3

Evolutionary Proteomics: Empowering Tandem Mass Spectrometry and Bioinformatics Tools for the Study of Evolution

Irving E. Vega[1], Dan Rittschof[2], Gary H. Dickinson[3] and Ian Musgrave[4]
[1]*Department of Biology and Protein Mass Spectrometry Core Facility, College of Natural Science, University of Puerto Rico – Río Piedras Campus, Puerto Rico,*
[2]*MSC Division, Nicholas School and Biology Department Duke University,*
[3]*Department of Oral Biology, University of Pittsburgh School of Dental Medicine,*
[4]*Discipline of Pharmacology, School of Medical Sciences, University of Adelaide, Adelaide,*
[1,2,3]*USA*
[4]*Australia*

1. Introduction

Darwin's and Wallace's evolution hypotheses had as their basis the survival of the fittest. The morphological characteristics that an organism displays are acted upon by selective forces, which control the likelihood that those particular characteristics will be transferred to the next generation. Morphological characteristics observed by scientists in the 19th century were understood by their successors in the 20th century as a phenotypic representation of the complex expression pattern of an organism's genome. Today, we recognize this fundamental concept as the interaction between genome and environment (Qui & Cho, 2008; Dick, 2011; Tzeng et al., 2011). Phenotype is environmentally influenced through molecular changes that begin with gene transcription (Kappeler & Meaney, 2010). Genome complexity, organized and established during embryonic development, is the stage where divergence between species begins (Wang et al., 2010).

Embryology is the intersection between evolutionary and developmental biology (Goodman & Coughlin, 2000). Developmental biology focuses on questions of how genetics control cell growth, differentiation and, ultimately, morphogenesis. For developmental biologists, the timing and regulation of genome expression is crucial to achieving differentiation. In contrast, how biological diversity is transformed through time is a key question for evolutionary biologists. Nevertheless, both disciplines base their theoretical concepts on the expression patterns of the genome.

Genomics as an approach ushered in the beginning of a new scientific era; a logical progression is using genomic approaches to study diversity at the organismal level. Significantly, shortly after completion of the human genome project, the scientific community discovered that only 2% of the human genome is composed of functional genes (Human Genome Project [HGP], 2008). It is fascinating that ~80% of the 20,000 to 25,000 genes in human DNA can be found in invertebrate genomes (Prachumwat & Li, 2008), suggesting that just a small set of genes is responsible for observed species diversity. Thus, a

series of questions arise in the genomics era: what makes us different from other organisms? If all organisms have a largely common set of genes, which genes are responsible for, or susceptible to molecular evolution? What molecular mechanisms maintain biological diversity among species?

Insights from the genomics era illustrated that the genome is like an orchestra. The orchestra uses the same instruments (genes) to generate different songs (organisms) that are distinct and unique. But a song can be modified from one music genre to another just by changes in tempo (phenotype) without losing its complete identity. Thus, the human's song comes in hundreds of different tempos without losing its identity, while the human song is very different from the mouse's song. Some scientists believe that epigenetics, factors causing differential gene expression, is the molecular mechanism that explains these variations in "tempo". After all, it has been demonstrated that epigenetic changes induce differential expression of genetic material and contributes to biological diversity (for examples see Levin & Moran, 2011; Day & Sweatt, 2011). However, changes in gene expression alone cannot fully explain diversity among species when most genes are shared. Thus, changes at the protein level are likely to play a role in inducing or maintaining species level biological diversity. Proteome dynamics in response to environmental cues or cellular insults can further contribute to biological diversity. The same protein, differentially modified or localized, could perform diverse functions in a specific cellular state. Although proteome dynamics could serve as the basis to understand species level biological diversity, the technology employed to identify and characterize protein changes require a theoretical context in order to inform molecular evolution.

Here, we put forth the concept that understanding biological commonalities through evolution could lead to the conceptualization of what contributes to the origin and maintenance of biological diversity. Proteomics approaches need not be limited by indexed and sequenced genomes from model organisms. The fact that most genes are conserved among species suggests that we could design experiments to search for evolutionary relationships from conserved proteins among members of genus, family, order, class or phylum. In this chapter, we discuss how tandem mass spectrometry and improved bioinformatics approaches could serve as powerful tools in the quest to uncover the molecular enigmas of evolution.

2. Tandem mass spectrometry: a historical perspective

Tandem mass spectrometry is an invaluable tool for identifying and sequencing proteins, and assessing their modifications. Protein sequencing methods provide insight into cellular and molecular mechanisms. Protein analysis has progressed continuously since the first reports by Dr. Edman, which illustrated that proteins are a sequential arrangement of amino acids (Edman, 1950). These key observations were soon followed by Dr. Sanger's sequencing of the first protein, insulin (Stretton, 2002). The technological revolution of the 1980's and 90's included crucial developments in mass spectrometry instrumentation by Dr. Fenn and Mr. Tanaka, which enabled the use of mass spectrometry for the study of biological macromolecules (Tanaka et al., 1988; Fenn, 2002). These developments, for which Dr. Fenn and Mr. Tanaka received the 2002 Nobel Prize in Chemistry, started a technical revolution that moved protein analysis from sequencing purified proteins one at a time in isolation, to the identification and quantification of multiple proteins in cellular extracts. These technological innovations, combined with a growing body of genomic knowledge and data, brought about the birth of *proteomics* (Patterson & Aebersold, 2003).

The monumental task of sequencing the genome of different species is an essential component for the identification of proteins and determination of their functions (Marshall, 2011). Genomic insight has made clear that protein diversity and biological function cannot be explained using just gene expression (Adami et al., 2000). Biological function will ultimately be determined by protein expression levels, in combination with each protein's structure, location and its interactions with other proteins and the environment. Posttranslational modifications (e.g. phosphorylation, glycosylation, hydroxylation) and processing (e.g. proteolytic cleavage, degradation), which alter protein sequence, structure and interactions, will further refine biological function. Proteomics as a field is essentially interested in characterizing each of these levels of complexity, for each protein, at each developmental stage. This type of approach would be impossible using traditional protein isolation, purification, and sequencing techniques.

With optimized sample preparation, tandem mass spectrometry instrumentation, and database selection and search algorithms (as described later in this chapter), multiple proteins within a complex mixture can be rapidly and definitively identified, and even quantified. The sub-field of *comparative proteomics* quantitatively compares the diversity, identity, and expression levels of proteins among samples, generally among individuals of a specific species exhibiting different phenotypes. Comparative proteomic approaches have great potential in medical diagnosis and treatment, as pathological conditions can be compared at the protein level to healthy individuals (see Sanchez et al., 2004 for detailed examples). A logical extension of comparative proteomics, *evolutionary proteomics*, assesses the diversity, identity, and expression levels of proteins among samples obtained from different species, in order to elucidate evolutionary patters and highly conserved mechanisms (for example see Budovskaya et al. 2005; Heyl et al. 2007). A unique example of evolutionary proteomics applications is the analysis of reproductive (seminal) fluids, since these fluids are directly acted upon by selective forces and show rapid diversification (Ramm et al., 2009; Marshall et al., 2011; reviewed in Findlay & Swanson 2010). Ramm et al. (2008) analyzed the array of proteins found in rodent seminal fluids among 18 muriod species. The study found significant variation in molecular mass of the same seminal proteins among species, suggestive of amino acid divergence, and showed evidence for sperm competition as a selective force.

To illustrate the potential of evolutionary proteomics in understanding the origin and maintenance of biological diversity, we present examples of highly conserved proteins that withstand evolution forces through multiple phyla. These proteins and biochemical mechanisms have been conserved through evolution because they serve a crucial function in an organism's survival and reproduction and can be adapted to serve multiple functions.

2.1 Highly conserved proteins

The integration of genomics and proteomics led to important discoveries including the identification of biomarkers and identification of conserved proteins from bacteria to humans. In the 1990's, the orthologous proteins of mammalian actin and tubulin were identified in bacteria, indicating that cytoskeletal structures are crucial components that preceded multicellular organisms (Bi & Lutkenhaus, 1991; Desai & Mitchison, 1998; Graumann, 2004). Therefore, other molecular changes through evolution, such as protein sequence and structure divergence, may explain the cytoskeletal differences between prokaryotes and eukaryotes. Understanding these types of changes through a proteomics approach will provide insight on molecular level changes that promote and/or enable

species divergence. Prerequisite knowledge of highly conserved proteins and biochemical mechanisms will guide hypothesis testing and enable narrow database searches. As in the case of actin and tubulin, proteins and biochemical mechanisms that are highly conserved through evolution are essential to the fitness of an organism (i.e. its likelihood to survive and reproduce). We present two examples of highly conserved proteins: collagen, and the proteins involved in blood coagulation. These proteins have been the focus of recent evolutionary proteomics-based reports and will be further detailed in the technical aspects of tandem mass spectrometry section of this chapter (section 3).

Collagen is found in all animals and is the most abundant protein in the majority of vertebrates. It serves as the key structural protein in vertebrates, composing the vast majority of the extracellular matrix, including bone matrix, skin and tendons. Collagen has a unique hierarchical helical structure that provides the mechanical strength and stability necessary to serve as a structural protein. At the amino acid level, collagen has a highly repetitive sequence of Xaa-Yaa-Glycine, where Xaa is most often proline and Yaa is most often hydroxyproline, although other amino acids can fill these positions. This repetitive sequence enables formation of a helical structure, composed of three polypeptides, stabilized by hydrogen bonds (i.e. tropocollagen molecules). Tropocollagen molecules pack together in a consistent staggered array producing fibers, which are cross-linked together, increasing mechanical strength. Collagen is highly adaptable, due to variations at the genetic, post-translational, and processing levels, enabling it to serve a wide range of structural functions. There are at least 28 different types of vertebrate collagen, which vary in function and/or distribution (Shoulders & Raines, 2009). Collagen variants likely evolved through a process of gene duplication and drift within duplicated genes (Boot-Handford & Tuckwell 2003).

The presence and structural importance of collagen is equally important in invertebrates. Collagen is present in all metazoans, although fibrillar forms of collagen appear to be lacking in arthropods and nematodes (Garrone, 1999; Boot-Handford & Tuckwell 2003). For example in sponges, the most primitive metazoans, spongin (short-chained collagens) are involved in both adhesion to the substrate and as a skeletal matrix (Garrone, 1999). Collagen also serves a key role in adhesion of marine mussels. Marine mussels adhere using a bundle of byssal threads, which extend from the foot of the organism to the substrate, where they display adhesive pads. Byssal threads are composed of a large collagen domain flanked by two elastin domains, which provide the byssus with both mechanical strength and flexibility (Coyne et al., 1997).

A second salient example of highly conserved proteins is those involved in blood coagulation, a process that stems the loss of blood during injury and without which an animal could not survive. In vertebrates, blood clot formation is brought about by two closely interrelated and converging proteolytic cascades, first described by Davie and Rantoff (1964) and MacFarlane (1964). Fibrinogen is proteolytically activated by this cascade, producing fibrin monomers, which form a network and are covalently cross-linked by a transglutaminase (factor XIII). In the blood coagulation cascade, inactive proteases (zymogens) are converted to their active form by limited proteolysis and then in turn activate the next protease in the cascade. All of the major coagulation cascade enzymes are trypsin-like serine proteases, which cleave on the carboxyl side of arginine residues (Neurath, 1984, 1986; Davie, 2003). Within this conserved system, variability in the non-proteolytic domains of these enzymes allow for substrate and cofactor binding specificity (Patthy, 1993; Neurath, 1999).

