample preparation[203]. Resulting peptide mass fingerprints were used for DB searching. Proteins unidentified by PMF were analyzed by nanoelectrospray tandem mass spectrometry on a modified MDS Sciex QSTAR Pulsar *i* quadruple time-of-flight (QqTOF) instrument, using uncoated borosilicate glass capillaries (1.2mm O.D. X 0.69mm I.D.) from Harvard Apparatus Ltd (capillaries were drawn in-house on a Sutter P-97 puller).

4.3.4 Database Searching

Peptide mass fingerprints were used for DB searching by Mascot[102] against the MSDB DB from NCBI (February, 2003), with a mass tolerance of 150 ppm; no restrictions were imposed for protein molecular weight; species set to “Green Plants”. Sets of tandem mass spectra from the analysis of unidentified proteins were first searched by Mascot against the above DB to identify proteins with peptides identical to those existing *in silico*, at a

precursor mass tolerance of 0.1 Da and fragment ion mass tolerance of 0.05 Da, as above. Mascot queries were generated from MS/MS spectra using the processing script Mascot v.1.6b2 as an extension of Bioanalyst QS software from Applied Biosystems (Foster City, CA). Mascot EST DB searching used "Other Green Plants" EST_others (Nov. 27, 2002). All tandem mass spectra were then analyzed by MS BLAST against the non-redundant nrdb protein DB at http://dove.embl-heidelberg.de/Blast2/msblast.html, as previously described[62]. Amino acid sequences were predicted with 0.1 Da tolerance for precursor masses and a 0.05 Da tolerance for fragment ions using Bioanalyst QS. Queries for MS BLAST DB searches were generated from MS/MS spectra using the ProBLAST v.1.0b11 data processing script as an extension of Bioanalyst QS[63]. In cases where scripted MS BLAST methods failed to identify a protein, an MS BLAST query was generated by manual interpretation of MS/MS spectra using Bioanalyst. If the analyzed protein remained unidentified after PMF, Mascot, and MS BLAST DB searching with queries derived automatically and manually from MS/MS data, peptide sequence tags were constructed from MS/MS spectra by manual interpretation using Bioanalyst QS for DB searching and MultiTag analysis of search results[91].

4.4 MS BLAST Specificity and Phylogenetic Analysis

4.4.1 Computing of MS BLAST Specificity and Phylogenetic Analysis.

The following analysis was performed by Bianca Habermann of MPI-CBG, Dresden. MS BLAST searches with standard settings were carried out for 1000 proteins from several model organisms (*S. cerevisiae*, *S. pombe*, *C. albicans*, *T. rubripes*, *R. norvegicus*, *M. musculus* and *H. sapiens*) using eight randomly selected peptides per DB entry. A non-redundant DB was prepared such that all sequences from the organism under analysis were omitted. Further trials using only the next closest species as a reference DB and not the total set of non-redundant proteins showed no significant increase or decrease in the success rate of identifications (data not shown). To estimate the success of MS BLAST searches, queries comprised of 8 peptides sequences were used, 10 amino acids in length and with 2 errors in each peptide to simulate ambiguities in spectrum interpretation. The top hit of each MS BLAST search was collected and tested for positive or negative identification, whereby threshold values for MS BLAST searching were calculated, essentially as published previously[62], with the exception that the reversed non-redundant DB downloaded from NCBI (release of August 2001) was used. A phylogenetic tree of a selected set of organisms from three subkingdoms was constructed based on mitochondrial small ribosomal RNA.

Multiple sequence alignments were made using the program *ClustalX*[204], and the phylogenetic trees were constructed with the programs *dnadist* and *fitch*, both from the Phylip package[205]. The estimated success rate of MS BLAST identification was correlated with phylogenetic distances between a model set of organisms and was applied to a larger set of organisms on a phylogenetic tree.

5 Publications

Shevchenko A., S. Sunyaev, A.J. Liska, P. Bork, and A. Shevchenko

Nanoelectrospray Tandem Mass Spectrometry and Sequence Similarity Searching for Identification of Proteins from Organisms with Unknown Genomes.

Methods in Molecular Biology, vol. *211*, 221-234, 2003

Sunyaev S., A.J. Liska, A. Golod, A. Shevchenko, and A. Shevchenko

MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry.

Analytical Chemistry, 75, 1307-1315, 2003

Liska A.J. and A. Shevchenko

Expanding the Organismal Scope of Proteomics: Cross-Species Protein Identification by Mass Spectrometry and its Implications.

Proteomics 3, 19-28 2003

Liska A.J. and A. Shevchenko

Combining Mass Spectrometry with Database Interrogation Strategies in Proteomics.

Trends in Analytical Chemistry 22, 291-298, 2003

Liska A.J. and A. Shevchenko

Identification of Proteins from Organisms with Unsequenced Genomes by Tandem Mass Spectrometry and Sequence-Similarity Database Searching Tools.

Cell Biology; A Laboratory Handbook (3rd Edition), Ed. Julio Celis, *in press*

Liska A.J.*, A. Popov*, S. Sunyaev, P. Coughlin, B. Habermann, A. Shevchenko, P. Bork, E. Karsenti, and A. Shevchenko

(*Equal Contribution)

Homology-Based Functional Proteomics by Mass Spectrometry: Application to the *Xenopus* Microtubule-Associated Proteome.

Molecular and Cellular Proteomics, submitted

Liska A.J., S. Sunyaev, I.N. Shilov, D.A. Schaeffer, and A. Shevchenko

Enhanced Error-Tolerant EST Database Searching by Tandem Mass Spectrometry and MultiTag Software.

Analytical Chemistry, *submitted*

Liska A.J.

Problem Selection as a Moral Dilemma in Proteomics.

Proteomics, submitted

Liska A.J.*, A. Katz*, A. Shevchenko, and U. Pick

(*Equal Contribution)

Homology-Based Proteomics by Mass Spectrometry Reveals Aspects of Salinity Adaptation in Dunaliella.

