Communicating Research to the General Public

At the March 5, 2010 UW-Madison Chemistry Department Colloquium, the director of the Wisconsin Initiative for Science Literacy (WISL) encouraged all Ph.D. chemistry candidates to include a chapter in their Ph.D. thesis communicating their research to non-specialists. The goal is to explain the candidate's scholarly research and its significance to a wider audience that includes family members, friends, civic groups, newspaper reporters, state legislators, and members of the U.S. Congress.

Ten Ph.D. degree recipients have successfully completed their theses and included such a chapter, less than a year after the program was first announced; each was awarded $500.

WISL will continue to encourage Ph.D. chemistry students to share the joy of their discoveries with non-specialists and also will assist in the public dissemination of these scholarly contributions. WISL is now seeking funding for additional awards.

Wisconsin Initiative for
Science Literacy

The dual mission of the Wisconsin Initiative for Science Literacy is to promote literacy in science, mathematics and technology among the general public and to attract future generations to careers in research, teaching and public service.

UW-Madison Department of Chemistry
1101 University Avenue
Madison, WI 53706-1396
Contact: Prof. Bassam Z. Shakhashiri
bassam@chem.wisc.edu
www.scifun.org

January 2011

Strategies for Protein and Peptide Characterization and Quantification using Electron-Transfer Dissociation Mass Spectrometry and Intrinsic Fluorescence

By

Jason D. Russell

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2012

Date of final oral examination: 6/21/12

This dissertation is approved by the following members of the Final Oral Committee:

Joshua J. Coon, Associate Professor, Chemistry
Lloyd M. Smith, Professor, Chemistry
Lingjun Li, Associate Professor, Chemistry
Richard D. Vierstra, Professor, Genetics
David J. Pagliarini, Assistant Professor, Biochemistry

Chapter 7

Thesis summary for non-specialists in collaboration with the Wisconsin Initiative for Science Literacy (WISL): Peptide and protein analysis by tandem mass spectrometry

Introduction

In less than 60 years since the discovery of the structure of DNA (deoxyribonucleic acid), the entire human genome, some 3-billion base pairs (3,000,000,000), has been sequenced providing the scientific community with a blueprint for studying heredity, disease, and fundamental biological processes.[1-5] Our genomes contain all of the heritable biochemical instructions needed for life. This information, contained in DNA, is passed on to us, in equal parts, from our mothers and fathers. Perhaps, one of the most surprising discoveries from the sequencing of the human genome was that there are only ~25,000 protein-coding genes. Genes are the portions of DNA that contain the chemical instructions for building proteins. In the years leading up to the publication of the human genome, it was thought that humans expressed at least 50,000-100,000 genes.[6] It was once believed that the complexity of an organism was the product of the number genes in its genome. It turns out that the number of genes does not necessarily correlate with the size of an organism, nor its biological complexity – and biological complexity itself difficult to define. For example, yeast (*Saccharomyces cerevisiae*) is a unicellular organism and has ~6000 protein-coding genes; while *Giardi lamblia*, a multicellular parasite, contains nearly the same number.[7,8] A nematode worm (*Caenorhabditis elegans*) has ~19,000 protein-coding genes, slightly less that of a domestic dog (19,300, *Canis lupus familiaris*).[9,10] So, while DNA is vital, it is not the only important biomolecule (contrary to what you may hear on contemporary television programs).

During the mid-to-late 20th century, scientists formulated what is famously referred to as the "central dogma of molecular biology".[11] In its simplest form, the central dogma states that the protein-coding regions (genes) of the genetic code contained in DNA is transferred to RNA (ribonucleic acid) through the process of DNA transcription. RNA is used as a template to make proteins through the process of RNA translation. This process is considered one way; proteins cannot naturally be reverse engineered and inserted back into the genetic code. An illustration of the relationship between DNA, genes, and protein is shown in **Figure 1**.[12]

What are proteins and why are they important?

Proteins are made of long chains of amino acids chemically linked together and folded into very specific three-dimensional structures. Proteins are required for nearly every cellular process and are the primary functional expression of an organism's genome. Proteins catalyze important chemical reactions, provide structural support, and are the cell's primary means of chemical signaling. Proteins are found in the cell, embedded within the cell and nuclear membranes, and are also found in extracellular fluids. Proteins are responsible for phenotype, characteristics that are directly observable or