A similar process has been shown to be involved in invertebrate blood coagulation. For example, the blood of horseshoe crabs coagulates through two converging cascades of trypsin-like serine proteases, contained within blood cells as inactive proteases, and released upon exposure to minute quantities of pathogens (reviewed in Muta & Iwanaga, 1996; Sritunyalucksana & Soderhall, 2000; Osaki & Kawabata, 2004; Theopold et al., 2004). Blood coagulation in crustaceans occurs though the action of a Ca^{2+} activated transglutaminase (Fuller & Doolittle 1971; Lorand, 1972; Kopacek et al., 1993), which is homologous to vertebrate factor XIIIa (Wang et al., 2001). Involvement of trypsin-like serine proteases in the crustacean blood coagulation process has also been shown (Durliat & Vranckx 1981; Madaras et al., 1981; Soderhall; 1981).

The activity of proteolytic enzymes, such as those involved in vertebrate and invertebrate blood coagulation, is widespread in biological systems (Neurath & Walsh, 1976; Neurath, 1986; Krem & Di Cera, 2002). These highly adaptable enzymes serve simple digestive function in primitive organisms, yet have evolved to regulate complex physiological control in higher organisms (Neurath, 1984; Krem & Di Cera, 2002). Apart from blood coagulation, proteolytic cascades of serine proteases comprise a variety of systems including the complement reaction, fibrinolysis, and dorsal-ventral patterning in drosophila (Neurath 1984; Krem & Di Cera, 2002). In each case, the proteolytic cascade enables amplification of a small stimulus into a physiological response (Neurath & Walsh 1976; Neurath, 1986). Proteolytic cascades are highly conserved because they work well, are adaptable, and can be regulated with inhibitors, cofactors and specific feedback mechanisms. Amino acid sequence analyses supports the hypothesis that the proteolytic cascades of vertebrate blood coagulation, horseshoe crab blood coagulation, and drosophila dorsal-ventral patterning, all evolved from a common ancestral cascade (Krem & Di Cera, 2002).

Variability on a common theme or conserved mechanisms at the molecular level may enable species divergence. For example at the transcription level, variation in Hox genes, a highly conserved group of genes that regulate body plan and structure through development, has been shown to drive morphological evolution (Heffer et al., 2010). Post-translation modifications, such as phosphorylation (Boekhorst et al., 2008) are highly conserved, and variations in the extent or specific site of phosphorylation parallel phylogenetic divergence. The identification and understanding of commonalities among distant species could provide the basis to dissect protein and, ultimately, specie divergence. Thus, with an appreciation of evolutionary proteomics approaches and highly evolutionarily conserved proteins, we move on to discuss technical aspects of tandems mass spectrometry, which will ultimately determine the success or failure of tandem mass spectrometry-based research.

3. Tandem mass spectrometry: technical aspects

Tandem mass spectrometry is defined as the sequential analysis of ions and their respective fragmentation patterns (Hunt et al., 1986; Mann & Kelleher, 2008). In the case of protein mass spectrometry analysis, this approach is also known as “bottom-up” proteomics. “Bottom-up” proteomics is based on the fragmentation pattern of an ionized peptide, which is unique to the corresponding sequence of that specific peptide. The uniqueness of those product ions facilitates the identification of the peptide and, consequently, the corresponding protein. This powerful discovery tool provides the opportunity to analyze protein samples from different organisms, extinct or living, in the quest to uncover the secrets of molecular evolution. To do so, there are several important technical aspects to be

considered in tandem mass spectrometry: 1) sample preparation, 2) ionization, 3) mass spectrometer capabilities and, 4) validation.

3.1 Sample preparation

The most important aspect in mass spectrometry analysis is sample preparation. The increase in sensitivity, below attomole (10^{-18} M) levels, brought about by advancements in technology makes sample preparation crucial, and inarguably the most important step in mass spectrometry analysis (reviewed in Patterson & Aebersol, 2003). The sensitivity of the current generation of mass spectrometers leaves no room for sample preparation error. In the quest to identify proteins across different species, the integrity, quality and purity of the samples are crucial. Different sample preparation paradigms have been recommended for tandem mass spectrometry analysis. These paradigms are not exempt from potential pitfalls and each project may require customization. The most important aspect is to prevent decomposition, adducts and contamination. The first step after homogenization of tissue or cells is the isolation of proteins from the membranous fraction. Lipids, from membrane or detergents, are a contaminant that affects the analysis of proteins through tandem mass spectrometry (see Table 2, Section 3.2). But, the most common contaminant in tandem mass spectrometry analysis is keratin. Keratin is the principle structural component of the human epidermis and is found in hair and nails. Thus, the first obstacle that a user of tandem mass spectrometry technology needs to overcome is avoiding the identification of his or her own keratin. Contamination with keratin results from touching sample tubes with bare hands, skin peels falling into the sample due to dandruff, hair contamination, or simply air flow in a crowded laboratory environment. Table 1 lists keratin sources and possible solutions to prevent this common contaminant. However, keratin contamination is just the first hurdle to overcome in the application of tandem mass spectrometry to molecular evolution studies.

Keratin Source	Possible Solution
Skin, hair, nails	Wash sample tubes with ethanol three times, dry on speed vacuum Always clean the surface area of benches and instruments to be used with ethanol Work under a laminar flow hood Always use gloves and lab coat with cuffed long sleeves Always wear disposable hair cover
Aerosol	Work under a laminar flow hood Use filter-pipette tips Separate a set of pipettes that are kept in a clean area Avoid crowed laboratory areas Maintain proper ventilation
Tissue Sample	Avoid contamination with skin or hair from the tissue sample source Rinse the tissue with buffer before homogenization Avoid preparing samples close to areas used for euthanization, surgery or dissection of animals

Table 1. Sources of keratin and possible solutions.

When sample preparation is ideal, the integration of tandem mass spectrometry analysis in molecular evolution studies could provide significant advances in the understanding of molecular conservation and divergence. The concept that most genetic information is conserved between species could serve to develop hypothesis-based approaches that uncover similarities between modern organisms and those extinct millions of years ago, and raises the question of what is the actual unit of evolution. The identification of similarities could provide valuable insights about conserved mechanisms and molecular divergence between distant species such as dinosaurs and birds or arthropods and humans.

However, the use of tandem mass spectrometry in molecular evolution studies remains controversial. This is due to the lack of available protein databases for extinct or living organisms that would enable comparison between taxa and the potential for contamination with "modern" proteins during sample preparation. These concerns were the basis of a debate over a report published in *Science* in 2007. Asara et al. (2007) targeted collagen as a highly conserved and resilient protein in vertebrates, proposing that based on its molecular characteristics it could be extracted from fossil bone. Asara et al. (2007) published the identification of collagen α1t1 from two extinct animals, mastodon (*Mammut americanun*) and T. rex dinosaur (*Tyrannosaurus rex*). The authors went further to indicate that the identified *T. rex's* collagen peptides were more similar to birds' (chicken) collagen than to other species, while mastodon collagen peptides were similar to mammals (dog, bovine, human and elephant). In the same *Science* issue, Schweitzer et al. (2007) provided evidence of collagen I in *T. rex's* cortical and medullary bone, using multiple techniques, atomic force microscopy (AFM), *in situ* immunohistochemisty, and TOF-SIMS, validating the results obtained by Asara et al. (2007). Both articles were subjected to an intense scrutiny and public debate within the scientific community (Buckley et al. 2008; Pevzner et al. 2008). Part of the scrutiny is valuable as it pushes researchers to ensure the integrity of their sample preparation and analysis, while other part hinders progress, with obstacles difficult to overcome. Interestingly, and not debated, the authors acknowledge the identification of human keratin as a contaminant. Although unlikely due to fast rate decomposition of soft tissue in comparison to bone, no one questioned if the keratin was actually a conserved peptide from the well preserved dinosaur's sample.

There were three major concerns in accepting the tandem mass spectrometry data published by Asara et al. (2007), namely: sample preparation, database search, and validation of the results (Buckley et al. 2008; Pevzner et al. 2008). The bioinformatics component in tandem mass spectrometry is a bottleneck and the most quietly accepted limitation. Thus, an entire section of this chapter will be dedicated to a discussion of bioinformatics with regard to its use in molecular evolution studies. Validation is an important issue in mass spectrometry and it is discussed within this section.

Concerns about sample preparation are essential, especially when the tissue under study is a fossil that has been exposed to nature, decomposition and other organisms (big and small) for millions of years. Thus questions of the purity of the *T. rex* samples are reasonable. We are fascinated that only the results obtained from the *T. rex* samples were questioned and not the mastodon, which was exposed to the same sources of contamination though for lesser time.

Protein identification from a fossil, such as *T. rex's* bone, is similar to searching for water on Mars. As scientists, we are trained to use existing knowledge in the quest to generate new knowledge. The use of tandem mass spectrometry for the analysis of protein extracted from a fossil is based on the assumption that the amino acids that form "modern" proteins are the

same as from ancient proteins and just the arrangement or amino acid order in the protein sequence has been subjected to evolution. This is similar to the assumption that the atmosphere and atoms prevalent on Mars are similar to those on Earth, and therefore there should be water on Mars. Yes, water in Mars and amino acids on proteins from extinct organisms may exist as we know them. But, it is reasonable to argue that some differences may exist. For example, one challenge in sample preparation for extinct organisms is the assumption that the chemistry and modifications of amino acids from millions of year old proteins is the same as those recognized today. This also brings up the importance of understanding the chemistry of amino acids, peptides, and proteins in the process of sample preparation. Extraction, isolation and purification of peptides and proteins depend on the chemical properties of amino acids and their interactions within the sequence. Sample preparation often includes treatment with compounds that induce modifications of the amino acids, such as hydroxylation, dehydration, oxidation and deamination (reviewed in Lubec & Afjehi-Sadat, 2007). These modifications change the molecular mass of the amino acids and, when not taken into consideration, affect the identification of peptides, due to deviations from the theoretical molecular mass registered in the database (see below). In reality, these considerations need to be taken into account for all samples not just for fossil samples.

Hypothesis based-approaches for molecular evolution studies of living organisms is a lesser challenge than for those of extinct animals, since issues of sample degradation and chemical changes over time are less important. An example of such a study was the identification of human coagulation factors in barnacle cement samples (Dickinson et al., 2009). The aggregation and cross-linking of barnacle cement proteins is crucial for the barnacle's survival since the cement anchors the organism to the surface, enabling feeding and reproduction. Based on the essential role of cement formation in the life cycle, it was hypothesized that the formation of this structure in an aqueous environment could be related to the molecular process involved in blood coagulation. As described in section 2.1, highly evolutionarily conserved proteins comprise the blood coagulation cascade. Blood coagulation is a life or death process that also involves the aggregation and cross-linking of soluble proteins.

Examination of the cement using AFM revealed a mesh of fibrous proteins, structurally similar to a fibrin blood clot. The first technical obstacle in analyzing the cement was sample preparation. Dickinson el al. (2009) developed a strategy to obtain proteins secreted by the barnacle during the process of cement release but before the secretion cured; the curing process renders most of the proteins insoluble. Complete proteins can only be collected prior to curing which involves at least one and probably other types of cross-linking (Dickinson et al., 2009). AFM of cement collected in this manner indicated the formation of fibrous structures upon polymerization, indistinguishable from those made by the barnacle *in situ*, validating the collection technique.