Plant Physiology, Submitted

Liska A.J., D. M. Rhoads, A. Shevchenko, and T. E. Elthon

Identification and Sequence Analysis of the Exogenous 32 kD NADH Dehydrogenase from Maize Mitochondria**.**

in preparation

Patent Application No. 0302774.5, United Kingdom: MultiTag Software

6 References

1. Blackstock, W. P.; Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins, *Trends Biotechnol*, *17*, 121-127 (1999)
2. Enard, W.; Khaitovich, P.; Klose, J.; Zollner, S.; Heissig, F.; Giavalisco, P.; Nieselt-Struwe, K.; Muchmore, E.*, et al.* Intra- and interspecific variation in primate gene expression patterns, *Science*, *296*, 340-343. (2002)
3. Mann, M.; Hendrickson, R. C.; Pandey, A. Analysis of proteins and proteomes by mass spectrometry, *Annu Rev Biochem*, *70*, 437-473 (2001)
4. Aebersold, R.; Goodlett, D. R. Mass spectrometry in proteomics, *Chem Rev*, *101*, 269-295. (2001)
5. Rubin, G. M.; Yandell, M. D.; Wortman, J. R.; Gabor Miklos, G. L.; Nelson, C. R.; Hariharan, I. K.; Fortini, M. E.; Li, P. W.*, et al.* Comparative genomics of the eukaryotes, *Science*, *287*, 2204-2215. (2000)
6. Venter, J. C. *et al.* The sequence of the human genome, *Science*, *291*, 1304-1351 (2001)
7. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature*, *409*, 860-921 (2001)
8. Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J. F.; Agarwal, P.; Agarwala, R.; Ainscough, R.*, et al.* Initial sequencing and comparative analysis of the mouse genome, *Nature*, *420*, 520-562 (2002)
9. Goff, S. A.; Ricke, D.; Lan, T. H.; Presting, G.; Wang, R.; Dunn, M.; Glazebrook, J.; Sessions, A.*, et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica), *Science*, *296*, 92-100 (2002)
10. Yu, J.; Hu, S.; Wang, J.; Wong, G. K.; Li, S.; Liu, B.; Deng, Y.; Dai, L.*, et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. indica), *Science*, *296*, 79-92 (2002)
11. Sequencing Consortium, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, *Nature*, *408*, 796-815 (2000)
12. Aparicio, S.; Chapman, J.; Stupka, E.; Putnam, N.; Chia, J. M.; Dehal, P.; Christoffels, A.; Rash, S.*, et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes, *Science*, *297*, 1301-1310. (2002)
13. Wilkins, M. R.; Williams, K. L. Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation, *J Theor Biol*, *186*, 7-15. (1997)
14. Lester, P. J.; Hubbard, S. J. Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics, *Proteomics*, *2*, 1392-1405 (2002)
15. Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching, *Anal Chem*, *71*, 2871-2882 (1999)
16. Fenyo, D. Identifying the proteome: software tools, *Curr Opin Biotechnol*, *11*, 391-395. (2000)
17. Pappin, D. J. C. Peptide mass fingerprinting using MALDI-TOF mass spectrometry, *Meth Mol Biol*, *211*, 211-217 (2002)

18. Gobom, J.; Mueller, M.; Egelhofer, V.; Theiss, D.; Lehrach, H.; Nordhoff, E. A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS, *Anal Chem*, *74*, 3915-3923. (2002)
19. Bantscheff, M.; Duempelfeld, B.; Kuster, B. An improved two-step calibration method for matrix-assisted laser desorption/ionization time-of-flight mass spectra for proteomics, *Rapid Commun Mass Spectrom*, *16*, 1892-1895 (2002)
20. Traini, M.; Gooley, A. A.; Ou, K.; Wilkins, M. R.; Tonella, L.; Sanchez, J. C.; Hochstrasser, D. F.; Williams, K. L. Towards an automated approach for protein identification in proteome projects, *Electrophoresis*, *19*, 1941-1949 (1998)
21. Egelhofer, V.; Gobom, J.; Seitz, H.; Giavalisco, P.; Lehrach, H.; Nordhoff, E. Protein identification by MALDI-TOF-MS peptide mapping: a new strategy, *Anal Chem*, *74*, 1760-1771. (2002)
22. Spengler, B. Post-source decay analysis in matrix-assisted laser desorption / ionization mass spectrometry of biomolecules, *J. Mass Spectrom*, *32*, 1019-1036 (1997)
23. Schnaible, V.; Wefing, S.; Resemann, A.; Suckau, D.; Bucker, A.; Wolf-Kummeth, S.; Hoffmann, D. Screening for disulfide bonds in proteins by MALDI in-source decay and LIFT-TOF/TOF-MS, *Anal Chem*, *74*, 4980-4988. (2002)
24. Krutchinsky, A. N.; Loboda, A. V.; Spicer, V. L.; Dworschak, R.; Ens, W.; Standing, K. G. Orthogonal injection of matrix-assisted laser desorption/ ionisation ions into a time-of-flight spectrometer through a collisional damping interface, *Rapid Commun. Mass Spectrom.*, *12*, 508-512 (1998)
25. Loboda, A. V.; Krutchinsky, A. N.; Bromirski, M.; Ens, W.; Standing, K. G. A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption / ionization source: design and performance, *Rapid Commun. Mass Spectrom.*, *14*, 1047-1057 (2000)
26. Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer., *Anal. Chem.*, *72*, 552-558 (2000)
27. Shevchenko, A.; Loboda, A.; Ens, W.; Standing, K. G. MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research, *Anal Chem*, *72*, 2132-2141. (2000)
28. Keough, T.; Youngquist, R. S.; Lacey, M. P. A method for high-sensitivity peptide sequencing using postsource decay matrix-assisted laser desorption ionization mass spectrometry, *Proc Natl Acad Sci U S A*, *96*, 7131-7136. (1999)
29. Bartet-Jones, M.; Jeffery, W. A.; Hansen, H. F.; Pappin, D. J. C. Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent, *Rapid Commun. Mass Spectrom.*, *8*, 737 - 742 (1994)
30. Krutchinsky, A. N.; Kalkum, M.; Chait, B. T. Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer, *Anal Chem*, *73*, 5066-5077. (2001)
31. Biemann, K. Contributions of mass spectrometry to peptide and protein structure, *Biomed Environ Mass Spectrom*, *16*, 99-111 (1988)
32. Shevchenko, A.; Chernushevic, I.; Wilm, M.; Mann, M. "De novo" sequencing of peptides recovered from in-gel digested proteins by nanoelectrospray tandem mass spectrometry, *Mol Biotechnol*, *20*, 107-118. (2002)