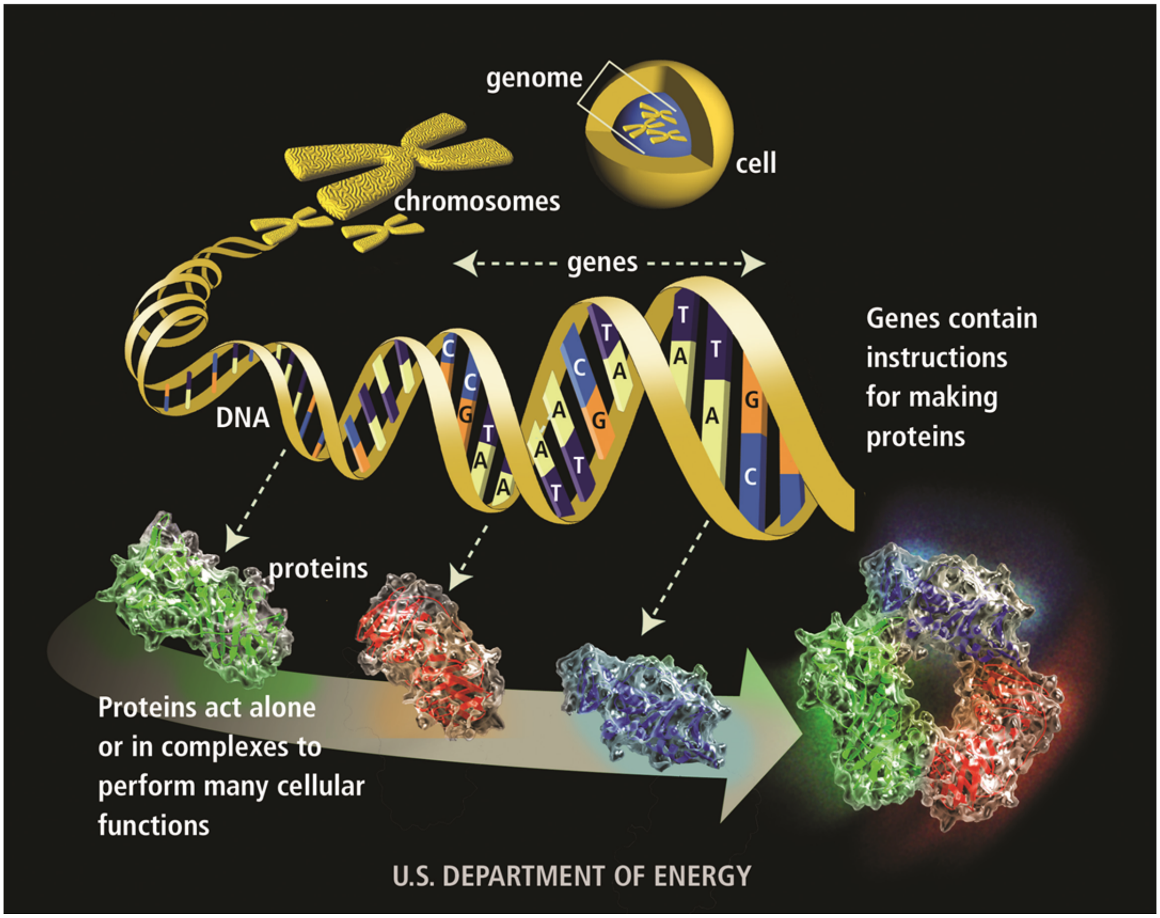

genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

Figure 1. The relationship between the genome, genes, and proteins. The genome contains all of an organism's genetic material. In humans, this material is stored in 23 pairs of chromosomes. The parts of the genome that codes for protein are called genes. Genes are transcribed from DNA to produce RNA which is then translated into protein. *Image courtesy of the Department of Energy Genomic Science program, http://genomicscience.energy.gov*[12]

can be measured through experimentation (*e.g.,* eye color, blood type, symptoms of Alzheimer's disease).

The genome, specifically genes, contains the chemical code for the type and order of amino acids for each protein. As an analogy, your genome can be thought of as a rough draft of a novel (your parents are co-authors). This rough draft contains all of the text needed for a fantastic story, but it needs serious editing. There are large portions of your genomic rough draft that contains material not essential to the story (less than 2% of the human genome codes for protein!).[3] Once edited, this rough draft becomes a proper novel with chapters that, in the grand scheme of the novel, serve in the telling of the main story. This novel represents all of your protein-coding genes and each chapter represents a specific biological function. Each paragraph of the novel can be thought of as a single protein where each word is a peptide (a portion of a protein) and each letter represents an amino acid. Paragraphs work together in each chapter to further the story, much like proteins work together to carry out a biological function. Chapters of the novel intertwine and culminate in what one hopes is an entertaining story; whereas, the assemblage of biological functions carried out by proteins are needed to sustain life.

Where does protein diversity arise? Protein species (proteoforms) can arise from the many chemical modifications known to occur on proteins, or the manner a gene is interpreted to create a protein (alternative splicing). There are also events that occur to change the DNA sequence in a gene region (mutations, substitutions) to create protein isoforms. These events can push the total number of the number of proteoforms and protein isoforms into the millions despite that there may be only 25,000 genes used to make protein.[13,14] Most of the changes to protein cannot be predicted or identified by studying the genome and require the direct analysis of protein.

Back to the novel example, proteoforms can be viewed as additional rounds of editorial revisions. These revisions may entail the substitution or deletion of a letter, a word, or even appending an entire sentence to the beginning or end of a paragraph. Sometimes changing word or adding a sentence does not affect the meaning of a paragraph – a neutral editorial change. The majority of the time, however, these revisions help to communicate the meaning in one paragraph to the next to help the flow of the story. In extreme cases, a small change, or multiple small changes, can affect the entire story. This is true in biology as well. Changes to a protein that results in chemical modifications, or changes to its sequence, often do not impact the manner in which the protein functions. These proteoforms do not change the biological story. However, chemical modification of proteins frequently helps to communicate signals through the cell and throughout the body such as the sensing of heat, cold, and pain. In extreme cases, a change to just a single amino acid in a protein can have life-altering biological effects. This is the case with sickle-cell anemia, a painful disease that changes the shape of red-blood cells causing a number of deleterious physiological effects.[15]

Organismal complexity and diversity stems from this extraordinary ability to engineer millions of proteoforms from thousands of genes. A goal of proteomics is to characterize these proteoforms in great detail and relate them to biological function. My colleagues and I are interested in studying changes to proteins, both physical modifications and changes in amount, to better understand the dynamic cellular processes they mediate through their interactions with other biomolecules (other proteins, DNA, messenger RNA, metabolites, etc.). Key to realizing this goal is the development of technology and methodology is to unambiguously identify proteins through sequencing, to map protein modifications, and to track changes in protein expression in a parallelizable fashion that can be scaled to fit the scope of the biological question.