Next, tandem mass spectrometry was used to uncover molecular similarities in proteins between barnacles and humans (Dickinson et al., 2009). The barnacle cement's extracted proteins were resolved in one dimensional SDS-PAGE and subjected to trypsin digestion. The purified peptides were then analyzed by tandem mass spectrometry and the obtained spectrum subjected to sequence analysis against the human database. This approach led to the identification of two conserved peptides that correspond to the protein Transglutaminase, factor XIIIA, which plays a crucial role in the process of blood coagulation (Dickinson et al., 2009). This result suggests that coagulation is a conserved

molecular mechanism crucial for the survival of all organisms and that has been preserved through evolution. Interestingly, the peptides identified correspond to only one factor in a complex and multifactorial molecular mechanism. Thus, the result obtained serves as foundation to uncover the similarities between blood coagulation and cement formation, but importantly the information gathered could be used to delineate the divergence of this process in organisms from two very distantly related taxa.

3.2 Ionization

The popular phrase "garbage in, garbage out" describes the importance of sample preparation in tandem mass spectrometry. However, there is one more important component that needs to be considered before the analytes reach the mass analyzer of a mass spectrometer: ionization efficiency. The process of sample ionization is carried out by two principal ionization sources used for tandem mass spectrometry analysis, namely Electrospray ionization (ESI) and Matrix-assisted laser desorption ionization (MALDI). These two sources of soft ionization were crucial in the integration of mass spectrometry for the analysis of biomolecules (Tanaka et al., 1988; Fenn, 2002). After all, only ions that are generated by an ionization method can be analyzed by a mass spectrometer.

Chemical	Source	Effect
Lipid	Cellular membrane, organelles	Affect resolution when liquid chromatography is used Produce prominent and persistent ions that could suppress or mask other ions
Detergent	Sample preparation buffers, surfactants	Affect resolution when liquid chromatography is used Produce prominent and persistent ions that could suppress or mask other ions Induce the formation of adducts
Salt	Sample preparation buffers	Induce the formation of adducts Ion suppression Ion overloading
Trifluoroacetic Acid (TFA)	Sample preparation solvent, reverse chromatography	Ionization suppression

Table 2. Chemicals and biomolecules that affect ionization and ion detection.

The efficiency of ESI to generate ions depends on three important factors: temperature, flow, and voltage. However, other factors may suppress ion formation as described in Table 2. A temperature of 200°C is constantly used at the ion transfer tube, which serves as the entrance to the mass analyzer. This high temperature allows the evaporation of highly volatile solvents in which analytes are dissolved (Table 3). The evaporation of the sprayed solvent containing the analytes generates the separation of one drop into smaller drops in a physical concept known as "Coloumbic explosion." This process continues until the analytes are completely dry and attracted to the mass analyzer where the ions are detected. Evaporation

of the volatile solvent is a prerequisite for ion detection and, therefore, directly related to sensitivity.

Based on this principle, flow rate is inversely related to sensitivity. The lower the flow rate (200-300 nL/min) the faster the solvent is evaporated promoting more ions to be dried out and detected by the mass analyzer. Thus, lower flow rate implies higher sensitivity. However, ions are also formed if voltage is applied to the sample. In ESI, voltages applied to the sample range from 1.0 to 2.5 kV. The combination of temperature, flow and voltage generates the spray of ionized analytes detected by the mass analyzer.

MALDI requires use of a chemical matrix in which the analytes are embedded (Lubec & Afjehi-Sadat, 2007). Table 3 shows matrices frequently used. Since laser energy is absorbed by the matrix and transferred to the analyte, the solubility of the analytes in the matrix affects the ionization efficiency. In practice, the three most common matrices are used to determine the efficiency of ionization at the level of ion detection. Thus, it is recommended to dissolve the same sample in each of the three matrices and evaluate results. Energy from the laser is equivalent to voltage used in ESI, and therefore requires fine tuning and optimization. The energy promotes ionization and transition to the gas phase. As for ESI, the MALDI matrices are volatile so that the ions formed can be detected by the mass analyzer (see below).

Ionization source	Solvent
ESI	(10-80%) Acetonitrile/ (90-20%) Water/ (0.1-0.2%) Formic Acid (10-80%) Methanol/ (90-20%) Water/ (0.1-1.0%) Acetic Acid
MALDI	2,5-Dihydroxybenzoic acid (DHB) 3,5-Dimethoxy-4-hydroxycinnamic acid (sinapinic acid) α-Cyano-4-hydroxycinnamic acid (CHCA) 3-amino-4-hydroxybenzoic acid

Table 3. Ionization source and common solvents.

ESI and MALDI both have their advantages and disadvantages (Lubec & Afjehi-Sadat, 2007). ESI provides the advantage of ionizing the analytes as they are eluted from a column. Liquid chromatography online to the ionization source provides a better resolution of the analytes, avoiding clustering and exceeding the dynamic range (i.e. a measure of the detection range of a detector; ratio of the largest to smallest detectable signal) of the instrument. In MALDI, the chromatography needs to be carried out off-line and the fractions analyzed individually. Off-line chromatography allows the analysis of only the desired fractions, while in on-line chromatography all elution steps are analyzed. Older instrumentation for on-line chromatography did not allow the recovery of fractions at a specific retention time for re-analysis. At present, there are instruments, such as NanoMate TriVersa (Advion), that collect fractions through out the chromatography gradient without interrupting the ionization. Although controversial, MALDI is recognized as a better ionization source for its capacity of assisted-energy transfer to the analyte, which reduces the fragmentation or decomposition of the analyte at the ionization source. MALDI ionizes certain peptides better than ESI, and vice versa. Thus, in molecular evolutionary studies is important, when possible, to take in consideration and use both ionization sources for tandem mass spectrometry analysis of a particular sample.

In summary, ionization efficiency plays a crucial role in tandem mass spectrometry. It is important to understand that the most abundant peptide in the prepared sample is not necessarily the most abundant ion in the mass spectrum, because each peptide that reaches the ionization source will have different ionization efficiency. Samples derived from preserved fossils or live organisms need to be prepared in accordance with the requirements of the ionization source used. Good sample preparation will increase the chances of protein identification or characterization. However, the lack of fine tuning of the ionization source could induce fragmentation, clustering and ionic suppression, converting sample preparation efforts in a futile exercise. Once the conditions are conducive to increasing the ionization efficiency, the mass spectrometer itself is the next component that requires optimization.

3.3 Mass spectrometers

Mass analyzers have progressed at a fast and steady pace (Patterson & Aebersold, 2003). Sir Joseph John Thomson would be delighted to see how the measurement of mass-to-charge (m/z) ratio is done today. Crookes tubes used by Dr. JJ Thomson are pieces of history in a science museum. From instruments with very low resolution and dynamic range in the 1940's (e.g. MS-2) to today's high resolution instruments, such as the 14-Tesla Fourier transform ion cyclotron resonance (FT-ICR) instrument, mass spectrometers continue pushing forward our capacity to discover and understand molecular processes. Increased mass accuracy, high resolving power, scanning speed and affordability are some of the attributes of modern instruments (Mann & Kelleher, 2008). Given the technological progress in mass spectrometry, this can become a reliable tool in the understanding of molecular evolution. Nonetheless, the efficiency of a mass spectrometer depends on both of the parameters discussed above, sample preparation and ionization efficiency.

The controversy surrounding the *T. rex's* peptides sequences discussed above provide the basis to understand why ion mass accuracy is an important criterion in the selection of a mass spectrometer for a specific application (Mann & Kelleher, 2008). Why is ion mass accuracy important? The identification of a peptide depends on the bioinformatics analysis of the mass spectra recorded from a specific sample. The mass spectrometer detects ions, which we know are from a digested protein(s) or from "chemical noise" in the sample. The requirement for ion mass accuracy is directly related to the bioinformatics tools available for data mining that depends on the mass of the precursor and product ions in a tandem mass spectrometry analysis. The debate on the confidence level in the correlation of the MS and MS/MS data with the database starts at the level of the MS spectrum (Pevzner et al., 2008). In the *T. rex* example, the authors used an instrument that is considered a low-resolution instrument, LTQ (linear ion trap), which has a lower mass accuracy than other available instruments (Asara et al., 2007). Although mass accuracy is intrinsically related to the resolving power of the mass spectrometer, it does not mean that a low-resolution mass spectrometer generates poor quality data. The main point here is that the user needs to consider the instrument's capabilities in regard to the research application(s) and goal(s).

The resolving power of a mass spectrometer is defined, in simple terms, as the mass analyzer's capability to set apart two ions with similar m/z ratios. At the mass spectrum level, the instrument resolving power can be seen as the "*valley*" or distance between two ions with similar m/z (Mann & Kelleher, 2008). Examples of low- and high-resolution instruments are listed in Table 4. The main difference among these instruments is the type of mass analyzer used. Low-resolution mass spectrometers are more affordable and widely

used for proteomics approaches in life sciences and biomedical research. Most of these instruments are very sensitive, have high scanning speed and robust performance. Linear ion trap mass analyzers are an example of a low-resolution mass spectrometer, but they compensate with a high scanning speed and dynamic range. On the other hand, high-resolution instruments generate accurate mass data, but with lower sensitivity and robustness. Although it is not always feasible, ideally selection between low- and high-resolution mass spectrometer is not just an issue of affordability (i.e. funds available), but also of application.

Resolution	Mass Analyzer	Instruments (examples)
Low	Quadrupole	Xevo TQ, Waters TSQ, ThermoElectron
	Ion trap	LCQ and LTQ, ThermoElectron 240-MS, Agilent amaZon, Bruker
	Time-of-flight	Autoflex, Bruker TripleTOF 5600, AB Sciex
High	Hybrid	LTQ Orbitrap, ThermoElectron Synapt G2-S, Waters
	Fourier Transform Ion Cyclotron Resonance	Apex Qh 9.4TFT-ICR and solarixFTICR, Bruker LTQ-FTICR, ThermoElectron

Table 4. Low- and high-resolution mass spectrometers.

For example, generation of data from complex protein samples could be achieved using an instrument with high scanning speed and sensitivity. The scanning speed is how fast the instrument can target an ion for fragmentation. In complex protein samples many peptides will elute from a column (LC-ESI) or ionize (MALDI) at the same time. Data acquisition from such a sample will depend on how many ions were selected and subject to fragmentation. In this case, a low-resolution mass spectrometer could extract the most information from the sample. However, for applications that require precise determination of m/z, such as quantification studies and identification of posttranslational modifications, high resolution mass spectrometers (e.g. Fourier Transform Ion Cyclotron Resonance, FT-ICR) are more appropriate. Based on scientific needs and marketing considerations, the industry produces hybrid mass spectrometers, which provide the best of both worlds. These instruments use a low-resolution mass analyzer (e.g. linear ion trap) to obtain information from complex samples, taking advantage of high scanning speed and sensitivity. Then, after the first data acquisition, ions can be transferred to FT mass analyzer where highly accurate mass is obtained from the selected ions. Thus, hybrid mass spectrometers are versatile instruments that should be considered as the most suitable tool in evolutionary proteomics approaches, since they are able to analyze complex mixtures of proteins as well as enable quantitative comparisons and high mass accuracy.

