Communicating Research to the General Public

At the March 5, 2010 UW-Madison Chemistry Department Colloquium, Prof. Bassam Z. Shakhashiri, the director of the Wisconsin Initiative for Science Literacy (WISL), encouraged all UW-Madison chemistry Ph.D. 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, program officers at appropriate funding agencies, state legislators, and members of the U.S. Congress.

Over 20 Ph.D. degree recipients have successfully completed their theses and included such a chapter.

WISL encourages the inclusion of such chapters in all Ph.D. theses everywhere through the cooperation of Ph.D. candidates and their mentors. WISL is now offering additional awards of $250 for UW-Madison chemistry Ph.D. candidates.

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

May 2014

METHOD OPTIMIZATION AND APPLICATION OF MASS SPECTROMETRY TO THE FIELD OF HEMATOLOGY

by

Emily M. Wilkerson

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 9/5/17

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Chemistry
David J. Pagliarini, Associate Professor, Biochemistry
Lloyd M. Smith, Professor, Chemistry
Ying Ge, Associate Professor, Chemistry

Chapter 5

A PROTEOMICS PRIMER FOR NON-SPECIALISTS: FOR THE WISCONSIN INITIATIVE FOR SCIENCE LITERACY

EMW wrote the chapter.

Introduction to Proteomics

Genes are the instructions for life and you may have been told that a certain trait was passed down from your parents. Your DNA carries the information for traits such as hair color, whether or not you have a widow's peak, your blood type, and many other aspects. The central dogma of biology (**Fig. 5.1**) outlines the flow of genetic information from DNA to RNA to proteins. RNA is the intermediate that acts a messenger and later a template for the production of proteins, which are long chains of amino acids.

The large scale study of proteins is called proteomics. In proteomics, scientists study the presence and/or relative quantities of proteins within a biological sample. A disease state may be caused by the absence, or conversely the over-expression of a certain protein. Therefore proteomics plays a critical role in current biological, biochemical, and medical investigations.

Mass Spectrometry

Mass spectrometry is a key analytical technique for probing differential protein expression, protein abundance, and protein modifications. In normal conditions, a protein may be at a certain level X, whereas under a different condition the same protein could be at 2X level. This is over-expression of the protein of interest.