How does my research relate to the study of proteins?

My research over the last 4 ½ years has been focused on the analysis proteins and peptides. My expertise is in the field of proteomics, but I am an analytical chemist, not a biochemist or molecular biologist. My research is not centered on understanding the functions of proteins - I leave that to the expert biologists. My research instead has been geared towards developing analytical tools, techniques, and methods to facilitate the characterization of proteins, their chemical modifications, and their abundances featuring mass spectrometry. As an analytical chemist I am concerned with the science of identification and measurement. I want to know what kind of proteins are in my sample and how much of them I have. The instrument I use to accomplish these tasks is the mass spectrometer. The mass spectrometer allows me to measure the mass of proteins and peptides and determine their amino acid sequences and chemical modifications. The methodologies I've developed allow me to analyze thousands of proteins and peptides in a single day. Armed with these analytical tools, I aim to allow biologists to better answer questions about their favorite protein or biological system.

The mass spectrometer – a molecular balance

A mass spectrometer is a scientific instrument that is able to measure the mass of molecules. Mass spectrometers allow the analysis of extremely small amounts of material, as little as 100 attomoles. In terms of mass, this is about 100 femtograms (100 x 10^{-15} grams). To put this in perspective, 100 femtograms is a more than billion times less than the mass of a single grain of table salt. Can't imagine the scale of 1,000,000,000-to-1? A 1-foot by 1-foot square in an area that measures 6-miles by 6-miles would represent the same scale. To be more precise, a mass spectrometer measures a molecule's mass-to-charge ratio (commonly represented as *m/z*). The key is that molecules must possess a net charge in order to be measured by the mass spectrometer. Charged molecules are referred to as ions and can be either positively or negatively charged depending on the chemical characteristics of the molecule. Mass spectrometers rely on the use of electric and/or magnetic fields to manipulate ions under vacuum. A vacuum is required to minimize ions from colliding with gas which would hamper their manipulation and detection. Ion manipulation includes the transmission of ions through various parts of a mass spectrometer, breaking ions apart, and performing *m/z* analysis.

Mass spectrometers come in many different flavors, but at the heart of every mass spectrometer is the mass analyzer. An instrument schematic of a type of mass spectrometer used in my research is shown in **Figure 2**. This mass spectrometer is an LTQ Orbitrap Velos manufactured by Thermo Fisher Scientific. It is considered a hybrid mass spectrometer because it has two different types of mass analyzers, an ion trap and an Orbitrap. Ion traps are often used as mass analyzers because they are fast and sensitive. The Orbitrap is used if resolution and mass accuracy are more important. To measure the mass of ions, they are introduced into the spectrometer inlet (I'll describe this process later). Ions are then transmitted through the mass spectrometer using ion optics which funnels ions into one of the two mass analyzers using radiofrequency (RF) ion guides, focusing lenses, and differential direct current (DC) offsets along the length of the instrument. This mass spectrometer has two other ion trapping devices, the C-trap and the collision cell. These devices do not perform mass analysis, but are used to transmit ions to the Orbitrap (C-trap) or to perform tandem mass spectrometry (collision cell) which will be discussed in the next section. The signal from a mass spectrometer, typically an electric current, is converted into a format easily interpreted by humans - a mass spectrum. **Figure 3** shows a typical mass spectrum of a mixture of peptides. The y-axis is the relative abundance of each ion and the x-axis is the mass-to-charge ratio of each ion. This particular mass spectrum might contain 50 or more peptides.

Identifying peptides and proteins by tandem mass spectrometry

For many technical reasons, it is most often easier to analyze peptides instead of a whole protein. Imagine you are reading the "genomic" novel from the example above. There is a paragraph that is 1,000 letters long, but there are no spaces between the letters. It is very difficult to read the paragraph as one long word. Instead, the paragraph is broken up into words and the words are easily read one at a time. An analogous situation arises for protein analysis by mass spectrometry. Our current technical limitations make it difficult to measure a protein that is 1,000 amino acids long. However, we can easily measure