A relevant issue raised above is the identification of proteins from fossil samples or distant taxa using the information and knowledge on "modern" proteins or genomes with limited known sequences. The identification of proteins depends on detection of the precursor ion

and its fragmentation pattern in the mass analyzer. The MS and MS/MS spectra are then compared with a database to extract the information that will lead to the identification of a peptide and, consequently, the corresponding protein. But, why cannot all peptides produced from a digested protein be identified? In addition to differential ionization efficiency between peptides, sensitivity, dynamic range and mass accuracy are crucial components that contribute to sequence coverage. For example, mass accuracy contributes to differentiating between isobaric amino acids, such as lysine (128.095) and glutamine (128.059). Another isobaric pair is leucine (131.094) and isoleucine (131.094), which are differentiated by fragmentation of their side chain. Amino acid modifications can also induce isobaric pairs, for example oxidized methionine (147.035) and phenylalanine (147.068). It is feasible to hypothesize that, in preserved fossil tissue or living organisms from different taxa, amino acids may be differentially modified or the peptides may contain unidentified amino acids. Thus, the question is: how many peptides are needed to correctly identify a protein present in a sample using tandem mass spectrometry? The rule of thumb is that two peptides are needed for a positive identification of a protein. Needless to say, more is always better, but when databases from living organisms are used to identify proteins extracted from fossil tissue this rule of thumb shortens the abysmal age differential between the extinct and the living. If we do identify matching peptides, how can we be certain that the identified protein is definitely in the sample? The answer is validation, validation, validation.

3.4 Validation, validation, validation…

Optimization and understanding the different components that are required for tandem mass spectrometry contributes to the generation of high quality and quantifiable data. The attention to detail during sample preparation contributes to avoiding contamination, undesirable modifications and decomposition of proteins. Calibration and fine tuning of the ionization source and mass analyzer increase the sensitivity and maximize the capabilities of the mass spectrometer. Despite all these considerations, the data generated by tandem mass spectrometry depend on the power of the computational tools and data mining capabilities available to the user. There is a phrase commonly used in the laboratory about the interpretation of MS data; "*if you did not detect a peptide, it does not mean that it is not there, but if a protein was identified you better validate it…*"

Validation of the results obtained by tandem mass spectrometry analysis is a requirement in today's proteomics approach. This fact recognizes that false positives could be detected despite the selection of the most stringent filter parameters and the requirement of powerful computational tools. However, the need for validation does not minimize the importance of tandem mass spectrometry as a discovery tool. It is just another example of how the scientific community finds ways to improve the quality and validity of their results. After all, a result could trigger a hypothesis, and the hypothesis can only be accepted as theory if it persists in scientific scrutiny over time.

Returning to the Asara et al. (2007) example and the controversy surrounding the identification of collagen peptides conserved from dinosaurs to chickens, the authors used different validation strategies to certify their tandem mass spectrometry analysis results, which were based on a predetermined search for collagen peptides. The authors search for collagen because it is a structurally strong protein that could withstand exposure to harsh environmental conditions and be readily detected in bone tissue samples. First, the authors generated a "database" that allowed them to search for the predicted collagen peptides in

the raw MS/MS-MS spectra obtained in the tandem mass spectrometry analysis of proteins extracted from the bone tissue. They generated a "database" using conserved sequences among different taxa and point-assisted mutation matrices to take into consideration amino acid divergence throughout evolution. The collection of peptides in the database was used to identify peptides in the protein sample from *T. rex* bone tissue. This hypothesis-based approach allowed the identification of 33% sequence coverage for collagen α1t1 and 16% collagen α2t1 (Asara et al., 2007). Validation of these results is necessary.

For validation of the identified collagen peptides from *T. rex*, Asara et al. (2007) used synthetic peptides. The synthesized collagen peptides were subjected to tandem mass spectrometry and the resulting fragmentation patterns (i.e. MS/MS) were compared to the identified peptide from the *T. rex* sample. This strategy demonstrated that product ions from the *T. rex* sample were similar to the corresponding synthetic peptide. Although this is a valid strategy, this validation depends on the same technique that was used to generate the data under scrutiny. The best validation strategies are described in a complementary report by Schweitzer et al. (2007), published back-to-back with Asara et al (2007) in the same *Science* issue. Schweitzer et al. (2007) used immunochemistry to locate collagen protein in the tissue using an antibody against avian collagen. The antibody clearly showed immunoreactivity in both cortical and medullary bone tissue from *T. rex*. In order to detected immunoreactivity, the fixed tissue would need to have an epitope recognized by the anti-avian collagen protein, suggesting that this protein may be present in the bone tissue. Taken together with AFM and TOF-SIMS analyses, the immunochemical analysis of *T. rex* tissue confirmed the tandem mass spectrometry data, indicating that at least, peptide sequences homologous to collagen were present in the dinosaurs sample.

The second example described earlier in this chapter was the identification of coagulation factors conserved from barnacle to human. Dickinson et al. (2009) demonstrated that the polymerization of barnacle cement is similar to blood coagulation. The authors validated the identification of Factor XIIIA (transglutaminase), using western blot and functional analyses. Immunoreactivity to human factor XIIIA antibody was found in barnacle cement at approximately 75 kDa, remarkably close to the 83 kDa mass of human factor XIIIA despite roughly a billion years of evolutionary divergence. Dickinson et al. (2009) went on to conduct two functional analyses: quantification of transglutaminase activity level and an amino acid composition analysis procedure that specifically probed the end-product of transglutaminase cross-linking (as described by Pisano et al., 1968, 1969). Both functional analyses confirmed the presence and activity of a transglutaminase in barnacle cement.

The report by Dickinson et al. (2009) further expanded the evolutionary hypotheses established by testing for other components of the blood coagulation cascade, trypsin-like serine proteases, not specifically identified by tandem mass spectrometry. Evidence of trypsin activity was shown through SDS-PAGE, i.e. a pattern of pro-forms of proteins, active proteins and proteolytic clips. Activity of trypsin in barnacle cement was demonstrated by trypsin activity quantification (using an arginine ester substrate), and the reliance of the polymerization process on proteolytic activity was shown through trypsin inhibition assays. Furthermore, western blot analysis showed immunoreactivity to bovine trypsin antibodies, indicating shared epitopes between barnacle and vertebrate trypsin. Variations in non-proteolytic domains of these enzymes are likely to enable substrate specificity. The results demonstrated that coagulation is a conserved biological process relevant to the fitness and survival of organisms as distant, in evolutionary terms, as barnacles and humans. This on-going work is then directed toward using these similarities as a basis to dissect the

molecular requirements for barnacle adhesion in order to identify functional divergence from the mammalian blood coagulation process.

Validation of tandem mass spectrometry results is based on the application or source of the sample used in the analysis. Protein identification could be validated by immunological techniques such as western blot, immunohistochemistry and immunocytochemistry. Also, immunoprecipitation or chromatography purification, followed by western blot analysis could be used to validate the identification of a novel protein as part of a known protein complex. Vega et al. (2008) demonstrated, using tandem mass spectrometry analysis of co-immunoprecipitated proteins, the association between the novel mouse protein EFhd2 and the human microtubule-associated protein tau (Tau) expressed in the brain of the tauopathy mouse model JNPL3 (Vega et al., 2008). The identified mouse's EFhd2 protein is 93% identical to its human counterpart. Thus, immunoprecipitation of Tau from human brain samples followed by western blot analysis validated its association with the human EFhd2 (Vega et al., 2008). This is not only an example of validation, but also how conserved proteins from different living organisms and distant taxa (i.e. mouse to human) could be used to identify proteins that were conserved through evolution. Thus, validation of results obtained by tandem mass spectrometry analysis is a way to reinforce an already powerful discovery tool.

4. Bioinformatics

As we have discussed in the preceding sections, two issues come to the forefront when performing evolutionary proteomics; is our protein of interest *accurately* identified and can we *reliably* infer the phylogeny of the organism in question? These questions become more acute when considering fossil material, where we will only have fragments of proteins to work with. We have already considered some aspects of this issue earlier in section 3.4. In this section we discuss bioinformatics approaches and how they can address issues of reliability.

4.1 Identifying proteins, databases and search strategies

4.1.1 Sensitivity: In both recovering sequences from fossil material (either bone or preserved tissue like Thylacine tissues) or from biological materials where we hope to have mechanistic insights through homology (e.g. barnacle glues), the issue of sensitivity is important. For fossil material, degradation means that peptides of interest will be at low abundance. For other materials, the proteins of interest may be of low abundance. Peptide MS can achieve exquisite sensitivity, in the attomole range or sometimes even below. This brings fossil material into the realm of the possible. However, while sensitivity is important, another issue is dynamic range. High-abundance proteins limit what we can observe with the given dynamic range of detection. For detection of peptides in complex mixtures, such as we see with a sample such as barnacle cement, the dynamic range is currently in the range of 10^3 to 10^4 (Mann & Kelleher 2008). This dynamic range issue may be why serine proteases were not detected in MS/MS scans of barnacle cement. For fossil material, exogenous contaminants such as bacterial proteins and keratin can be of high abundance compared to the desired protein (Edwards, 2011), making detection of low abundance material difficult.

4.1.2 Search Programs: The most common approach to identifying proteins in modern MS/MS data is to search un-interpreted MS/MS data directly. There are a large number of

programs, both commercial and free, which can implement these searches. Commercial programs include SEQUEST, the original MS/MS database searching program, Phenyx, ProteinLynx and many others. A selection of popular free programs is given in Table 5.

Program	URL
InSPect (web based, requires registration)	http://proteomics.ucsd.edu/LiveSearch/
Mascot (free web version and licensed version)	http://www.matrixscience.com/
OMSSA (web based)	http://pubchem.ncbi.nlm.nih.gov/omssa/
Sonar (web based)	http://hs2.proteome.ca/prowl/knexus.html
X!-tandem (computer based –Linux and Windows)	http://www.thegpm.org/tandem/

Table 5. Freely available *de novo* MS/MS peptide identification software.

Each program uses a variety of strategies to determine sequence identity. SEQUEST uses correlation factors to match peptides while the Open Mass Spectrometry Search Algorithm (OMSSA), Mascot and X!-tandem use a probabilistic approach. As well, there are a bewildering variety of options available; all allow a choice of digest conditions, with varying options. For example OMSSA has a menu of 20 different digest conditions, including no digestion, while Sonar has only seven (and does not have a "no digestion" option). Similarly, there are a number of options for modifications, both fixed and variable. InSpect has the fewest options available of the free programs. All have a number of database options to search, which will be discussed in the next section.

4.1.3 Data formats: There are a wide variety of data formats produced by MS/MS instruments, many of which are proprietary and not readable by many of the programs mentioned here. These formats will need to be converted into a common format. Usually the vendor of the instrument will provide some data export capability, but there are also some open source tools at Proteowizard (http://proteowizard.sourceforge.net/) and the Trans Proteomic Pipeline (http://tools.proteomecenter.org/software.php). Conversion of these files may cause some loss of information (e.g. loss of metadata when converting to DTA format), which can potentially impact on the identification results. File formats include DTA, a simple text file format with no metadata, PKL, MGF (Mascot Generic Format) and a variety of formats to encode metadata using XML (mzXML, mzData and mzML). This impacts the programs you can use to search for matches, SEQUEST uses the DTA format, X! Tandem is set up to use DTA, PKL or MGF files, and OMMSA can handle DTA, XML encapsulated DTA and PKL or MGF files. The PRIDE database of MS spectra (http://www.ebi.ac.uk/pride/init.do) , which can be helpful to validate search strategies, uses the mzData format.