33. Wilm, M.; Mann, M. Analytical properties of the nano electrospray ion source., *Anal. Chem.*, *66*, 1-8 (1996)
34. Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. Femtomole sequencing of proteins from polyacrylamide gels by nano- electrospray mass spectrometry, *Nature*, *379*, 466-469. (1996)
35. Wilm, M.; Neubauer, G.; Mann, M. Parent ion scans of unseparated peptide mixtures, *Anal Chem*, *68*, 527-533 (1996)
36. Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M. Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode, *Anal Chem*, *73*, 1440-1448. (2001)
37. Ekroos, K.; Chernushevich, I. V.; Simons, K.; Shevchenko, A. Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer, *Anal Chem*, *74*, 941-949. (2002)
38. Geromanos, S.; Philip, J.; Freckleton, G.; Tempst, P. InJection adaptable fine ionization source ('JaFIS') for continuous flow nano-electrospray, *Rapid Commun Mass Spectrom*, *12*, 551-556 (1998)
39. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat Biotechnol*, *19*, 242-247. (2001)
40. Xiang, F.; Anderson, G. A.; Veenstra, T. D.; Lipton, M. S.; Smith, R. D. Characterization of microorganisms and biomarker development from global ESI-MS/MS analyses of cell lysates, *Anal Chem*, *72*, 2475-2481. (2000)
41. Conrads, T. P.; Alving, K.; D.Veenstra, T.; Belov, M. E.; Anderson, G. A.; Anderson, D. J.; Lipton, M. S.; Pasa-Tolic, L.*, et al.* Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ^{15}N- metabolic labeling., *Anal Chem*, *73*, 2132-2139 (2001)
42. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat Biotechnol*, *17*, 994-999 (1999)
43. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics, *Mol Cell Proteomics*, *1*, 376-386. (2002)
44. Medzihradszky, K. F.; Leffler, H.; Baldwin, M. A.; Burlingame, A. L. Protein identification by in-gel digestion, high-performance liquid chromatography, and mass spectrometry: peptide analysis by complementary ionization techniques, *J Am Soc Mass Spectrom*, *12*, 215-221. (2001)
45. Shevchenko, A.; Loboda, A.; Ens, W.; Schraven, B.; Standing, K. G.; Shevchenko, A. Archived polyarylamide gels as a resource for proteome characterization by mass spectrometry., *Electrophoresis*, *22*, 1194-1203 (2001)
46. Krutchinsky, A. N.; Zhang, W.; Chait, B. T. Rapidly switchable matrix-assisted laser desorption/ionization and electrospray quadrupole-time-of-flight mass spectrometry for protein identification., *J Am Soc Mass Spectrom*, *11*, 493-504 (2000)
47. Baykut, G.; Fuchser, J.; Witt, M.; Weiss, G.; Gosteli, C. A combined ion source for fast switching between electrospray and matrix-assisted laser desorption/ionization in Fourier transform ion cyclotron resonance mass spectrometry, *Rapid Commun Mass Spectrom*, *16*, 1631-1641 (2002)

48. Chan, T. W.; Duan, L.; Sze, T. P. Accurate mass measurements for peptide and protein mixtures by using matrix-assisted laser desorption/ionization Fourier transform mass spectrometry, *Anal Chem*, *74*, 5282-5289. (2002)

49. Ge, Y.; Lawhorn, B. G.; ElNaggar, M.; Strauss, E.; Park, J. H.; Begley, T. P.; McLafferty, F. W. Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry, *J Am Chem Soc*, *124*, 672-678. (2002)

50. Schwartz, J. C.; Senko, M. W.; Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer, *J Am Soc Mass Spectrom*, *13*, 659-669. (2002)

51. Collings, B. A.; Campbell, J. M.; Mao, D.; Douglas, D. J. A combined linear ion trap time-of-flight system with improved performance and MS(n) capabilities, *Rapid Commun Mass Spectrom*, *15*, 1777-1795 (2001)

52. Shevchenko, A.; Chernushevich, I.; Wilm, M.; Mann, M. De Novo peptide sequencing by nanoelectrospray tandem mass spectrometry using triple quadrupole and quadrupole - time -of-flight instruments, *Meth Mol Biol*, *146*, 1-16 (2000)

53. Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; Haller, P.*, et al.* Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation, *Rapid Commun Mass Spectrom*, *15*, 1214-1221 (2001)

54. Shevchenko, A.; Keller, P.; Scheiffele, P.; Mann, M.; Simons, K. Identification of components of trans-Golgi network-derived transport vesicles and detergent-insoluble complexes by nanoelectrospray tandem mass spectrometry., *Electrophoresis*, *18*, 2591-2600 (1997)

55. Uttenweiler-Joseph, S.; Neubauer, G.; Christoforidis, S.; Zerial, M.; Wilm, M. Automated de novo sequencing of proteins using the differential scanning technique, *Proteomics*, *1*, 668-682. (2001)

56. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry, *J Comput Biol*, *6*, 327-342 (1999)

57. Taylor, J. A.; Johnson, R. S. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry, *Anal Chem*, *73*, 2594-2604. (2001)

58. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, *25*, 3389-3402. (1997)

59. Pearson, W. R. Flexible sequence similarity searching with the FASTA3 program package, *Methods Mol Biol*, *132*, 185-219 (2000)

60. Huang, L.; Jacob, R. J.; Pegg, S. C.; Baldwin, M. A.; Wang, C. C.; Burlingame, A. L.; Babbitt, P. C. Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches, *J Biol Chem*, *276*, 28327-28339. (2001)

61. Mackey, A. J.; Haystead, T. A. J.; Pearson, W. R. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences, *Mol Cell Proteomics*, *1*, 139-147 (2002)

62. Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. Charting the proteomes of organisms with unsequenced genomes by MALDI- quadrupole time-of-flight mass spectrometry and BLAST homology searching, *Anal Chem*, *73*, 1917-1926. (2001)

63. Nimkar, S.; Loo, J. A., Orlando FL 2002; Abstract 334.

64. Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.*, et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, *415*, 141-147. (2002)

65. Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.*, et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature*, *415*, 180-183. (2002)

66. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists, *Cell*, *92*, 291-294 (1998)