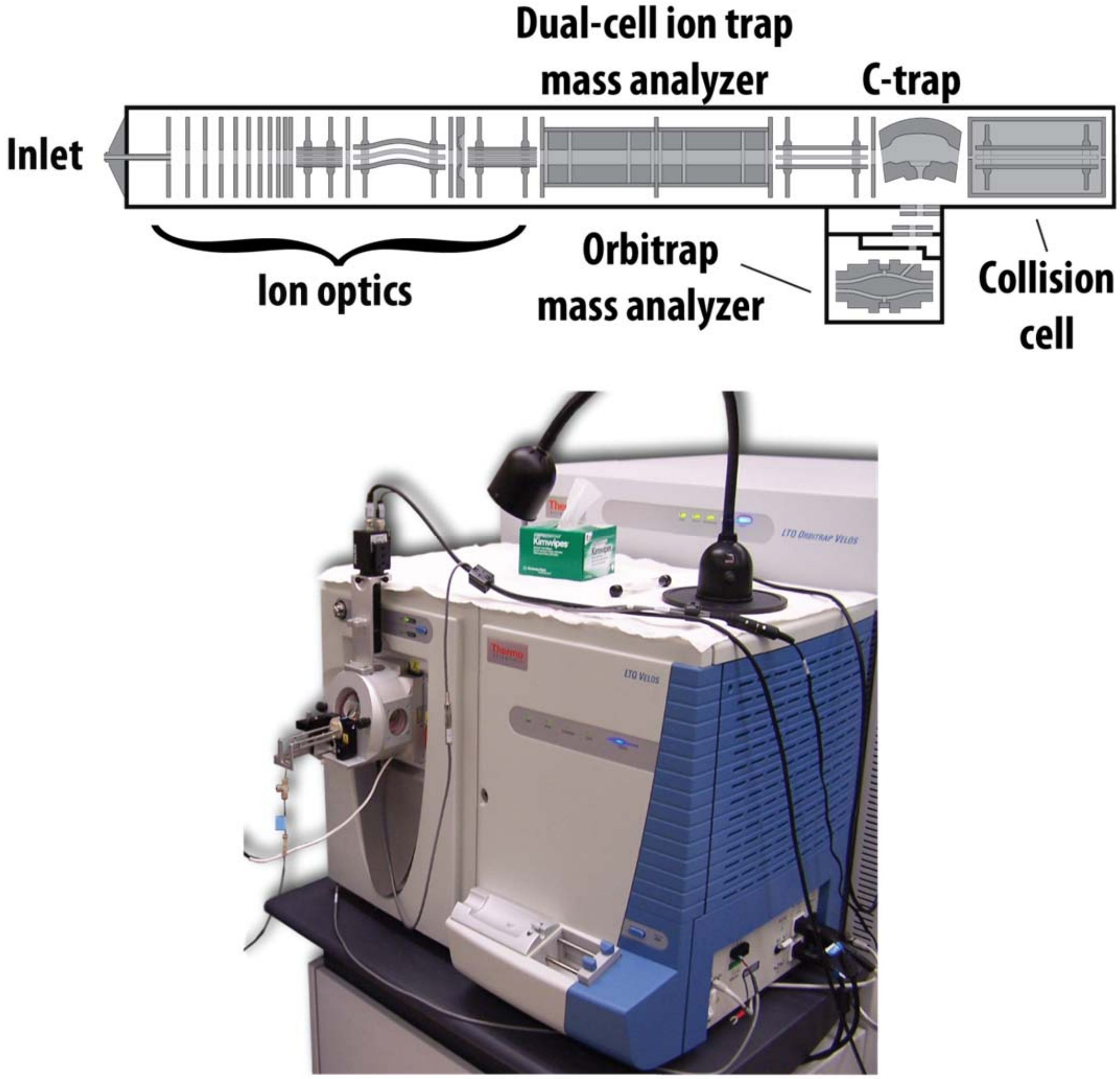


Figure 2. Instrument schematic of a LTQ Orbitrap Velos mass spectrometer manufactured by Thermo Fisher Scientific (top). Actual image of the mass spectrometer (bottom).