Given the variety of programs and formats available, which do you choose? Head-to-Head comparisons have been performed on a limited number of these programs using validated MS spectra; Boutilier et al (2005) compared SEAQUEST, Mascot, Sonar and Pepsea on an LCQ and a high resolution mass spectrometer. They found limited overlap between the programs, and this overlap was different with each instrument used. One approach to overcome this limitation is to use multiple search engines. This comes with a significant computational cost, but if differing programs identify the same peptides then there is greater confidence in the result (Boutilier et al. 2005). This approach was used by Asara et

al., (2007) where both Mascot and SEQUEST scores were used as part of the process to validate *T. rex* collagen sequences.

4.1.4 Databases: Database choice is a trade-off between sensitivity and the time it takes to search the database. Smaller, selective databases will take a short time to search but may miss important peptides, while larger databases take much longer to search and may produce results of lower statistical significance.

Database	URL
RefSeq Protein	http://www.ncbi.nlm.nih.gov/RefSeq/
UniProt/Swiss Prot	http://www.uniprot.org/
Human Proteomics Database	http://www.hprd.org/
UniRef	http://www.ebi.ac.uk/uniref/
International Protein Index (closes Sept 2011)	http://www.ebi.ac.uk/IPI/IPIhelp.html

Table 6. Databases suitable for MS/MS peptide identification in fossil and phylogenetic material.

The National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov /RefSeq/) and the European Bioinformatics Institute (http://www.ebi.ac.uk/) house two major curated sequence databases, along with search engines and other bioinformatics tools. These can be searched online through the MS search program, or the databases downloaded locally for faster searching if the MS/MS predictor program allows searching local databases (Table 6). For meaningful searches, it is imperative that high quality, curated databases be used. Within even these narrow limits the choice of database is, as mentioned above, a trade-off between speed of searching and low false positive rate versus more exclusive searches (Duncan, 2010; Edwards, 2011; Cottrell, 2011).

For fossil material, smaller databases may be appropriate. For example, the SwissProt section of the UniProtKB database is a non-redundant database. This means that it contains a consensuses sequence for each protein, the known variants being encapsulated in a single entry. Searching SwissProt will give fast results, at the risk of missing proteins with a low representation. In the case of fossil bone, where we are typically trying to recover collagen sequences, the speed and low error rate of such databases may be acceptable. More comprehensive databases, where all the known sequences are represented (such as UniRef100 or NCBInr) may be more appropriate with material from contemporary samples where low abundance proteins are important.

Within a database, it may be worthwhile to restrict the search taxonomically to organisms with the highest potential match. Most of the programs mentioned above have some capacity to do this. Some (e.g. SEQUEST and X!-tandem), can run custom databases if the peptides are likely to be poorly represented in larger databases. In follow up to the Dickinson et al. (2009) report, the barnacle cement sequences were compared to a database of custom non-redundant sequences from *Pacifastacus leniusculus*, a crustacean in which coagulation has been well studied (Dickinson, Vega, Rittschof, Musgrave, *unpublished*).

Anyone wishing to undertake identification of fossil proteins should make use of the PRIDE database of MS spectra (http://www.ebi.ac.uk/pride/init.do) which contains MS spectra for six extinct species, including *T. Rex* and Mammoth, to validate their search strategies against these fossil spectra before proceeding.

4.1.5 When matches are found: With programs like SEQUEST, the final scores are given as cross correlation. This does not typically provide enough information to decide if the protein identifications are valid. Usually the SEQUEST output is passed through a series of validation stages, typically the program Scaffold (http://www.proteomesoftware.com/Proteome_software_prod_Scaffold.html) to validate the peptide identities, and ProteinProphet (http://proteinprophet.sourceforge.net/: Nesvizhskiiet al., 2003) to convert the cross correlation scores to probabilities. Programs like Mascot and X!-tandem return probabilities that the matches are wrong (an empirical expectation value in the case of X!-tandem and a theoretical P-value in the case of Mascot). While X!-tandem and Mascot do not explicitly need post-processing, conversion of their statistics to the probability of a true match with ProteinProphet can be helpful, especially when comparing outputs. As stated above, when matches are found the rule of thumb is validation, validation, validation.

4.1.6 Contaminants: Searching your results against a contaminants database is very important, especially with fossil samples where contamination is a perennial problem (see section 3.1 above for methods to reduce contamination; see also Edwards, 2011 and Cottrell, 2011). There are two collections of contaminants databases available for researchers to download. The Max Planck Institute of Biochemistry, Martinsried, maintains a file of proteins selected from the International Protein Index (http://www.biochem.mpg.de/en/rd/maxquant/110606_backup/Downloads/Downloads.html). The Global Proteome Machine Organization maintains a common Repository of Adventitious Proteins (cRAP) which contains proteins selected from UniProt (http://www.thegpm.org/crap/index.html). Additionally, most algorithms (e.g. BioWorks, Proteome Discover) include filters for the most prevalent peptides produced from the major proteases used in tandem mass spectrometry analysis. For example, by enabling this feature, peptides from trypsin (or even keratin) could be filtered from the data. On the other hand, an exclusion list with the m/z ratios of the most prominent ions from expected contaminants in the sample preparation or solvent used could be created as part of the method. The mass analyzer will not waste time fragmenting ions with the m/z ratios listed. Although this pre-analysis exclusion strategy will generate "*cleaner*" data, it has the disadvantage that in low-resolution instruments peptides with similar m/z ratios of the one excluded will not be analyzed.

4.1.7 False discovery rates: When running peptides through very large databases, there will inevitably be false positive matches (above and beyond the contaminant issue above). Estimates of the False Discovery Rate (FDR) are now used alongside the statistical measures for indentifying peptides (Cottrell, 2011; Edwards, 2011). The FDR must be computed for each sequencing run, and can be computed by running the spectra against a decoy. The decoy run must be with identical search parameters to the real run, and the number of matches from the decoy database gives an estimate of the number of false positives in the run against the genuine database. Decoys can be sequences from the genuine database which have been either reversed or shuffled. Some search engines can run the genuine and decoy databases simultaneously. Cutoff for the FDR can be between 1% and 10%, however, when running fossil sequences it is best to be as conservative as possible (Ramos-Fernández et al., 2008).

4.2 Finding phylogenies

With your peptide sequences in hand, you can now approach building phylogenies. This is no trivial task, for any group of organisms, there exist an enormous number of possible phylogenetic trees which much be searched to find the optimal tree. As well, homoplasy (convergent evolution, parallel evolution and site reversal) can confuse the analysis even more. With MS/MS derived sequences, there is an additional problem if we do not have the full protein sequence since some potentially useful phylogenetic sequence information will be missing. Nonetheless, with careful attention useful phylogenies can be obtained.

Program	URL
BLAST (standalone or web-based versions)	http://blast.ncbi.nlm.nih.gov/Blast.cgi
COMPASS (web based multiple Sequence alignment)	http://prodata.swmed.edu/compass/compass.php
MAFFT (web-based multiple Sequence alignment and phylogeny)	http://mafft.cbrc.jp/alignment/server/index.html
MUSCLE (web-based multiple Sequence alignment and phylogeny)	http://www.drive5.com/muscle/
EMBL-EBI ClustalW (web-based phylogeny)	http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/
Mobile@Pasture (web-based Multiple phylogenetic methods)	http://mobyle.pasteur.fr/cgi-bin/portal.py?#welcome

Table 7. Phylogenetics programs suitable for evolutionary MS/MS.

The first step in producing phylogenies is to produce multiple alignments with the database of sequences with which you wish to form a phylogeny. This aspect of the process is often given less attention than the others, but the accuracy of the alignments can have a significant impact on the phylogeny derived from them (Ogden & Rosenberg, 2006). The problem is particularly acute with fossil material, as coverage of the sequence will be incomplete (e.g. only 32% of mammoth α1t1 collagen sequence was recovered; Asara et al., 2007), and alignment matching may have significant mis-matches. Programs like MUSCLE, COMPASS or MAFFT which use iterative alignment gaps can have better selectivity than simple CLUSTAL-W alignments. MAFFT also has the advantage of being tolerant of large gaps, such as found in the MS/MS peptides recovered.

Once an alignment is produced then a phylogeny can be derived. There are many different approaches to creating the phylogeny. Methods include Maximum Likelihood, Maximum Parsimony, Bayesian, Neighbor Joining and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). As with the choice of databases, the choice of a phylogenetic algorithm is a trade-off between speed and accuracy. In head to head comparisons, Maximum Likelihood and Bayesian approaches outperformed Maximum Parsimony and Neighbor Joining (Ogden & Rosenberg, 2006) and Bayesian and UPGMA outperform Maximum Likelihood (Douady et al 2003, Kahn et al., 2008). In general, with the large gaps in MS/MS sequence data, Neighbor Joining does poorly. Unfortunately, at present, Maximum Likelihood approaches can be prohibitively slow.

A possible approach is to use multiple tree-making programs to determine the consensus. This approach was used by Organ et al., (2008) who used two independent implementations of Maximum Likelihood and one Neighborhood Joining scheme to determine the taxonomic status of peptides retrieved from mammoth and *T. Rex*. All retrieved the *T. Rex* sequences as being closer to chicken than to mammals, amphibians and fish; however, Neighborhood Joining misplaced the alligator and mammoth sequences (Organ et al., 2008). UPGMA does a much better job of developing relationships, but places mammoth in *Canis*.

Choosing the sequences to enter into your program can be critical; the tree constructed from just aligning the top sequences picked up with BLAST and PHI-BLAST places *T. Rex* next to *Rattus*. Excessive mammalian sequences can bias the result. A balance of taxonomic forms is needed as well as a moderate sized-search space (Liska & Shevenko, 2003). A problem with studying the other example described earlier, barnacle cement, is that that representative proteins from related organisms are quite limited.

As for peptide prediction, there are a large number of programs available to perform phylogenies, with many web-based systems (see the Table 7 for a short list of common programs). Mobile@Pasture is a one stop shop, where a variety of different phylogenetic methods are available (Maximum Likelihood, Bayesian and UPGMA) often with more than one variation. The web versions of MUSCLE and MAFFT both come with inbuilt phylogeny tests, but they are quite limited (Neighbor Joining and UPGMA). Again, anyone contemplating constructing fossil phylogenies should test their programs against the FASTA dataset in the supplemental data from Organ et al (2008).