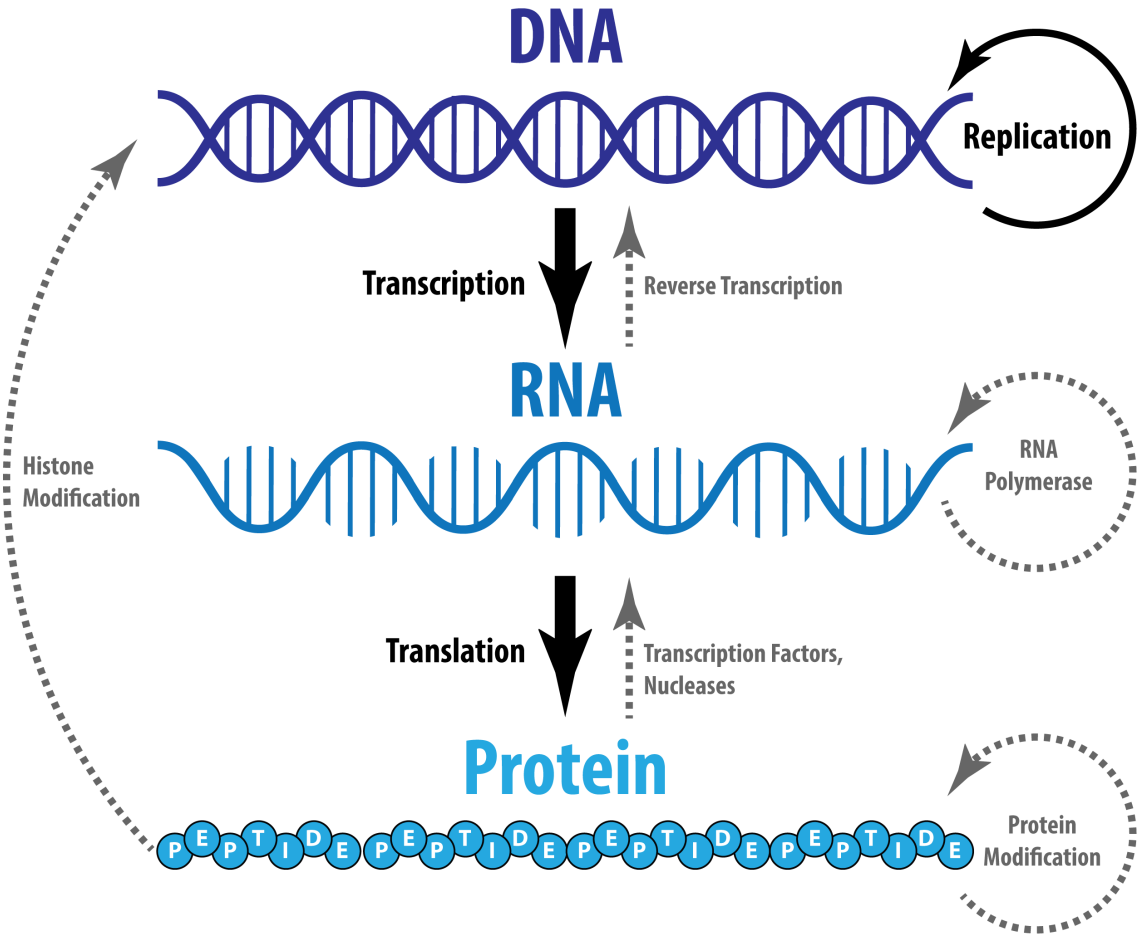

Figure 5.1: Central Dogma of Biology. Flow of genetic information from DNA to RNA to protein. There are two phases, transcription and translation. During transcription DNA is converted into RNA, which is the messenger. RNA then is used as a template for protein formation. Proteins can be modified by other molecules or other proteins. These modifications are called post-translational modifications because they occur after the protein is formed.