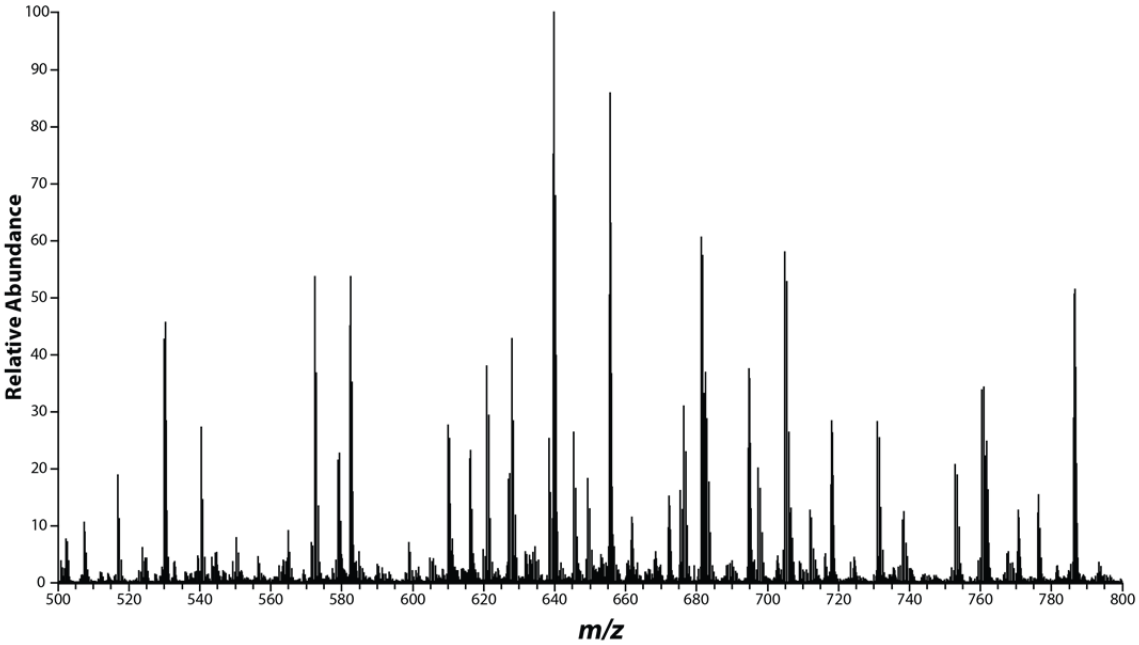
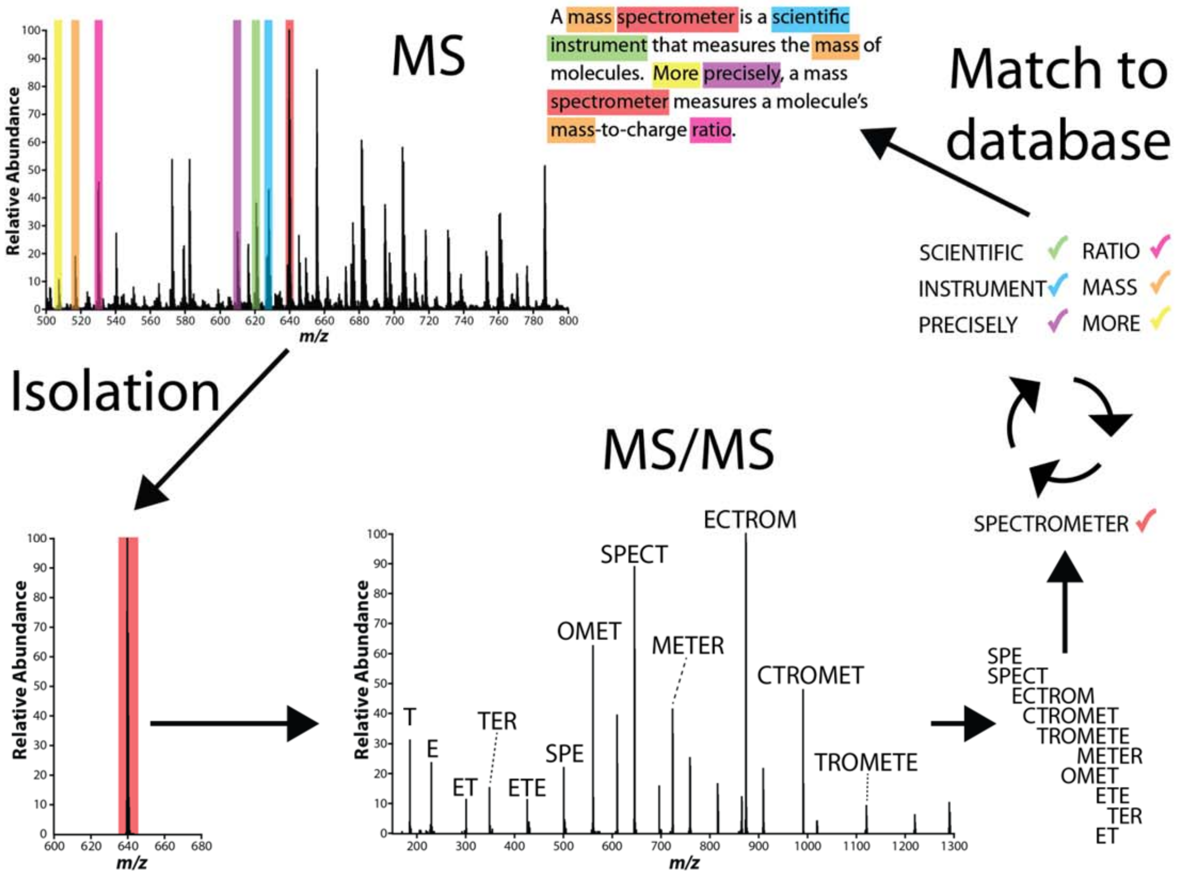


Figure 3. A mass spectrum of a mixture of peptides (*m/z* = mass-to-charge ratio). Each cluster of peaks represents a peptide. There are at least 50 peptides in this mass spectrum.

peptides that are 10-25 amino acids in length, akin to reading small words instead attempting to read a single word that is a paragraph long. Proteins can be broken down into peptides by using enzymes that cut along the protein in well-defined locations. This is analogous to proper spacing of the paragraph to define single words.

Simply measuring the mass of the peptide does not tell us very much. We need to know the amino acid sequence of the peptide to confirm its identity and determine the protein from which it was derived. This is accomplished by performing tandem mass spectrometry (MS/MS). First, peptide masses are recorded by the mass spectrometer (MS). Then, thousands of ions from a single peptide species can be isolated from all of the other peptides. This population of relatively pure peptide (referred to as precursor ions) can be subjected to collisions with trace amounts of gas in the mass spectrometer or reacted with other molecules causing them to break apart at predicable locations. The masses of these fragments (product ions) can be measured by the mass spectrometer producing a product ion mass spectrum ((MS/MS spectrum). Generally, a single peptide will break into two fragments. But, thousands of identical peptide ions are broken apart at different locations and measured simultaneously. This will hopefully produce fragments between each amino acid in the peptide. Once the fragments are identified, they can be reassembled to reveal the identity of the peptide from the ordering of amino acids. The MS/MS spectrum is like a fingerprint unique to the peptide sequence. Once the sequence of the peptide is known, it is often possible to determine the protein from which the peptide originated. This entire process might be more easily explained by going step-by-step through an example presented in **Figure 4.**



A mass spectrometer is a scientific instrument that measures the mass of molecules. More precisely, a mass spectrometer measures a molecule's mass-to-charge ratio.

Figure 4. Process of peptide identification by tandem mass spectrometry. Peptide masses (words) are measured and recorded. In this example, the peptide at *m/z* 640 (red) is isolated and subjected to MS/MS producing a product ion spectrum. The peptide fragments can be assembled and interpreted to yield the peptide sequence “SPECTRUM”. This process is repeated several times to identify more peptides. These peptides are compared are then used to identify the protein (paragraph) from which the peptides originated.