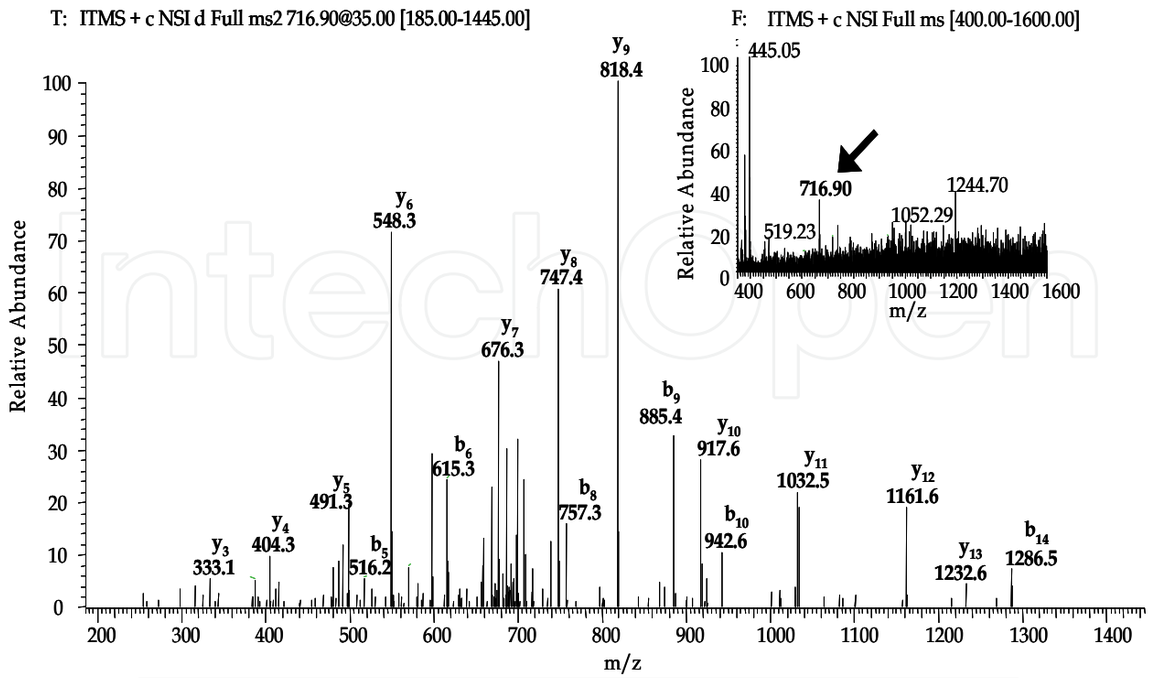
5. Future directions

All aspects discussed above are important to be considered by the novice and contemplated by the expert user of tandem mass spectrometry. They provide the basis to suggest future directions that can enhance the use of tandem mass spectrometry in molecular evolution studies. Special attention is directed to the subjects discussed below, with the understanding that they are just some of the important areas to be considered in the quest to move from hypothesis-based to discovery-based approaches in evolutionary proteomics.

5.1 Protein identification by *de novo* sequencing

The identification of proteins from tandem mass spectrometry data heavily relies on databases generated from sequenced genomes (Colinge & Bennet, 2007; Kumar & Mann, 2009). As described above, this dependence hinders the capacity of tandem mass spectrometry as a discovery tool. Additionally, many of the MS/MS data generated evade the process of identification due to amino acid modifications not taken in consideration, algorithms and filters used, among others. Analyzing each spectrum generated by tandem mass spectrometry eliminates the high throughput capabilities of the instrument. Discovery-based approaches are hindered by the limitation of not being able to detect non-conserved peptides in a sample from extinct or living organisms using available databases. These facts call to automatize the search for novel sequences in MS/MS data through a process called *de novo* sequencing.

Today, identification of protein sequences by tandem mass spectrometry depends mostly on database search algorithms. As discussed above, the ions detected in a MS spectrum are selected for induced fragmentation, generating product ions in a second MS spectrum or MS/MS. For example, figure 1 shows a MS spectrum (insert) where an ion corresponding to



b ion		Amino Acid		y ion
-----	1	E	15	-----
201.1	2	A	14	1303.6
272.1	3	A	13	1232.6
401.2	4	E	12	1161.5
516.2	5	D	11	1032.5
615.3	6	V	10	917.5
686.3	7	A	9	818.4
757.3	8	A	8	747.4
885.4	9	Q	7	676.3
942.4	10	G	6	548.3
1029.4	11	S	5	491.2
1100.5	12	A	4	404.2
1157.5	13	G	3	333.2
1286.5	14	E	2	276.2
-----	15	K	1	147.1

Fig. 1. MS and MS/MS spectra of a specific peptide. The figure illustrates the MS spectrum (insert). The arrow indicates the m/z ratio of the ion corresponding to the 2+ charge state of the peptide (black arrow). The MS/MS spectrum illustrates the product ions of fragmentation. The specific peptide fragments identified by database search analysis are indicated. These correspond to the expected product ions as shown in the table.

the 2+ charge state (black arrow) of the peptide listed on the table was detected. Fragmentation of a peptide is produced by different techniques (see above). In this example, collision-induced dissociation (CID) was used. CID is based on the movement of helium atoms inside the linear-ion trap that collide against the selected ion (peptide). The peptide bond is rigid based on its partial double-bond character. The instrument is tuned to induced fracture at the peptide bond. An analogy that explains this concept well is an accident that has probably happened to all of us at one time in our life. Remember when you threw a baseball against a glass window? Yes, it broke! However, if the baseball had been thrown, at

the same speed, to the glove of a friend, the glove did not break. Well, the glove is flexible while the glass is rigid. The collision of helium atoms against the peptide induces its fragmentation. A fragmentation event at the peptide bond generates two product ions, known as b and y ions. The b ion is the fragment containing the carboxyl terminal of the peptide bond and the y ion contains the amino terminal of the peptide bond. Thus, if a peptide of 15 amino acids (as the one shown in figure 1, table) is fragmented at the fifth amino acid of the sequence, b5 ion and y10 ion are generated. The products of the fragmentation events are shown in the MS/MS spectrum and analyzed against the database. The database search algorithms identified the corresponding sequence based on a comparison with those in the database. Dependence of the comparison against known sequences hinders the development of discovery-based approaches.

De Novo sequencing allows the subtraction of sequence information directly from MS/MS spectra without the need to confront the data against a pre-established database (Dancik et al., 1999). Programs developed for *de novo* sequencing started in the late 1980s. At the time several computational problems needed to be overcome. First, the algorithms needed to overcome the limitations of being instrument or method specific. Ion fragmentation is conducted by different techniques. Collision-induced dissociation (CID), electron capture dissociation (ECD), infrared multiphoton dissociation (IRMPD) and the combination of these are some of the techniques used for the fragmentation of ions to extract sequencing data. The difference among them is not only in the method used, but also in the fragmentation efficiency of a particular sequence. This brings the second computational problem. How do you assign sequence information to incomplete fragmentation products? Not a trivial question to address even with today's advances in computational biology. Thirdly, the scoring system used to determine levels of confiability to select the correct sequence from all possible alternatives needs to be robust and verifiable. Needless to say, new algorithms need to overcome these three major problems.

A new generation of *de novo* sequencing algorithms has become available. DeNoS in 2005 claimed to be the first algorithm to sequence peptides with >95% reliability (Savitski et al. 2005). Other more commonly known algorithms are PEAKS and Lutefisk, which can be used to analyze data obtained from both CID and ECD fragmentation (B. Ma et al., 2003; Johnson & Taylor, 2002). Similarly, newly developed algorithms such as ADEPTS and ScanRanker provide a platform not only to extract unassigned MS/MS spectra but also to increase the confidence of peptide sequence alignments (He & B. Ma, 2010; Z.Q. Ma et al., 2011). Ultimately, "shotgun protein sequencing" takes advantages of sequence alignment to group unidentified MS/MS spectra and identify modified and unmodified peptide variants, generating a new method to study unknown proteins from tandem mass spectrometry approaches (Bandeira, 2011). Undoubtedly, these approaches should serve to strengthen discovery-based approaches to identifying conserved and novel proteins from extinct and living organisms from different taxa.

5.2 Identification of biologically active peptides

As with most truly powerful scientific approaches, the immediate future of tandem mass spectrometry is studies in which the sophisticated user applies his/her knowledge and expertise to cross-disciplinary investigations of regulatory and physiological processes that are now in reach. Routinely these studies provide insight as to the nature of evolutionarily conserved as well as evolutionarily variable biochemical processes. The advancement of the

specific field of enquiry as well as linking to molecular and biochemical approaches provides insight into how evolution works. A relatively simple example of this kind of problem is the processing of pro-forms of proteins into their active protein and signaling peptides. As described in sections 2.1 and 3.4, proteolytic activation of inactive proteins is the basis for a wide range of physiological processes including blood coagulation, the complement reaction, fibrinolysis, dorsal-ventral patterning in drosophila, and adhesion in barnacles. At one level these are some of the best understood processes in physiology and biology. At the evolutionary level they are a frontier.

The following example of the potential cross disciplinary use of Mass Spectrometry is signaling from marine chemical ecology. The signals are based upon the action of trypsin like serine proteases, evolutionarily ancient and highly conserved enzymes, which are important in sequencing proteins in Tandem Mass spectrometry. The signal molecules are pheromones or resource cues that are found in at least 4 phyla of marine invertebrates and that have strong similarity if not homology to the vertebrate complement cascade (Pettis et al., 1993; Rittschof & Cohen, 2004). These signaling peptides are all peptides whose carboxyl terminal sequences end in arginine or lysine. The peptides coordinate a wide variety of processes such as prey location by gastropods, hermit crab shell location, larval release in decapod crustaceans, and induction of larval settlement in barnacles, and have been called keystone molecules by some ecologists (Zimmer & Butman, 2000).

Although these peptides have been known to exist for over 30 years, none have been sequenced. Moreover, only in the case of barnacle settlement pheromones have the precursors to the peptides, in this case, settlement inducing protein complex (SIPC; Dreanno et al., 2006a; 2006b; 2007) been identified. These peptides which are active at nM (10^{-9}) to pM (10^{-12}) should be low hanging fruit for modern techniques which have aM (10^{-18}) sensitivity. Knowing sequences and identifying precursors would enable the development of theory of the evolution of peptide signaling and shed light on the evolution of activation of the pro-forms of proteins, including which processes are highly conserved and which can tolerate change and enable diversity to evolve. Cross-disciplinary interest would include proteomics, enzymology, biochemistry, physiology, neurophysiology, behavior and ecology.

5.3 Identification of posttranslational modifications: A case for substituted amino sugars

Another area with high potential for the application of Tandem Mass Spectrometry is sequencing and determining substituted structures of complex amino sugar polymers associated with glucose amino glycans, such as heparin, chondroitan sulfate, related polymers and their products. These polymers from vertebrates are a hot topic in materials science (Ornitz, 2000). The related structures in estuarine and marine fish and invertebrates are the source of substituted amino sugar signal molecules (Forward & Rittschof, 2000; Rahmen et al., 2000). Again, in marine systems the actual structure of the native signal molecules has not been determined. Rather, indirect enzymatic methods and purified molecules from higher vertebrates, have been used in conjunction with bioassays to provide insight as to the structure of active signals.

We see an interesting parallel between studying complex amino sugar polymers and peptide sequencing. Sequencing of complex amino sugars is more difficult than peptide sequencing because there are usually at least two kinds of common linkages that can be degraded by enzymes and the sugars themselves can be substituted at a variety of sites with

sulfates or acetates (Rittschof & Cohen, 2004). The first step in characterizing the sugar back bone is to strip off the substituted groups. This is frustrating as these substitutions confer biological activity. If techniques were developed using enzymes to sequence the polymers with their substituents, one could begin to discern the evolutionary relationships between the classes of substituted sugars and their conserved and variable regions.