67. Ibba, M.; Soll, D. Aminoacyl-tRNA synthesis, *Annu Rev Biochem*, *69*, 617-650 (2000)

68. Deshaies, R. J. SCF and Cullin/Ring H2-based ubiquitin ligases, *Annu Rev Cell Dev Biol*, *15*, 435-467 (1999)

69. Lyapina, S.; Cope, G.; Shevchenko, A.; Serino, G.; Tsuge, T.; Zhou, C.; Wolf, D. A.; Wei, N.*, et al.* Promotion of NEDD8-CUL1 conjugate cleavage by COP9 signalosome, *Science*, *292*, 1382-1385 (2001)

70. Seol, J. H.; Shevchenko, A.; Shevchenko, A.; Deshaies, R. J. Skp1 forms multiple protein complexes, including RAVE, a regulator of V- ATPase assembly, *Nat Cell Biol*, *3*, 384-391. (2001)

71. Wei, N.; Deng, X. Making sense of the COP9 signalosome. A regulatory protein complex conserved from Arabidopsis to human, *Trends Genet*, *15*, 98-103 (1999)

72. Cope, G. A.; Deshaies, R. J. COP9 signalosome: a multifunctional regulator of SCF and other cullin-based ubiquitin ligases, *Cell*, *114*, 663-671 (2003)

73. Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.*, et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae., *Nature*, *403*, 623-627 (2000)

74. Mathe, C.; Sagot, M. F.; Schiex, T.; Rouze, P. Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Res*, *30*, 4103-4117. (2002)

75. Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.*, et al.* Complementary DNA sequencing: expressed sequence tags and human genome project, *Science*, *252*, 1651-1656. (1991)

76. Langen, H.; Gray, C.; Roder, D.; Juranville, J. F.; Takacs, B.; Fountoulakis, M. From genome to proteome: protein map of Haemophilus influenzae, *Electrophoresis*, *18*, 1184-1192. (1997)

77. Chang, W. W.; Huang, L.; Shen, M.; Webster, C.; Burlingame, A. L.; Roberts, J. K. Patterns of protein synthesis and tolerance of anoxia in root tips of maize seedlings acclimated to a low-oxygen environment, and identification of proteins by mass spectrometry, *Plant Physiol*, *122*, 295-318. (2000)

78. Porubleva, L.; Vander Velden, K.; Kothari, S.; Oliver, D. J.; Chitnis, P. R. The proteome of maize leaves: use of gene sequences and expressed sequence tag data for identification of proteins with peptide mass fingerprints, *Electrophoresis*, *22*, 1724-1738. (2001)

79. Peltier, J. B.; Friso, G.; Kalume, D. E.; Roepstorff, P.; Nilsson, F.; Adamska, I.; van Wijk, K. J. Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins, *Plant Cell*, *12*, 319-341 (2000)

80. Gomez, S. M.; Nishio, J. N.; Faull, K. F.; Whitelegge, J. P. The chloroplast grana proteome defined by intact mass measurements from LC-MS., *Mol Cell Proteomics*, *1*, 46-59 (2002)

81. Decker, G.; Wanner, G.; Zenk, M. H.; Lottspeich, F. Characterization of proteins in latex of the opium poppy (Papaver somniferum) using two-dimensional gel electrophoresis and microsequencing, *Electrophoresis*, *21*, 3500-3516. (2000)

82. Kuster, B.; Mortensen, P.; Andersen, J. S.; Mann, M. Mass spectrometry allows direct identification of proteins in large genomes, *Proteomics*, *1*, 641-650. (2001)

83. Koc, E. C.; Burkhart, W.; Blackburn, K.; Moseley, A.; Koc, H.; Spremulli, L. L. A proteomics approach to the identification of mammalian mitochondrial small subunit ribosomal proteins, *J Biol Chem*, *275*, 32585-32591. (2000)

84. Tournebize, R.; Popov, A.; Kinoshita, K.; Ashford, A. J.; Rybina, S.; Pozniakovsky, A.; Mayer, T. U.; Walczak, C. E.*, et al.* Control of microtubule dynamics by the antagonistic activities of XMAP215 and XKCM1 in Xenopus egg extracts, *Nat Cell Biol*, *2*, 13-19. (2000)

85. 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, *Meth Mol Biol*, *211*, 221-234 (2003)

86. Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags, *Anal Chem*, *66*, 4390-4399. (1994)

87. Andersen, J. S.; Lyon, C. E.; Fox, A. H.; Leung, A. K.; Lam, Y. W.; Steen, H.; Mann, M.; Lamond, A. I. Directed proteomic analysis of the human nucleolus, *Curr Biol*, *12*, 1-11. (2002)

88. Neubauer, G.; King, A.; Rappsilber, J.; Calvio, C.; Watson, M.; Ajuh, P.; Sleeman, J.; Lamond, A.*, et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex, *Nat Genet*, *20*, 46-50. (1998)

89. Feller, W. An introduction to Probability Theory and its Applications, *John Wiley & Sons, Inc.*, *vol. II* (1966)

90. Baudouin-Cornu, P.; Surdin-Kerjan, Y.; Marliere, P.; Thomas, D. Molecular evolution of protein atomic composition, *Science*, *293*, 297-300. (2001)

91. Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A.; Shevchenko, A. MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry, *Anal Chem*, *75*, 1307-1315 (2003)

92. Blackshear, P. J.; Lai, W. S.; Thorn, J. M.; Kennington, E. A.; Staffa, N. G.; Moore, D. T.; Bouffard, G. G.; Beckstrom-Sternberg, S. M.*, et al.* The NIEHS Xenopus maternal EST project: interim analysis of the first 13,879 ESTs from unfertilized eggs, *Gene*, *267*, 71-87. (2001)

93. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, *Anal Chem*, *74*, 5383-5392 (2002)

94. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry, *Anal Chem*, *75*, 4646-4658 (2003)

95. Dreger, M. Proteome analysis at the level of subcellular structures, *Eur J Biochem*, *270*, 589-599. (2003)

96. Jung, E.; Heller, M.; Sanchez, J. C.; Hochstrasser, D. F. Proteomics meets cell biology: the establishment of subcellular proteomes, *Electrophoresis*, *21*, 3369-3377. (2000)

97. Murray, A. W.; Kirschner, M. W. Cyclin synthesis drives the early embryonic cell cycle, *Nature*, *339*, 275-280. (1989)