To relate this to the previous examples, peptides will be represented as words and the protein represented by a paragraph. A mixture of peptides (words) is analyzed by the mass spectrometer producing a mass spectrum (MS). There are seven peptide precursors that we are interested in sampling (each peptide precursor is a different color). To determine the peptide sequence of the precursor at *m/z* of 640, it is first isolated. This removes all of the other peptides and leaves a pure population (several thousand ions) of the peptide that occurs at *m/z* 640. The isolated peptide population is subjected to MS/MS and the masses of the product ions are recorded as the product ion spectrum. Some of the product ions are single amino acids (letters) and some are strings of amino acids. The product ions can be assembled like a puzzle to reveal that the peptide at *m/z* 640 is “spectrometer”. The process of isolation and fragmentation is repeated for other *m/z* values yielding the peptides “scientific”, “instrument”, “precisely”, “ratio”, “mass”, and “more”. These peptides (words) can be used to identify the protein (paragraph) from which they originated because it is unlikely another protein has all of these same peptides.

One of the major areas of my research was developing technology and methodology to make peptides fragment more efficiently. There were many product ions for the peptide identified as “spectrum”. This is not always the case. If this peptide produced fewer product ions we might not be so certain about assigning a sequence to the MS/MS spectrum. Sequence coverage, the fraction of the peptide sequence that can be explained by product ions, is an important experimental figure of merit. Sequence coverage is important because there may be another protein that contains the peptide “specturm”. The masses of “spectrum” and “specturm” are identical, but they can be differentiated by their unique MS/MS fingerprint if a sufficient number of product ions from the right most portions of the peptides can be measured.

Getting peptide ions into the mass spectrometer

Peptides and proteins are amphoteric, they can accept protons (H^+) or donate protons to become positively or negatively charged in solution. They are ions in solution, but they are nonvolatile and do not readily transfer as ions from solution into the atmosphere. Gaseous ions are required to do mass spectrometry so one must find a way to transfer these ions from solution into the atmosphere. There are currently a number of methods that serve to ionize and transfer biomolecules from solution into the atmosphere to form gaseous ions, but the method I used in my research was electrospray ionization (ESI).[16] In this process, a liquid containing proteins and peptides are electrosprayed from the tip of a capillary (small diameter tube) in which an electrical potential ranging from a few hundred to several thousand volts has been applied. The electrical potential creates a high electric field at the capillary tip. For the production of positive ions (cations), the capillary acts as a positive electrode. Cations in the liquid, including positively charged peptides and proteins, migrate to the surface of the liquid as negative ions migrate in the opposite direction to neutralize the imposed electric field. The concentration of cations at the surface destabilizes as ions are drawn towards the counter electrode (the mass spectrometer) by the electrical potential gradient. Yet, these cations cannot escape the liquid due to surface tension. The liquid is drawn out of the capillary tip and forms a cone. This has been named the Taylor cone after the scientist that described the process of cone formation in the presence of the electric field.[17] The Taylor cone is drawn out into a filament due to the electric field. The high amount of charge destabilizes the filament and droplets erupt from the filament rich in surface charge.[18] The droplets produced from the Taylor cone proceed towards the mass spectrometer in the electric field gradient. Solvent evaporates from the droplets as they soon reach a charge density limit that results in coulombic fissions (explosions) producing even smaller droplets (**Fig. 5**). This process repeats continually until solvent-free ions are produced and enter the inlet of the mass spectrometer.

Large-scale analysis of peptides – putting it all together

Rarely did my research involve the analysis of a single peptide or protein. Most often I analyzed complex mixtures containing thousands of proteins that were enzymatically digested to produce a sample containing peptides numbering in the tens-of-thousands (or more!). To handle such a complex samples, the proteomics community has developed workflows to accommodate such samples. Part of my research was tailoring these workflows and modifying them to meet my research needs. One such workflow is shown in **Figure 6**. For this example, yeast cells were cultured and the cells were broken open (lysed) and the protein was extracted. The protein was enzymatically digested with the protease trypsin which specifically cuts proteins wherever they have an arginine or lysine amino acid in the sequence. The digested proteins, now peptides, were fractionated using a form of liquid chromatography called strong-cation exchange (SCX). The principle of