Glucose amino glycans (GAG's) are a very important group of structural polymers which have evolved signaling functions. They are the ultimate in post synthesis modification of proteins and promise to entertain researchers for decades. Understanding their structures, their conserved and variable components and their relations throughout biology is a formidable challenge that will require coordinated use of techniques and approaches from a wide range of disciplines, from medicine and agriculture to materials science and chemical engineering. Mass spectrometry would inform the enquiry.

5.4 Funding issues

The use of tandem mass spectrometry in molecular evolution studies relies on a multidisciplinary and interdisciplinary approach. As it is obvious from all the issues discussed above, tandem mass spectrometry transcends many disciplines that range from computational biology and bioinformatics to chemistry and biochemistry. The integration of all of this knowledge is required to fulfill all of the capabilities that this technology puts in our hands. Therefore, the complex nature of tandem mass spectrometry as a tool is a cohesive force among scientists from diverse disciplines, whose professional goals and motivation is discovery and pushing forward the boundaries of knowledge. However, there is always a force in the opposite direction that pulls scientists away from each other, FUNDING.

Why should funding agencies consider investing in the improvement of tools for molecular evolution studies? First, the discovery of conserved mechanisms could provide valuable information to dissect the pathways that are crucial for cellular function. This information is essential for understand the pathophysiology of diseases that leads to cell death. Additionally, drug development could benefit from this information since highly conserved proteins could be "hubs" in a protein network involved in a variety of cellular processes (Zhu et al., 2007). Second, the identification of conserved proteins paves the way to dissect differences among organisms of different taxa, contributing to the understanding of divergence and speciation. The dissected differences may contribute to the understanding of molecular processes in a specific organism. Third, identification of conserved proteins and, consequently, molecular mechanisms may uncover new experimental tools and applications that benefit human kind. From the Dickinson et al. (2009) example, which showed similarities between barnacle cement and human blood coagulation, elucidating the proteome involved in barnacle adhesion may allow this organism to be used as an invertebrate model to discover and validate anti-coagulant drugs and to develop biomedical glues that could be later used in humans. This knowledge can also be employed in *preventing* adhesion of barnacles; the settlement of barnacles on ships and other man-made objects in the sea (i.e. biofouling) cost maritime industries billions of dollars annually. Anticoagulants that have been well-developed by the pharmaceutical industry may prove to effectively prevent barnacle adhesion when employed in a ship paint, which would greatly improve fuel efficiency (Rittschof et al., 2011).

This is not a request without precedent. NIH invested $30 million in a project to unravel the human brain's connections. This effort brought together neuroscientists, microscopists,

chemists, biochemists, neurologists, radiologists, and others in order to construct the connection map that directs our mind. A similar effort to unravel the mysteries of molecular evolution could bring together evolutionists, developmental biologists, mass spectrometrists, bioinformaticians, computational biologists, statisticians, biochemists, cellular biologists, chemists and neuroscientists. The benefits of such an effort go beyond our inclination to understand the past, transcending to better understand our present.

6. Conclusions

In this chapter, the use of tandem mass spectrometry in molecular evolution studies has been discussed from a user perspective. Important considerations, from sample preparation to data mining, were explained to bear in the mind of those that want to venture on in these studies and to entice the creativity of the expert to push the boundaries forward. The examples used are recent debates and discoveries using hypothesis-based approaches to understand the past and conserved molecular mechanisms. Although controversial, the identification of conserved proteins from extinct organisms led us to think that, as stated above, our *song* has been modified by changes in *tempo* to adjust in space and time. Essential mechanisms that render the organism more fit to survive or respond to its environment open our understanding on how conservation is an integral component of evolution. In the future, scientists will continue joining forces in the quest to develop more instrumental and bioinformatics tools to uncover the mysteries of the past for the benefit of the present and future.

7. Acknowledgements

IEV acknowledge the support of NIH-COBRE grant 5P20RR016439 and UPR Institutional funds that contributed to the establishment of the Protein Mass Spectrometry Facility, and NIH-NINDS funds (SC1NS066988) that support, in part, this work. DR and GHD acknowledge support of the U.S. Office of Naval Research, Coatings and Biofouling Program, for continued support of interdisciplinary research.

8. References

Adami, C., Ofria, C., & Collier, T.C. (2000) Evolution of biological complexity. *Proceedings of the National Academy of Sciences of the United States of America,* 97, pp. 4463-4468

Asara, J.M., Schweitzer, M.H., Freimark, L.M., Phillips, M., & Cantley, L.C. (2007) Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science*, 316, pp. 280-285

Bandeira, N. (2011) Protein Identification by spectral networks analysis. *Methds Mol. Biol.* 694, pp. 151-168

Bi, E., & Lutkenhaus, J. (1991) FtsZ ring structure associated with division in *Escherichia coli*. *Nature*, 354, pp. 161-164

Boekhorst, J., van Breukelen, B., Heck, A.J.R., & Snel., B. (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biology,* 9, pp. R144.

Boot-Handford, R.P. & Tuckwell, D.S. (2003). Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest. *BioEssays,* 25 pp. 142-151

Boutilier, K., Ross, M., Podtelejnikov, A.V., Orsi, C., Taylor, R., Taylor, P., & Figeys, D. (2005) Comparison of different search engines using validated MS/MS test datasets. *Analytica Chimica Acta*, 534, pp. 11-20

Budovskaya, Y.V., Stephan, J.S., Deminoff, S.J. & Herman, P.K. (2005) An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase. *Proceedings of the National Academy of Sciences of the United States of America,* 102, pp. 13933-139398

Buckley, M., Walker, A., Ho, S. Y. W., Yang, Y., Smith, C., Ashton, P., Thomas Oates,J., Cappellini, E., Koon, H., Penkman, K., Elsworth, B., Ashford, D., Solazzo, C., Andrews, P., Strahler, J., Shapiro, B., Ostrom, P., Gandhi, H., Miller, W., Raney, B., Zylber, M.I., Gilbert, M.T.P., Prigodich, R.V., Ryan, M., Rijsdijk, K.F., Janoo, A., & Collins, M.J. (2008) Comment on "Protein sequence from Mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science*, 319, pp. 33

Colinge, J., & Bennett, K.L. (2007) Introduction to Computational Proteomics. *PLoS Computational Biology*, 3, pp. 1151-1160

Cottrell, J.S. (2011) Protein identification using MS/MS data. *Journal of Proteomics*, 74, pp. 1842-1851

Coyne, K.J., Xiao-Xia, Q., & Waite, J.H. (1997). Collagen in mussel byssus: a natural block copolymer. *Science*, 277, pp. 1830-1832

Dancik, V., Addona, T.A., Clauser, KR., Vath, J.E., & Pevzner, P.A. (1999) De Novo sequencing via tandem mass spectrometry. *J. Computational Biol.*, 6, pp. 327-342

Davie, E. (2003) JBC Centennial 1905-2005: 100 years of biochemistry and molecular biology. A brief historical review of the waterfall/cascade of blood coagulation. *Journal of Biological Chemistry,* 278, pp. 50819-50832

Davie, E. W., & Rantoff, O. D. (1964) Waterfall sequence for intrinsic blood clotting. *Science,* 145, pp. 1310-1312

Day, J.J., & Sweatt, J.D. (2011) Epigenetic mechanisms in cognition. *Neuron*, 70, pp. 813-829

Desai, A., & Mitchison T.J. (1998) Tubulin and FtsZ structures: functional and therapeutic implications. *Bioessays*, 20, pp. 523-527

Dick, D.M. (2011) Gene-environment interaction in psychological traits and disorders. *Annual Review of Clinical Psychology.* 7, pp. 383-409

Dickinson, G.H., Vega, I.E., Wahl, K.J., Orihuela, B., Beyley, V., Rodriguez, E.N., Everett, R.K., Bonaventura, J., Rittschoff, D. (2009) Barnacle cement: a polymerization model based on evolutionary concepts. *Journal of Experimental Biology*, 212, pp. 3499-3510

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., & Douzery, E.J. (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20, pp. 248-254

Duncan, M.W., Aebersold, R., & Caprioli, R.M. (2010) The pros and cons of peptide-centric proteomics. *Nature Biotechnology,* 28, pp. 659-664

Dreanno, C., Matsumura, K., Dohmae, N., Takio, K., Hirota, H., Kirby, R.R. & Clare, A.S. (2006a) An alpha(2)-macroglobulin-like protein is the cue to gregarious settlement of the barnacle *Balanus amphitrite*. *Proceedings of the National Academy of Sciences of the United States of America*, 103, pp. 14396-14401

Dreanno, C., Kirby, R.R., & Clare, A.S. (2006b) Locating the barnacle settlement pheromone: spatial and ontogenetic expression of the settlement-inducing protein complex of

Balanus amphitrite. *Proceedings of the Royal Society B-Biological Sciences*, 273, pp. 2721-2728

Dreanno, C., Kirby, R.R., & Clare, A.S. (2007) Involvement of the barnacle settlement-inducing protein complex (SIPC) in species recognition at settlement. *Journal of Experimental Marine Biology and Ecology,* 351, pp. 276-282

Durliat, M., & Vranckx, R. (1981) Action of various anticoagulants on hemolymphs of lobsters and spiny lobsters. *Biological Bulletin* 160, pp. 55-68

Edman, P. (1950) Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica*, 4, pp. 283-293

Edwards, N.J. (2011) Protein identification from tandem mass spectra by database searching. *Methods in molecular biology*, 694, pp. 119-138

Findlay, G.D., & Swanson, W.J. (2010) Proteomics enhances evolutionary and functional analysis of reproductive proteins. *Bioessays,* 32, pp. 26-36

Forward, R.B., & Rittschof, D. (2000) Alteration of photoresponses involved in diel vertical migration of a crab larva by fish mucus and degradation products of mucopolysaccharides. *Journal of Experimental Marine Biology and Ecology,* 245, pp. 277-292

Fuller, G. M., & Doolittle, R. F. (1971) Studies of invertebrate fibrinogen. II. Transformation of lobster fibrinogen into fibrin. *Biochemistry*, 10, pp. 1311-1315

Garrone, R. (1999). Collagen, a common thread in extracellular matrix evolution. *Proceedings of the Indian Academy of Sciences: Chemical Sciences*, 111, pp. 51-56

Goodman, C.S., & Coughlin, B.C. (2000) The evolution of evo-devo biology. *Proceedings of the National Academy of Sciences of the United States of America,* 97, pp. 4424-4425

Graumann, P.L. (2004) Cytoskeletal elements in bacteria. *Current Opinion in Microbiology*, 7, pp. 565-571

He, L., & Ma, B. (2010) ADEPTS: advanced peptide de novo sequencing with a pair of tandem mass spectra. *J. Bioinform Comput. Biol.,* 8, pp. 981-994

Heffer, A., Shultz, J.W., & Pick, L. (2010) Surprising flexibility in a conserved Hox transcription factor over 550 million years of evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 107, pp. 18040-18045

Human Genome Project (March 26, 2008) The science behind the human genome project, In: *From the Genome to the Proteome*, August 25, 2011, Available from: http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml

Hunt, D.F., Yates III, J.R., Schabanowitz, J., Winston, S., & Hauer C.R. (1986) Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 83, pp. 6233-6237

Johnson, R.S., & Taylor, J.A. (2002) Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Molecular Biotechnology,* 22, pp. 301-315