98. Cassimeris, L.; Spittle, C. Regulation of microtubule-associated proteins, *Int Rev Cytol*, *210*, 163-226 (2001)

99. Lohka, M. J.; Maller, J. L. Induction of nuclear envelope breakdown, chromosome condensation, and spindle formation in cell-free extracts, *J Cell Biol*, *101*, 518-523. (1985)

100. Karsenti, E.; Vernos, I. The mitotic spindle: a self-made machine, *Science*, *294*, 543-547. (2001)

101. Graf, J. D.; Kobel, H. R. Genetics of Xenopus laevis, *Methods Cell Biol*, *36*, 19-34 (1991)

102. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, *20*, 3551-3567 (1999)

103. Creasy, D. M.; Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data, *Proteomics*, *2*, 1426-1434. (2002)

104. Minella, O.; Mulner-Lorillon, O.; Poulhe, R.; Belle, R.; Cormier, P. The guanine-nucleotide-exchange complex (EF-1 beta gamma delta) of elongation factor-1 contains two similar leucine-zipper proteins EF-1 delta, p34 encoded by EF-1 delta 1 and p36 encoded by EF-1 delta 2, *Eur J Biochem*, *237*, 685-690. (1996)

105. Mikami, A.; Paschal, B. M.; Mazumdar, M.; Vallee, R. B. Molecular cloning of the retrograde transport motor cytoplasmic dynein (MAP 1C), *Neuron*, *10*, 787-796. (1993)

106. Andersen, S. S.; Karsenti, E. XMAP310: a Xenopus rescue-promoting factor localized to the mitotic spindle, *J Cell Biol*, *139*, 975-983. (1997)

107. Andersen, S. S.; Buendia, B.; Dominguez, J. E.; Sawyer, A.; Karsenti, E. Effect on microtubule dynamics of XMAP230, a microtubule-associated protein present in Xenopus laevis eggs and dividing cells, *J Cell Biol*, *127*, 1289-1299. (1994)

108. Vernos, I.; Heasman, J.; Wylie, C. Multiple kinesin-like transcripts in Xenopus oocytes, *Dev Biol*, *157*, 232-239. (1993)

109. Houliston, E.; Le Guellec, R.; Kress, M.; Philippe, M.; Le Guellec, K. The kinesin-related protein Eg5 associates with both interphase and spindle microtubules during Xenopus early development, *Dev Biol*, *164*, 147-159. (1994)

110. Sawin, K. E.; LeGuellec, K.; Philippe, M.; Mitchison, T. J. Mitotic spindle organization by a plus-end-directed microtubule motor, *Nature*, *359*, 540-543. (1992)

111. Gudkov, A. V.; Kazarov, A. R.; Thimmapaya, R.; Axenovich, S. A.; Mazo, I. A.; Roninson, I. B. Cloning mammalian genes by expression selection of genetic suppressor elements: association of kinesin with drug resistance and cell immortalization, *Proc Natl Acad Sci U S A*, *91*, 3744-3748. (1994)

112. Nagase, T.; Ishikawa, K.; Miyajima, N.; Tanaka, A.; Kotani, H.; Nomura, N.; Ohara, O. Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro, *DNA Res*, *5*, 31-39. (1998)

113. Hamill, D. R.; Howell, B.; Cassimeris, L.; Suprenant, K. A. Purification of a WD repeat protein, EMAP, that promotes microtubule dynamics through an inhibition of rescue, *J Biol Chem*, *273*, 9285-9291. (1998)

114. De Marco, V.; Burkhard, P.; Le Bot, N.; Vernos, I.; Hoenger, A. Analysis of heterodimer formation by Xklp3A/B, a newly cloned kinesin- II from Xenopus laevis, *Embo J*, *20*, 3370-3379. (2001)

115. Zhang, J.; Han, G.; Xiang, X. Cytoplasmic dynein intermediate chain and heavy chain are dependent upon each other for microtubule end localization in Aspergillus nidulans, *Mol Microbiol*, *44*, 381-392. (2002)

116. Smith, D. J. The complete sequence of a frog alpha-tubulin gene and its regulated expression in mouse L-cells, *Biochem J*, *249*, 465-472. (1988)

117. Good, P. J.; Richter, K.; Dawid, I. B. The sequence of a nervous system-specific, class II beta-tubulin gene from Xenopus laevis, *Nucleic Acids Res*, *17*, 8000. (1989)

118. Assmann, V.; Jenkinson, D.; Marshall, J. F.; Hart, I. R. The intracellular hyaluronan receptor RHAMM/IHABP interacts with microtubules and actin filaments, *J Cell Sci*, *112*, 3943-3954. (1999)

119. Tsukiyama, T.; Palmer, J.; Landel, C. C.; Shiloach, J.; Wu, C. Characterization of the imitation switch subfamily of ATP-dependent chromatin-remodeling factors in Saccharomyces cerevisiae, *Genes Dev*, *13*, 686-697. (1999)

120. Trachtulcova, P.; Janatova, I.; Kohlwein, S. D.; Hasek, J. Saccharomyces cerevisiae gene ISW2 encodes a microtubule-interacting protein required for premeiotic DNA replication, *Yeast*, *16*, 35-47. (2000)

121. Earle, E.; Saxena, A.; MacDonald, A.; Hudson, D. F.; Shaffer, L. G.; Saffery, R.; Cancilla, M. R.; Cutts, S. M.*, et al.* Poly(ADP-ribose) polymerase at active centromeres and neocentromeres at metaphase, *Hum Mol Genet*, *9*, 187-194. (2000)

122. Liang, P.; MacRae, T. H. Molecular chaperones and the cytoskeleton, *J Cell Sci*, *110*, 1431-1440. (1997)

123. Lange, B. M.; Bachi, A.; Wilm, M.; Gonzalez, C. Hsp90 is a core centrosomal component and is required at different stages of the centrosome cycle in Drosophila and vertebrates, *Embo J*, *19*, 1252-1262. (2000)

124. Etkin, L. D.; el-Hodiri, H. M.; Nakamura, H.; Wu, C. F.; Shou, W.; Gong, S. G. Characterization and function of Xnf7 during early development of Xenopus, *J Cell Physiol*, *173*, 144-146. (1997)