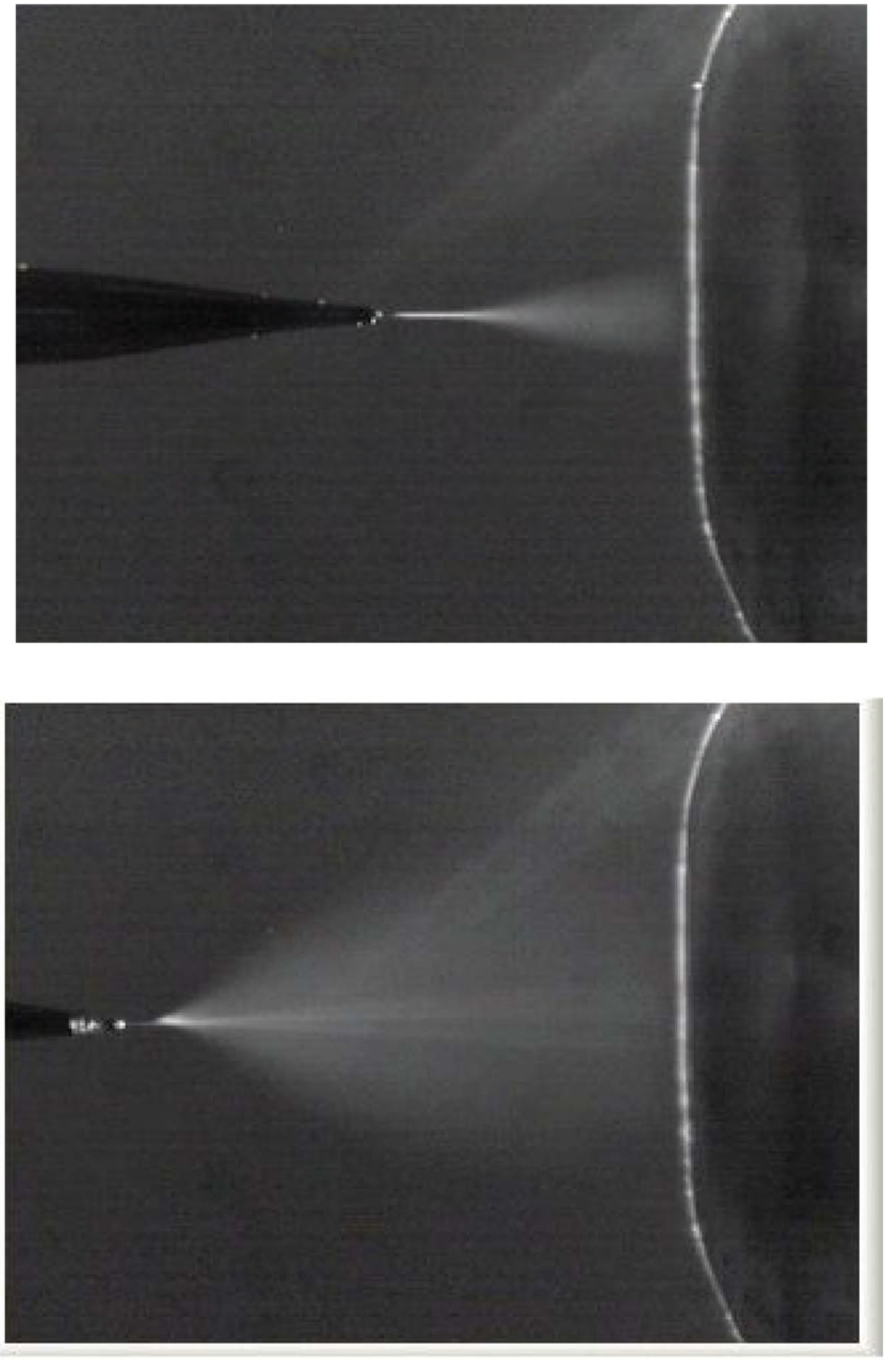


Figure 5. Electrospray ionization (ESI) from the tip of a capillary toward the inlet of a mass spectrometer. The opening of the tip is approximately 1-2 µm in diameter. Formation of a single Taylor cone (top) and multiple Taylor cones (bottom). *Images courtesy of Arne Ulbrich*.