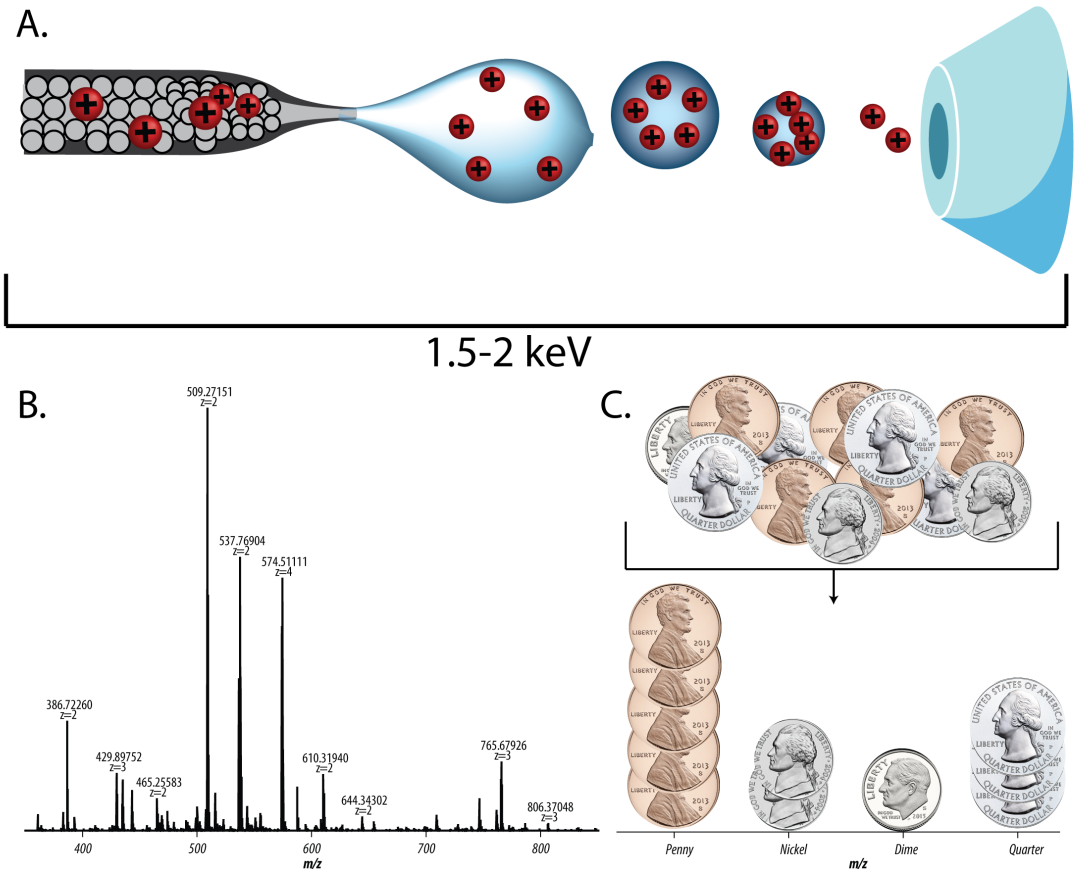

Figure 5.2: Overview of Mass Spectrometry. **(a)** Overview of ionization where liquid elutes from tip of a column or needle. Then the liquid turns into gas by the time it enters the mass spectrometer. The analytes, molecules of interest, are now charged and in the gas phase, so they are gas ions. **(b)** Example of a mass spectrum readout. The horizontal axis is the mass-to-charge ratio and then the height is dependent on how much of that analyte is present, which I will refer to as intensity **(c)** Coin analogy to describe mass spectrometry-based proteomics where the mass spectrometer is the coin sorting machine that enables a mixture of pocket change to be characterized in terms of the amount of each type of coins, which are analogous to peptides or proteins

Proteins are not all at the same level in our bodies, which is why we use proteomics to measure protein abundance. The modification of proteins with other molecules can act as a signal for the protein to regulate a certain activity in the body.

The two basic parts of the mass spectrometer are the ionization source and the mass analyzer. The mass spectrometer detects ions (positively or negatively electrically charged molecules) in the gas phase in terms of their mass-to-charge ratio (*m/z*). Electrospray ionization, one specific type of ionization, applies high voltage to liquid (**Fig. 5.2A**). This creates an aerosol where the liquid evaporates by the time the ions enter the mass spectrometry.

To help understand how the mass spectrometer operates we will compare the mass spectrometer to a coin sorting machine. You start off with a mix of coins from a bag, piggy bank, or other container. However, to be measured by the machine they need to be in a different form. So the ionization process of peptides is similar to the process of taking the coins out of the wrappers, plastic bags, or coin banks where they were housed. The goal is to have the coins, or peptides/proteins for proteomics, in a format that the machine can analyze.

Once the ions are in the mass spectrometer, it then can determine their composition. The coin sorting machine is able to determine how many pennies, nickels, and dimes are present and detect the differences between the different coins. For the

mass spectrometer, the mass analyzer determines the mass-to-charge ratio of each different component and the relative abundance of each component (**Fig. 5.2**B). Now comparing the mass spectrum in **Fig. 5.2**B to the histogram of coins in **Fig. 5.2C** we can see some similarities. The height of the peaks in the mass spectrum are characteristic of the intensity or how much of that given peptide is present.

Mass Spectrometry-based Proteomics

Mass spectrometry-based proteomics is split into two different approaches to study the proteome - top down and bottom up. In top down, intact proteins are ionized into the mass spectrometer, which allows researchers to see their full mass. In bottom-up proteomics the proteins are digested into smaller peptides and then we rely on data analysis to piece together the proteins from their observed peptide components. For the remaining explanation I will focus predominantly on bottom-up proteomics.

As mentioned before, proteins are long chains of amino acids. There are 20 different amino acids that can be combined in different sequences of different lengths. The basic structure of each amino acid is the same, with one specific region varying between the different amino acids, which is called the R group. Each amino acid has a characteristic mass due to the differences in this R group region. These

amino acid masses are used to determine the sequence of proteins or peptides observed by the mass spectrometer.

In the coin analogy the machine is able to distinguish between a nickel and a penny. However if we wanted more information about what makes a nickel and a penny different, we would have to do further analysis. Likewise in the initial analysis in the mass spectrometer we can determine the mass-to-charge of the peptide, but we don't know the sequence of the peptide.

In mass spectrometry-based proteomics we fragment, or break apart, the peptides to determine the sequence of the peptide. A peptide of interest can be isolated, then energy applied, to make it break apart into pieces. For example the peptide with the sequence PEPTIDE would be broken into P, PE, PEP, PEPT, PEPTI, and PEPTID. We could also generate E, DE, IDE, TIDE, PTIDE, EPTIDE by breaking off an amino acid from the front end of the peptide. All of the peptide fragments will also have a mass-to-charge that is measured and recorded during a subsequent scan producing a tandem mass spectrum. The difference between the mass-to-charge of PE and PEP would be the mass of the amino acid P.