125. Bashour, A. M.; Bloom, G. S. 58K, a microtubule-binding Golgi protein, is a formiminotransferase cyclodeaminase, *J Biol Chem*, *273*, 19612-19617. (1998)

126. Robertson, K.; Hensey, C.; Gautier, J. Isolation and characterization of Xenopus ATM (X-ATM): expression, localization, and complex formation during oogenesis and early development, *Oncogene*, *18*, 7070-7079. (1999)

127. Hekmat-Nejad, M.; You, Z.; Yee, M. C.; Newport, J. W.; Cimprich, K. A. Xenopus ATR is a replication-dependent chromatin-binding protein required for the DNA replication checkpoint, *Curr Biol*, *10*, 1565-1573. (2000)

128. Chung, D. W.; Zhang, J. A.; Tan, C. K.; Davie, E. W.; So, A. G.; Downey, K. M. Primary structure of the catalytic subunit of human DNA polymerase delta and chromosomal location of the gene, *Proc Natl Acad Sci U S A*, *88*, 11197-11201. (1991)

129. Ofulue, E. N.; Candido, E. P. Molecular cloning and characterization of the Caenorhabditis elegans elongation factor 2 gene (eft-2), *DNA Cell Biol*, *10*, 603-611. (1991)

130. Heine, H.; Delude, R. L.; Monks, B. G.; Espevik, T.; Golenbock, D. T. Bacterial lipopolysaccharide induces expression of the stress response genes hop and H411, *J Biol Chem*, *274*, 21049-21055. (1999)

131. Wormington, W. M. Developmental expression and 5S rRNA-binding activity of Xenopus laevis ribosomal protein L5, *Mol Cell Biol*, *9*, 5281-5288. (1989)

132. Amaldi, F.; Beccari, E.; Bozzoni, I.; Luo, Z. X.; Pierandrei-Amaldi, P. Nucleotide sequences of cloned cDNA fragments specific for six Xenopus laevis ribosomal proteins, *Gene*, *17*, 311-316. (1982)

133. Kwon, H. J.; Bae, S.; Son, Y. H.; Chung, H. M. Expression of the Xenopus homologue of the receptor for activated C- kinase 1 (RACK1) in the Xenopus embryo, *Dev Genes Evol*, *211*, 195-197. (2001)

134. Wittmann, T.; Hyman, T. Recombinant p50/dynamitin as a tool to examine the role of dynactin in intracellular processes, *Methods Cell Biol*, *61*, 137-143 (1999)

135. Mack, G. J.; Compton, D. A. Analysis of mitotic microtubule-associated proteins using mass spectrometry identifies astrin, a spindle-associated protein, *Proc Natl Acad Sci U S A*, *98*, 14434-14439. (2001)

136. Popov, A. V.; Severin, F.; Karsenti, E. XMAP215 is required for the microtubule-nucleating activity of centrosomes, *Curr Biol*, *12*, 1326-1330. (2002)

137. Adams, I. R.; Kilmartin, J. V. Localization of core spindle pole body (SPB) components during SPB duplication in Saccharomyces cerevisiae, *J Cell Biol*, *145*, 809-823. (1999)

138. Sang Lee, J.; Gyu Park, S.; Park, H.; Seol, W.; Lee, S.; Kim, S. Interaction network of human aminoacyl-tRNA synthetases and subunits of elongation factor 1 complex, *Biochem Biophys Res Commun*, *291*, 158-164. (2002)

139. Suprenant, K. A.; Tempero, L. B.; Hammer, L. E. Association of ribosomes with in vitro assembled microtubules, *Cell Motil Cytoskeleton*, *14*, 401-415 (1989)

140. Sciortino, S.; Gurtner, A.; Manni, I.; Fontemaggi, G.; Dey, A.; Sacchi, A.; Ozato, K.; Piaggio, G. The cyclin B1 gene is actively transcribed during mitosis in HeLa cells, *EMBO Rep*, *2*, 1018-1023. (2001)

141. Hochegger, H.; Klotzbucher, A.; Kirk, J.; Howell, M.; le Guellec, K.; Fletcher, K.; Duncan, T.; Sohail, M*., et al.* New B-type cyclin synthesis is required between meiosis I and II during Xenopus oocyte maturation, *Development*, *128*, 3795-3807. (2001)

142. Perez, L. H.; Antonio, C.; Flament, S.; Vernos, I.; Nebreda, A. R. Xkid chromokinesin is required for the meiosis I to meiosis II transition in Xenopus laevis oocytes, *Nat Cell Biol*, *4*, 737-742. (2002)

143. Groisman, I.; Huang, Y. S.; Mendez, R.; Cao, Q.; Theurkauf, W.; Richter, J. D. CPEB, maskin, and cyclin B1 mRNA at the mitotic apparatus: implications for local translational control of cell division, *Cell*, *103*, 435-447. (2000)

144. Mayfield, J. A.; Fiebig, A.; Johnstone, S. E.; Preuss, D. Gene families from the Arabidopsis thaliana pollen coat proteome, *Science*, *292*, 2482-2485 (2001)

145. Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hays, L.; Schieltz, D*., et al.* Proteomic survey of metabolic pathways in rice, *Proc Natl Acad Sci U S A*, *99*, 11969-11974 (2002)

146. Amiour, N.; Merlino, M.; Leroy, P.; Branlard, G. Proteomic analysis of amphiphilic proteins of hexaploid wheat kernels, *Proteomics*, *2*, 632-641 (2002)

147. Eng, J.; McCormack, A.; Yates, J., *Journal of the American Society of Mass Spectrometry*, *5*, 976-989 (1994)

148. van Wijk, K. J. Challenges and prospects of plant proteomics, *Plant Physiol*, *126*, 501-508. (2001)

149. Hasegawa, P. M.; Bressan, R. A.; Zhu, J.; Bohnert, H. J. Plant cellular and molecular responses to high salinity, *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, *51*, 463-499 (2000)

150. Gong, Z.; Koiwa, H.; Cushman, M. A.; Ray, A.; Bufford, D.; Kore, E. S.; Matsumoto, T. K.; Zhu, J.*, et al.* Genes that are uniquely stress regulated in salt overly sensitive (sos) mutants, *Plant Physiol*, *126*, 363-375. (2001)

151. Kawasaki, S.; Borchert, C.; Deyholos, M.; Wang, H.; Brazille, S.; Kawai, K.; Galbraith, D.; Bohnert, H. J. Gene expression profiles during the initial phase of salt stress in rice, *Plant Cell*, *13*, 889-905 (2001)

152. Kreps, J. A.; Wu, Y.; Chang, H. S.; Zhu, T.; Wang, X.; Harper, J. F. Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress, *Plant Physiol*, *130*, 2129-2141 (2002)

153. Seki, M.; Ishida, J.; Narusaka, M.; Fujita, M.; Nanjo, T.; Umezawa, T.; Kamiya, A.; Nakajima, M.*, et al.* Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray, *Funct Integr Genomics*, *2*, 282-291 (2002)

154. Pick, U. In *Salinity: Enviroment-Plants Molecules,* Acad. Pub.Dordrecht ed.; Lauchli, A., Luthge, U., Eds.; Kluwer, 2002, pp 97-112.