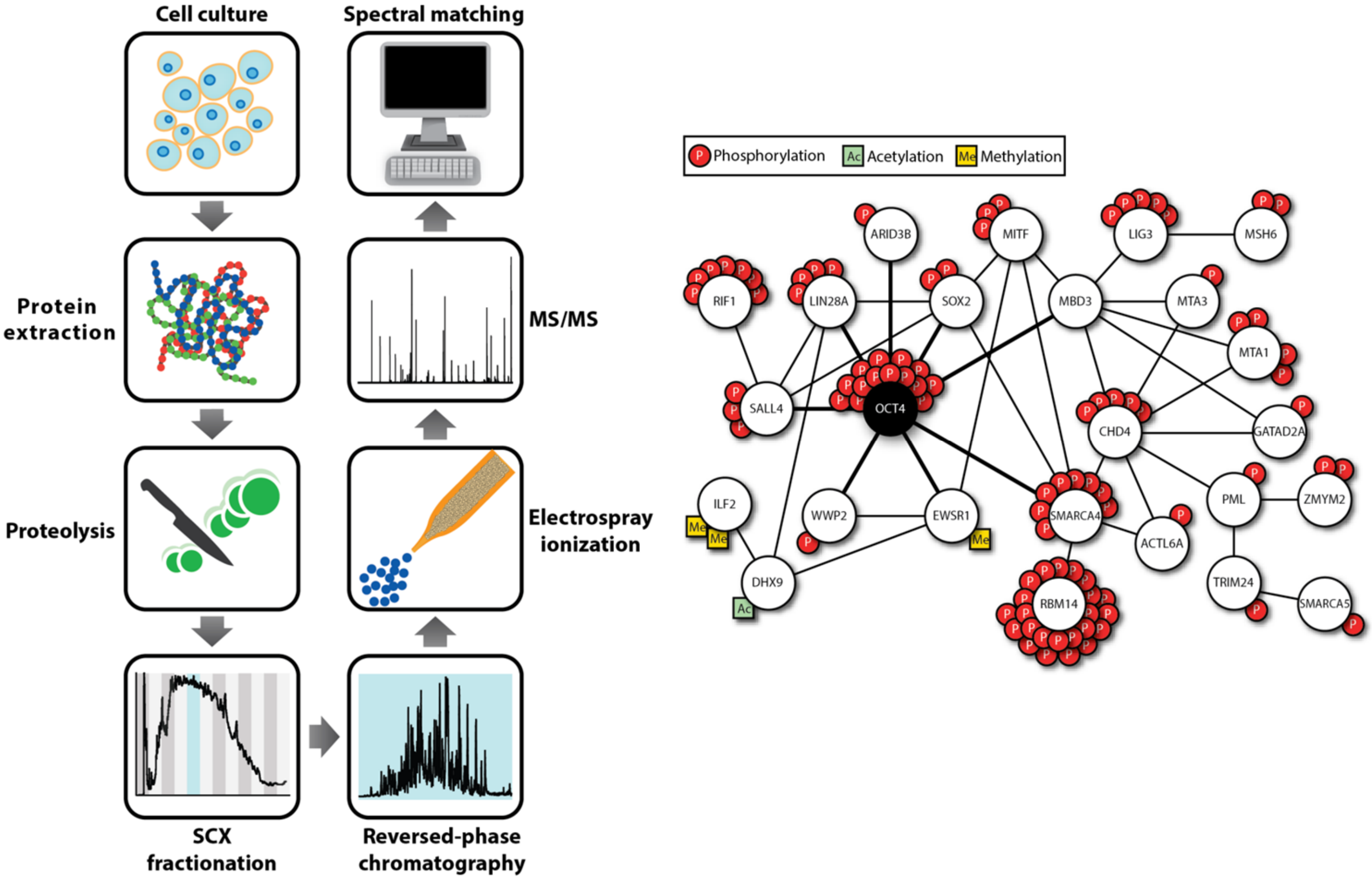


Figure 6. Workflow for a large-scale analysis of peptide mixtures (left). Protein interaction network and chemical modifications for transcription factors from human embryonic stem cells (right).[19]

this type of chromatography is that positively charged proteins stick to the negatively charged ion-exchange chromatography material and are gradually eluted from the column in order of increasing number of positive charges. This ion-exchange material is similar to that of household water softeners. Fractions of the eluting peptides are collected at intervals. Now the peptides are divided among many smaller samples. So, instead of analyzing all of the peptides at once, they can be spread across multiple samples making the overall analysis much more sensitive. The peptides were next separated using another form of liquid chromatography (reversed-phase chromatography) that is compatible with electrospray ionization. Peptides elute from this type of chromatography column based upon hydrophobicity, meaning that the peptides that are most water soluble will generally elute from the column first. Peptides that are less water soluble ("sticky" or "greasy") elute later. As the peptides were eluted from the column they were ionized by electrospray. Tandem mass spectrometry was performed on peptide precursors as they were sampled by the mass spectrometer. In this entire experiment, MS/MS was performed 65,001 times! That means 65,001 MS/MS spectra were produced. Luckily, software has been developed that automatically attempts to match an MS/MS spectrum to a peptide sequence. There were 13,555 MS/MS spectra that matched to peptides. These 13,555 peptides were mapped to 1,349 yeast proteins.

What can you do with information from these peptide experiments? In a collaboration with my more biologically inclined colleagues, we examined proteins called transcription factors in human embryonic stem cells.[19] Transcription factors are proteins that regulate gene expression. We were interested in how particular transcription factors helped to maintain cells in a stem-cell like state rather than differentiating (changing) into different cell types. We were able to map in great detail the interactions of many important transcription factors with other proteins and with each other (**Fig. 6**). We also identified chemical modifications that were important for these interactions. These experiments brought us a wealth of information about stem cell biology that would be nearly impossible to discover with any other technique available today.

References

1. Watson, J.D., Crick, F.H.C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737-738.

2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* The sequence of the human genome. *Science* **2001**, *291*, 1304-1351.

3. International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860-921.

4. Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S. A vision for the future of genomics research. *Nature* **2003**, *422*, 835-847.

5. International Human Genome Sequencing Consortium Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431*, 931-945.

6. Pertea, M., Salzberg, S. Between a chicken and a grape: estimating the number of human genes. *Genome Biology* **2010**, *11*, 206.

7. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. *et al.* Life with 6000 Genes. *Science* **1996**, *274*, 546-567.

8. Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A.A., Cande, W.Z., Chen, F., Cipriano, M.J. *et al.* Genomic minimalism in the early diverging intestinal parasite giardia lamblia. *Science* **2007**, *317*, 1921-1926.

9. The C. elegans Sequencing Consortium Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science* **1998**, *282*, 2012-2018.

10. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, *438*, 803-819.

11. Crick, F. Central dogma of molecular biology. *Nature* **1970**, *227*, 561-563.

12. Human Genome Program, U.S.D.o.E., "From the cell to protein machines", http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer2001/1.shtml, https://public.ornl.gov/site/gallery/detail.cfm?id=403&topic=&citation=&general=from%20genes%20to%20proteins&restsection=all#credits, accessed June 6th, 2012.

13. Orchard, S., Hermjakob, H., Apweiler, R. Annotating the human proteome. *Mol. Cell. Proteomics* **2005**, *4*, 435-440.

14. Tipton, J.D., Tran, J.C., Catherman, A.D., Ahlf, D.R., Durbin, K.R., Kelleher, N.L. Analysis of intact protein isoforms by mass spectrometry. *J. Biol. Chem.* **2011**, *286*, 25451-25458.

15. Stuart, M.J., Nagel, R.L. Sickle-cell disease. *The Lancet 364*, 1343-1360.

16. Fenn, J., Mann, M., Meng, C., Wong, S., Whitehouse, C. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246*, 64-71.

17. Kebarle, P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J. Mass Spectrom.* **2000**, *35*, 804-817.

18. Cech, N.B., Enke, C.G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev.* **2001**, *20*, 362-387.

19. Brumbaugh, J., Hou, Z., Russell, J.D., Howden, S.E., Yu, P., Ledvina, A.R., Coon, J.J., Thomson, J.A. Phosphorylation regulates human OCT4. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *in press*.