The process of sequencing peptides, called de novo peptide sequencing, is time consuming due to large number of spectra collected by the mass spectrometer per experiment. Therefore search algorithms and software programs are used by

researchers to sift through the large amount of raw data that a mass spectrometer produces. For example, a single 90-minute analysis of one sample can easily collect more than 150,000 spectra containing peptide sequencing information. The data analysis tools not only help tackle the large amount of data, but are also used by bioinformaticians to develop scoring algorithms to limit the reporting of false positive peptides.

Application of Mass Spectrometry-based Proteomics

Building on the basics of mass spectrometry-based proteomics, I will now describe how we use the information about peptides, peptide sequence, and abundance to tackle biological questions. Sequencing of the human genome was a large collaboration published in 2001 and completed in 2004 that identified over 20,000 protein-coding genes. Following the flow of genetic information, it might be logical to sequence the entire human proteome. However, that task has proven to be very complex - two recent draft maps of the human proteome were published in 2014. The capabilities of proteomics are constantly improving in terms of speed, sensitivity, and throughput. The abilities to answer biologically relevant questions extend well beyond a complete human proteome. Proteomics can be used to determine protein function, learn more about the effect of a drug, or better understand a

specific biological pathway. Two examples of applications of proteomics follow:

Protein-Protein Interactions Each of our cells is full of proteins that work together to fulfill certain functions such as helping us transport nutrients and fight infections. Often times proteins work together to carry out a specific function, forming networks of protein-protein interactions. It is possible to determine the function of a protein of unknown function by determining what proteins it interacts with. Mass spectrometry is a useful technique for probing these interactions in conjunction with a biochemical technique called affinity enrichment. During an affinity enrichment we target a protein of interest. In selecting this protein of interest we also capture any proteins that interact with our target. Mass spectrometry can then be used to identify those proteins interacting with their target.

I used this technique in collaboration with the Pagliarini group at University of Wisconsin - Madison to probe protein-protein interactions in a specific cellular component called the mitochondria. There are at least 228 proteins of unknown function in the mitochondria and at least 25 that are associated with human disease. Our study focused on 50 of the proteins of unknown function. We used affinity enrichment mass spectrometry to identify high confidence protein-protein interactions. High confidence interactions were determined using a statistical algorithm to separate out actual interactions from non-specific background. Biochemical

follow-up provided further understanding of the protein function. We identified the protein C17orf89 as a key player in an important pathway, complex I. For the majority of patients with complex I deficiency, doctors and scientists do not fully understand the molecular cause. Thus placing additional proteins into the pathway functions helps us understand missing components of biological pathways.

Quantitative Proteomics Protein abundance is important not only in the context of a healthy, undisturbed cell, but also in understanding what happens when that cell is perturbed. Quantitative proteomic techniques allow us to measure changes between different experimental conditions and determine whether proteins are up- or down- regulated. The normal abundance of a protein X might be 100, however if it is down-regulated it would be less than 100 and up-regulated would be greater than 100. For example when you contract a cold, your immune system begins to fight the infection. This means that some proteins may need to be produced at a much higher rate than they would be if you were healthy to help fight the infection.

During my graduate research I used this technique to better understand how eosinophils, a type of white blood cell, change upon activation by a protein interleukin-5. Interleukin-5 acts as a trigger for the eosinophils to undergo significant change in preparation for migration to a site of inflammation. Activated eosinophils are associated with asthma and allergic responses. Previous research had shown that

the activation of eosinophils causes a global shape change of the eosinophil. We used quantitative proteomic techniques to determine which proteins were up- and down- regulated. We measured the cells after only 5 minutes of activation, so there were minimal changes in protein expression. Given the short time span we were particularly interested in what proteins were becoming phosphorylated (a type of post-translational modification). We found 220 sites of phosphorylation that were significantly changing (based on statistics). We found patterns in the phosphorylation results that pointed towards specific protein pathways being impacted by the activation. Our collaborators are now looking further into several of the up-regulated proteins to better understand their biological significance.

Outcomes of Proteomics

Mass spectrometry based proteomics can be applied to better understand human samples of biomedical relevance. Ideally proteomics will answer biomedical problems and will provide meaningful results that can be used to improve treatment, clinical recommendations, or basic scientific understanding of a sample type. As mass spectrometry technologies continue to improve, the clinical relevance of proteomics increases as analyzing a large number of samples becomes not only more feasible, but also more efficient, comprehensive, and accessible.