155. Fisher, M.; Gokhman, I.; Pick, U.; Zamir, A. A salt-resistant plasma membrane carbonic anhydrase is induced by salt in Dunaliella salina, *J Biol Chem*, *271*, 17718-17723 (1996)

156. Pick, U. Dunaliella - A model extremophilic alga, *Israel Journal of Plant Sciences*, *46*, 131-139 (1998)

157. Santoni, V.; Molloy, M.; Rabilloud, T. Membrane proteins and proteomics: un amour impossible?, *Electrophoresis*, *21*, 1054-1070 (2000)

158. Wingler, A.; Lea, P. J.; Quick, W. P.; Leegood, R. C. Photorespiration: metabolic pathways and their role in stress protection, *Philos Trans R Soc Lond B Biol Sci*, *355*, 1517-1529 (2000)

159. Mori, S.; Castoreno, A.; Lammers, P. J. Transcript levels of rbcR1, ntcA, and rbcL/S genes in cyanobacterium Anabaena sp. PCC 7120 are downregulated in response to cold and osmotic stress, *FEMS Microbiol Lett*, *213*, 167-173 (2002)

160. Andreishcheva, E. N.; Zviagil'skaia, R. A. Adaptation of yeasts to salt stress, *Prikl Biokhim Mikrobiol*, *35*, 243-256 (1999)

161. Blomberg, A. Osmoresponsive proteins and functional assessment strategies in Saccharomyces cerevisiae, *Electrophoresis*, *18*, 1429-1440 (1997)

162. Chen, R. H.; Miettinen, P. J.; Maruoka, E. M.; Choy, L.; Derynck, R. A WD-domain protein that is associated with and phosphorylated by the type II TGF-beta receptor, *Nature*, *377*, 548-552 (1995)

163. Jiang, J.; Clouse, S. D. Expression of a plant gene with sequence similarity to animal TGF-beta receptor interacting protein is regulated by brassinosteroids and required for normal plant development, *Plant J*, *26*, 35-45 (2001)

164. Dunand-Sauthier, I.; Walker, C.; Wilkinson, C.; Gordon, C.; Crane, R.; Norbury, C.; Humphrey, T. Sum1, a component of the fission yeast eIF3 translation initiation complex, is rapidly relocalized during environmental stress and interacts with components of the 26S proteasome, *Mol Biol Cell*, *13*, 1626-1640 (2002)

165. Caldas, T. D.; El Yaagoubi, A.; Richarme, G. Chaperone properties of bacterial elongation factor EF-Tu, *J Biol Chem*, *273*, 11478-11482 (1998)

166. Caldas, T.; Laalami, S.; Richarme, G. Chaperone properties of bacterial elongation factor EF-G and initiation factor IF2, *J Biol Chem*, *275*, 855-860 (2000)

167. Sadka, A.; Himmmelhoch, S.; Zamir, A. A 150 kilodalton cell-surface protein is induced by salt in the halotolerant green-alga Dunaliella salina, *Plant Physiol*, *95*, 822-831 (1991)

168. Fisher, M.; Gokhman, I.; Pick, U.; Zamir, A. A structurally novel transferrin-like protein accumulates in the plasma membrane of the unicellular green alga Dunaliella salina grown in high salinities, *J Biol Chem*, *272*, 1565-1570 (1997)

169. Hayashi, M.; Hirai, K.; Unemoto, T. Sequencing and the Alignment of Structural Genes in the Nqr Operon Encoding the Na+-Translocating Nadh-Quinone Reductase From Vibrio-Alginolyticus, *Febs Letters*, *363*, 75-77 (1995)

170. Katz, A.; Pick, U. Plasma membrane electron transport coupled to Na+ extrusion in the halotolerant alga Dunaliella, *Biochimica Et Biophysica Acta-Bioenergetics*, *1504*, 423-431 (2001)

171. Rost, B. Enzyme function less conserved than anticipated, *J Mol Biol*, *318*, 595-608. (2002)

172. Suckau, D.; Resemann, A.; Schuerenberg, M.; Hufnagel, P.; Franzen, J.; Holle, A. A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics, *Anal Bioanal Chem*, *376*, 952-965 (2003)

173. Gewolb, J. Genomics. Animals line up to be sequenced, *Science*, *293*, 409-410. (2001)

174. Gewolb, J. Genome research. DNA sequencers to go bananas?, *Science*, *293*, 585-586. (2001)

175. Muse, S. V. Examining rates and patterns of nucleotide substitution in plants, *Plant Mol Biol*, *42*, 25-43. (2000)

176. Iliopoulos, I.; Tsoka, S.; Andrade, M. A.; Janssen, P.; Audit, B.; Tramontano, A.; Valencia, A.; Leroy, C.*, et al.* Genome sequences and great expectations, *Genome Biol*, *2* (2001)

177. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases, *Proc Natl Acad Sci U S A*, *90*, 5011-5015. (1993)

178. Mann, M.; Hojrup, P.; Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases, *Biol Mass Spectrom*, *22*, 338-345. (1993)

179. Yates, J. R., 3rd; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification, *Anal Biochem*, *214*, 397-408. (1993)

180. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein identification by mass profile fingerprinting, *Biochem Biophys Res Commun*, *195*, 58-64. (1993)

181. Gay, S.; Binz, P. A.; Hochstrasser, D. F.; Appel, R. D. Peptide mass fingerprinting peak intensity prediction: Extracting knowledge from spectra, *Proteomics*, *2*, 1374-1391. (2002)

182. Parker, K. C. Scoring methods in MALDI peptide mass fingerprinting: ChemScore, and the ChemApplex program, *J Am Soc Mass Spectrom*, *13*, 22-39. (2002)

183. Spengler, B.; Luetzenkirchen, F.; Metzger, S.; Chaurand, P.; Kaufmann, R.; Jeffery, W.; Bartlet-Jones, M.; Pappin, D. Peptide sequencing of charged derivatives by postsource decay MALDI mass spectrometry, *Int J Mass Spectrom*, *169*, 127-140 (1997)

184. Choudhary, J. S.; Blackstock, W. P.; Creasy, D. M.; Cottrell, J. S. Interrogating the human genome using uninterpreted mass spectrometry data, *Proteomics*, *1*, 651-667. (2001)

185. Marina, A.; Garcia, M. A.; Albar, J. P.; Yague, J.; Lopez de Castro, J. A.; Vazquez, J. High-sensitivity analysis and sequencing of peptides and proteins by quadrupole ion trap mass spectrometry, *J Mass Spectrom*, *34*, 17-27. (1999)

186. Huang, P.; Wall, D. B.; Parus, S.; Lubman, D. M. On-line capillary liquid chromatography tandem mass spectrometry on an ion trap/reflectron time-of-flight mass spectrometer using the sequence tag database search approach for peptide sequencing and protein identification, *J Am Soc Mass Spectrom*, *11*, 127-135. (2000)

187. Bafna, V.; Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database, *Bioinformatics*, *17*, S13-21. (2001)

188. Field, H. I.; Fenyo, D.; Beavis, R. C. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database, *Proteomics*, *2*, 36-47. (2002)

189. Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data, *Proteomics*, *2*, 1406-1412. (2002)

190. Choudhary, J. S.; Blackstock, W. P.; Creasy, D. M.; Cottrell, J. S. Matching peptide mass spectra to EST and genomic DNA databases, *Trends Biotechnol*, *19*, S17-22. (2001)

191. Nielsen, M. L.; Bennet, K. L.; Larsen, B.; Moniatte, M.; Mann, M. Peptide End Sequencing by Orthogonal MALDI Tandem Mass Spectrometry, *Journal of Proteome Research*, *1*, 63-71 (2002)

192. Schlosser, A.; Lehmann, W. D. Patchwork peptide sequencing: extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution, *Proteomics*, *2*, 524-533. (2002)

193. Qin, J.; Herring, C. J.; Zhang, X. De novo peptide sequencing in an ion trap mass spectrometer with 18O labeling, *Rapid Commun Mass Spectrom*, *12*, 209-216 (1998)

194. Wattenberg, A.; Organ, A. J.; Schneider, K.; Tyldesley, R.; Bordoli, R.; Bateman, R. H. Sequence dependent fragmentation of peptides generated by MALDI quadrupole time-of-flight (MALDI Q-TOF) mass spectrometry and its implications for protein identification, *J Am Soc Mass Spectrom*, *13*, 772-783. (2002)

195. MacCoss, M. J.; Wu, C. C.; Yates, J. R., 3rd Probability-based validation of protein identifications using a modified SEQUEST algorithm, *Anal Chem*, *74*, 5593-5599. (2002)

196. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, *Anal Chem*, *74*, 5383-5392. (2002)

197. Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Boucherie, H.*, et al.* Linking genome and proteome by mass spectrometry: large-scale

identification of yeast proteins from two dimensional gels, *Proc Natl Acad Sci U S A*, *93*, 14440-14445. (1996)

198. Saalbach, G.; Erik, P.; Wienkoop, S. Characterisation by proteomics of peribacteroid space and peribacteroid membrane preparations from pea (Pisum sativum) symbiosomes, *Proteomics*, *2*, 325-337. (2002)
199. Liska, A. J.; Popov, A.; Sunyaev, S.; Shevchenko, A.; Habermann, B.; Bork, P.; Karenti, E.; Shevchenko, A. Homology-Based Proteomics by Tandem Mass Spectrometry: Application the Xenopus Microtubule-Associated Proteome, *submitted* (2003)
200. Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels., *Anal. Chem.*, *68*, 850-858 (1996)
201. Lasek, R. J.; Brady, S. T. Attachment of transported vesicles to microtubules in axoplasm is facilitated by AMP-PNP, *Nature*, *316*, 645-647. (1985)
202. Liska, A.; Katz, A.; Shevchenko, A.; Pick, U. Homology-Based Proteomics by Mass Spectrometry Reveals Aspects of Salinity Adaptation of Dunaliella, *Plant Physiol*, *submitted* (2003)
203. Havlis, J.; Thomas, H.; Sebela, M.; Shevchenko, A. Fast-response proteomics by accelerated in-gel digestion of proteins, *Anal Chem*, *75*, 1300-1306 (2003)
204. Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res*, *25*, 4876-4882. (1997)
205. Felsenstein, J. *Distributed by the author*, 1993.

Declaration concerning the PhD thesis "**Homology-Based Functional Proteomics By Mass Spectrometry Advanced Informatic Methods**" submitted by Adam J. Liska.

Herein, I declare that I have produced this manuscript without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This manuscript has not previously been presented in identical or similar form to any other German or foreign examination board. Experimental work performed by collaborators is indicated as such.

The thesis work was conducted from July 2001 to October 2003 under the supervision of Dr. Andrej Shevchenko at the *Max Planck Institute of Molecular Cell Biology and Genetics* in the biological mass spectrometry laboratory.

I declare that I have not undertaken any previous unsuccessful doctorate proceedings.

I declare that I recognize the doctorate regulations of the Faculty of Sciences of the Technische Universität Dresden.

Adam J. Liska

Erklärung

Die vorliegende Arbeit wurde im Zeitraum von Juli 2001 bis October 2003 am Max Plank Institut für Molekulare Zellbiologie und Genetik Dresden unter der wissenschaftlichen Betreuung von Dr. Andrei Shevchenko angefertigt. Es haben keine früheren erfolglosen Promotionsversuche stattgefunden. Die Promotionsordnung der Fakultät Mathematik und Naturwissenschaften der TU Dresden vom 20. März 2000 erkenne ich an.

Dresden, den 10. October 2003

Adam J. Liska

Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Dresden, den 10. October 2003

Adam J. Liska