Kappeler, L., & Meaney, M.J. (2010) Epigenetics and parental effects. *Bioessays,* 32, pp. 818-827

Khan, H.A., Arif, I.A., Bahkali, A.H., Al Farhan, A.H., & Al Homaidan, A.A. (2008) Bayesian, maximum parsimony and UPGMA models for inferring the phylogenies of antelopes using mitochondrial markers. *Evolutionary Bioinformatics Online*, 4, pp. 263-270

Kopacek, P., Hall, M., & Soderhall, K. (1993) Characterization of a clotting protein, isolated from plasma of the fresh water crayfish *Pacifastacus leniusculus*. *European Journal of Biochemistry*, 213, pp. 591-597

Kumar, C., & Mann, M. (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets, *FEBS Letters*, 583, pp. 1703-1712

Levin, H.L., & Moran, J.V. (2011) Dynamic interactions between transposable elements and their hosts. *Nature Reviews Genetics*, 12, pp. 615.627

Liska, A.J., & Shevchenko, A. (2003) Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, 3, pp. 19-28

Lorand, L. (1972) Fibrinoligase: the fibrin-stabilizing factor system of blood plasma. *Annals of the New York Academy of Sciences*, 202, pp. 6-30

Lubec, G., & Afjehi-Sadat L. (2007) Limitations and pitfalls in protein identification by mass spectrometry, *Chemical Reviews*, 107, pp. 3568-3584

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003) PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry,* 17, pp. 2337–2342

Ma, Z.Q., Chambers, M.C. Ham, A.J., Cheek, K.L., Whitwell, C.W., Aerni, H.R., Schilling, B., Miller, A.W., Caprioli, R.M., & Tabb, D.L. (2011) Scan Ranker: Quality assessment of tandem mass spectra via sequence tagging. *Journal of Proteome Research*, 10, pp. 2896-2904

MacFarlane, R. G. (1964) An enzyme cascade in the blood clotting mechanism, and its function as a biochemical amplifier. *Nature,* 202, pp. 498-499

Madaras, F., Chew, M. Y., & Parkin, J. D. (1981) Purification and characterization of the sand crab (*Ovalipes bipustulatus*) coagulogen (fibrinogen). *Thrombosis and Haemostasis*, 45, pp. 77-81

Mann, M., & Kelleher, N.L. (2008) Precision proteomics: the case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences of the United States of America*, 105, pp. 18132-18138

Marshall, E. (2011) Waiting for the Revolution. *Science*, 331, pp. 526-529

Marshall, J.L., Huestis, D.L., Garcia, C., Hiromasa, Y., Wheeler, S., Noh, S., Tomich, J.M., Howard, D.J. (2011) Comparative proteomics uncovers the signature of natural selection acting on the ejaculate proteomes of two cricket species isolated by postmating, prezygotic phenotypes. *Molecular Biology and Evolution,* 28, 423-435

Muta, T., & Iwanaga, S. (1996) The role of hemolymph coagulation in innate immunity. *Current Opinion in Immunology*, 8, pp. 41-47

Nesvizhskii, A.I., Keller, A., Kolker, E., & Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 75, pp. 4646-4658

Neurath, H. (1984) Evolution of proteolytic enzymes. *Science,* 224, pp. 350-357

Neurath, H. (1986) The versatility of proteolytic enzymes. *Journal of Cellular Biochemistry,* 32, pp. 35-49

Neurath, H. (1999) Proteolytic enzymes, past and future. *Proc. Natl. Acad. Sci.,* 96, pp. 10962-10963

Neurath, H., & Walsh, K. A. (1976) Role of proteolytic enzymes in biological regulation (a review). *Proceedings of the National Academy of Sciences of the United States of America,* 73, pp. 3825-3832

Ogden, T.H., & Rosenberg, M.S. (2006) Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology,* 55, pp. 314-328

Osaki, T., & Kawabata, S. (2004) Structure and function of coagulogen, a clottable protein in horseshoe crabs. *Cellular and Molecular Life Sciences*, 61, pp. 1257-1265

Organ, C.L., Schweitzer, M.H., Zheng. W., Freimark, L.M., Cantley, L.C., & Asara, J.M. (2008) Molecular phylogenetics of mastodon and *Tyrannosaurus rex*. *Science,* 320, pp. 499

Ornitz, D.M. (2000) FGFs, heparin sulfate and FGFRs: complex interactions essential for development. *Bioessays*, 22, pp. 108-112

Patternson, S.D., & Aerbersold, R.H. (2003) Proteomics: the first decade and beyond. *Nat. Genet*., 33 Suppl., pp. 311-323

Patthy, L. (1993) Modular design of proteases of coagulation, fibrinolysis, and complement activation: implications for protein engineering and structure - function studies. Pages 10-22 in L. Lorand and K. G. Mann, editors. Methods in Enzymology. Academic Press, Inc., San Diego.

Pettis, R.J., Erickson, B.W., Forward, R.B. & Rittschof, D. (1993) Superpotent synthetic tripeptide mimics of the mud-crab pumping pheromone. *International Journal of Peptide and Protein Research*, 42, pp. 312-319

Pevzner, P.A., Kim, S., & Ng, J. (2008) Comment on "Protein sequence from Mastodon and Tyrannosaurus rex revealed by mass spectrometry. *Science*, 321, pp. 1040

Pisano, J. J., Finlayso, J. S., & Peyton, M. P. (1968). Cross-link in fibrin polymerized by factor XIII epsilon-(gamma-glutamyl)lysine. *Science* 160, pp. 892-89

Pisano, J. J., Finlayso, J. S. & Peyton, M. P. (1969). Chemical and enzymatic detection of protein cross-links. Measurement of epsilon-(gamma-glutamyl)lysine in fibrin polymerized by factor XIII. *Biochemistry* 8, pp. 871-876

Prachumwat, A., & Wen-Hsiung, L. (2008) Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. *Genome Research*, 18, pp. 221-232

Qi, L., & Cho, Y.A. (2008) Gene-environment interaction and obesity. *Nutrition Reviews,* 66, pp. 684-694

Rahman, Y.J., Forward, R.B., & Rittschof, D. (2000) Responses of mud snails and periwinkles to environmental odors and disaccharide mimics of fish odor. *Journal of Chemical Ecology*, 26, pp. 679-696

Ramos-Fernandez, A., Paradela, A., Navajas, R., & Albar, J.P. (2008) Generalized method for probability-based peptide and protein identification from tandem mass spectrometry data and sequence database searching. *Molecular and Cellular Proteomics*, 7, pp. 1748-1754

Ramm, S.A., McDonald, L., Hurst, J.L., Beynon, R.J., & Stockley, P. (2009) Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition. *Molecular Biology and Evolution,* 26, pp. 189-198

Rittschof, D., Cohen, J.H. (2004) Crustacean peptide and peptide-like pheromones and kairomones. *Peptides,* 25, pp. 1503-1516

Rittschof, D., Dickinson, G.H., Orihuela, B., & Holm, E.R. (2011) Anticoagulants as antifouling agents. US Pub. No. 2011/0041725 A1.

Sadreyev, R., & Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *Journal of Molecular Biology*, 326, pp. 317-336

Sanchez, J.C., Corthals, G.L., & Hochstrasser, D.F., editors (2004) Biomedical applications of proteomics. Weinheim: Wiley-VCH. 435 p.

Savitski, M.M., Nielsen, M. L., Kjeldsen, F., & Zubarev, R. A. (2005) Proteomics-Grade de Novo Sequencing Approach. *Journal of Proteome Research,* 4, pp. 2348–2354

Schweitzer, M.H., Suo, Z., Avci, R., Asara, J.M., Allen, M.A., Arce, F.T. & Homer J.R. (2007) Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. *Science*, 316, pp. 277-280

Shoulders, M.D., & Raines, R.T. (2009). Collagen structure and stability. *Annual Review of Biochemistry*, 78, pp. 929-958

Soderhall, K. (1981) Fungal cell-wall beta-1,3-glucans induce clotting and phenoloxidase attachment to foreign surfaces of crayfish hemocyte lysate. *Developmental and Comparative Immunology,* 5, pp. 565-573

Stretton, A.O.W. (2002) The first sequence: Fred Sanger and Insulin. *Genetics*, 162, pp. 527-532

Sritunyalucksana, K., & Soderhall, K. (2000) The proPO and clotting system in crustaceans. *Aquaculture,* 191, pp. 53-69

Tanaka, K., Hiroaki, W., Yutaka, I., Satoshi, A., Yoshikazu, Y., Tamio, Y., & Matsuo, T. (1988) Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2, pp. 151-153

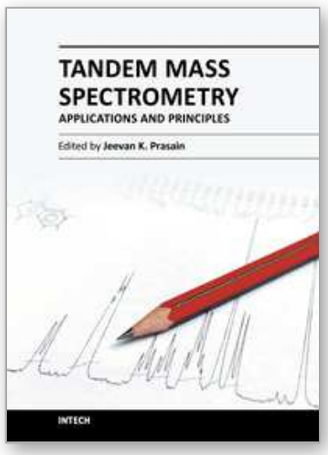
Theopold, U., Schmidt, O., Soderhall, K., & Dushay, M. S. (2004) Coagulation in arthropods: defense, wound closure and healing. *Trends in Immunology,* 25, pp. 289-294

Tzeng, J.Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M.I., Sale, M.M., Worrall, B.B., Hsu, F.C., Thomas, D.C. & Sullivan, P.F. (2011) Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *American Journal of Human Genetics*, 89, pp. 277-288

Wang, R., Liang, Z., Hall, M. & Soderhall, K. (2001) A transglutaminase involved in the coagulation system of the freshwater crayfish, *Pacifastacus leniusculus*. Tissue localisation and cDNA cloning. *Fish and Shellfish Immunology*, 11, pp. 623-637

Zhu, X., Gerstein, M., & Snyder M. (2007) Getting connected: analysis and principles of biological networks. *Genes and Development*, 21, pp. 1010-1024

Zimmer, R.K., & Butman, C.A. (2000) Chemical signaling processes in the marine environment. *Biological Bulletin*, 198, pp. 168–87



Tandem Mass Spectrometry - Applications and Principles

Edited by Dr Jeevan Prasain

ISBN 978-953-51-0141-3
Hard cover, 794 pages
Publisher InTech
Published online 29, February, 2012
Published in print edition February, 2012

Tandem Mass Spectrometry - Applications and Principles presents comprehensive coverage of theory, instrumentation and major applications of tandem mass spectrometry. The areas covered range from the analysis of drug metabolites, proteins and complex lipids to clinical diagnosis. This book serves multiple groups of audiences; professional (academic and industry), graduate students and general readers interested in the use of modern mass spectrometry in solving critical questions of chemical and biological sciences.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Irving E. Vega, Dan Rittschof, Gary H. Dickinson and Ian Musgrave (2012). Evolutionary Proteomics: Empowering Tandem Mass Spectrometry and Bioinformatics Tools for the Study of Evolution, Tandem Mass Spectrometry - Applications and Principles, Dr Jeevan Prasain (Ed.), ISBN: 978-953-51-0141-3, InTech, Available from: http://www.intechopen.com/books/tandem-mass-spectrometry-applications-and-principles/evolutionary-proteomics-empowering-tandem-mass-spectrometry-and-bioinformatics-tools-for-the-study-o

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen