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 50 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

Integrated Proteomic Strategies for Proteoform Discovery

by

Leah V. Schaffer

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2020

Date of final oral examination: 06/12/2020

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

Brian G. Fox, Professor, Biochemistry
Ying Ge, Professor, Cell and Regenerative Biology and Chemistry
Lingjun Li, Professor, Pharmaceutical Sciences and Chemistry
Lloyd M. Smith, Professor, Chemistry

7 CHAPTER FOR THE NONSCIENTIFIC PUBLIC

This chapter was written in collaboration with the Wisconsin Initiative for Science Literacy.

7.1 Introduction

The purpose of this chapter is to explain the research in my dissertation to a broader, non-scientific audience. I am currently writing this chapter from home due to the ongoing COVID-19 pandemic. Through the last few months, we have watched the importance of science currently play out in real time, as scientists work tirelessly to understand this novel virus and develop vaccines or possible treatments. One thing that drives me personally as a scientist is knowing that increasing humanity's collective knowledge about the world around us can have real, tangible benefits for humans and truly improve or even save people's lives. With each passing year, I believe more strongly in the importance of listening to scientific experts when making decisions and policies. It is therefore important to me as a scientist to fulfill the responsibility of communicating my research to others and to convey why it is important. I would like to thank the Wisconsin Initiative for Science Literacy at the University of Wisconsin-Madison for providing this important platform and for their sponsorship and support throughout the writing of this chapter.

7.2 What are proteoforms?

In all the cells in our body, proteins are the main molecules that carry out biological functions. Proteins help digest the food we eat, are responsible for the shape of our cells, and act in our immune system to prevent us from getting sick. Humans have about 20,000 genes that contain the instructions to create about 20,000 different proteins. For example, two well-known proteins are hemoglobin and insulin. Hemoglobin proteins transport oxygen in our blood, whereas insulin proteins signal to different cells in our body to take in sugar from our blood. These two different proteins are derived from two different genes

Humans have a surprisingly low number of genes compared to other organisms. We have around the same number of protein-encoding genes as a chicken or a dog, and we have fewer than a rice plant. Our complexity as humans is therefore not solely due to our number of genes, but rather is possible due to diversity at the protein-level. Just as a chain is made of different types of links attached together, proteins (chains) are made of different types of building blocks called amino acid residues (links), as shown in **Figure 7.1**. There are about twenty different types of amino acids that make up proteins. In this way, proteins are like chains made of different combinations of twenty different types of links. Some links are heavier than others and some might have certain properties, like being more water resistant. The overall properties of the chain depend on the length of the chain and the order and number of the different types of links.

Proteins derived from the same gene can vary from one another. Proteins can have a change in their sequence (such as substituting one type of amino acid for another), can be clipped short, and can be chemically modified. There are many different types of chemical modifications, each of which are added to certain amino acids based on their structure. These modifications can act as a signal for other proteins to interact and carry out a function together. When a modification is added, it can also cause a protein to change its shape, turning it "on" or "off" to perform a function in the cell.

We can imagine these different sources of protein variation using the chain of links analogy. A chain might have a small bronze link substituted with a medium silver link. The chain might have a section in the middle cut out and the two new ends joined together. The chain might be clipped at different links to form several different smaller chains. An extra, brightly-colored link (modification) might be added onto one or more of the main chain links.

Each of these biological processes results in a different proteoform, which is a

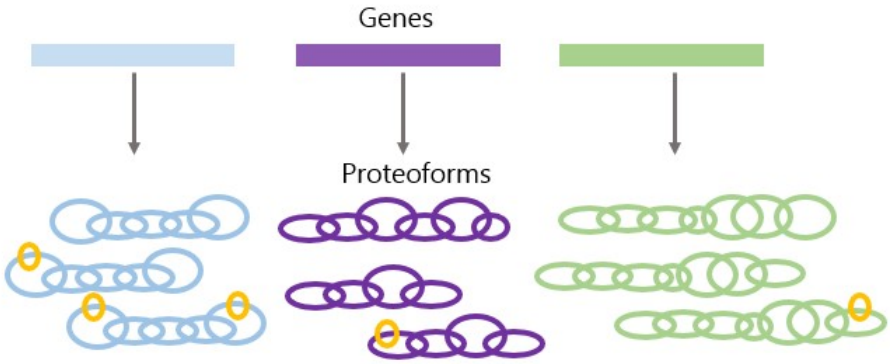

Figure 7.1. Each gene in the human genome can be the source of different proteoforms (the forms of a protein). Similar to a chain with different links, each proteoform has a unique sequence (the main link of chains) and profile of modifications (yellow links on the main chain).

specific form of a protein that has a distinct sequence and profile of modifications. Different proteoforms from the same gene can have completely different functions in the cell. An unmodified proteoform might activate a process in the cell; if a chemical modification is added to the proteoform by another proteoform the new, modified proteoform might then begin to repress the same process. Some studies have shown that different proteoforms from the same gene can have biological functions as different as those from different genes. Therefore, it is important to identify and quantify proteoforms in a biological system of interest in order to understand the system. Scientists are currently analyzing how proteoforms change in diseases such as cancer, diabetes, and heart disease in order to understand the causes and effects of these diseases.

7.3 Analysis of proteoforms using mass spectrometry

Proteoforms are typically analyzed using an instrument called a mass spectrometer. In mass spectrometric analyses, molecules become ionized (charged), and the mass

spectrometer measures the mass-to-charge (m/z) ratios of the ions. The result of data processing is a mass spectrum (MS), which shows the intensity of each m/z value. Data analysis programs can use these m/z values to determine the masses of the original molecules, which is similar to how much the molecules weigh. The ability to determine the mass of a molecule is extremely powerful when trying to identify it.

In a typical mass spectrometry experiment to identify proteoforms (**Figure 7.2**), we load a sample of proteoforms onto a column that separates the proteoforms over the course of ~1-2 hours. At the other end of the column, the proteoforms are ionized and introduced into the mass spectrometer. The instrument takes a mass spectrum of the intact proteoforms that just exited the column and were ionized (MS1). Then, the most abundant m/z peak in the MS1 is selected and fragmented in the instrument, and a second mass spectrum is acquired of the resulting fragments (MS2). This process is repeated ~3 times, another MS1 is taken, and 3 new peaks from the new MS1 are selected to be fragmented. The cycle repeats over the course of the entire 1-2 hour run as different proteoforms exit the column and enter the mass spectrometer. Data analysis programs use the mass of a proteoform from an MS1 spectrum and the masses of its fragments from an MS2 spectrum to determine a proteoform's identity.

For example, in the chain of links analogy described above, say there are three types of chains known to possibly exist in the world, described in a special database (**Figure 7.3**). There is a chain made of two big links and one small link, a chain made of three small links, and a chain made of four medium links. We then find an unknown chain that we want to identify, meaning we want to know which of these three possible chains it could be. We weigh the intact chain (MS1) and immediately determine that it cannot be the chain consisting of three small links because this unknown chain weighs too much to be that one. However, we still don't know if this unknown chain is the one with two big links and one small link or the other one with

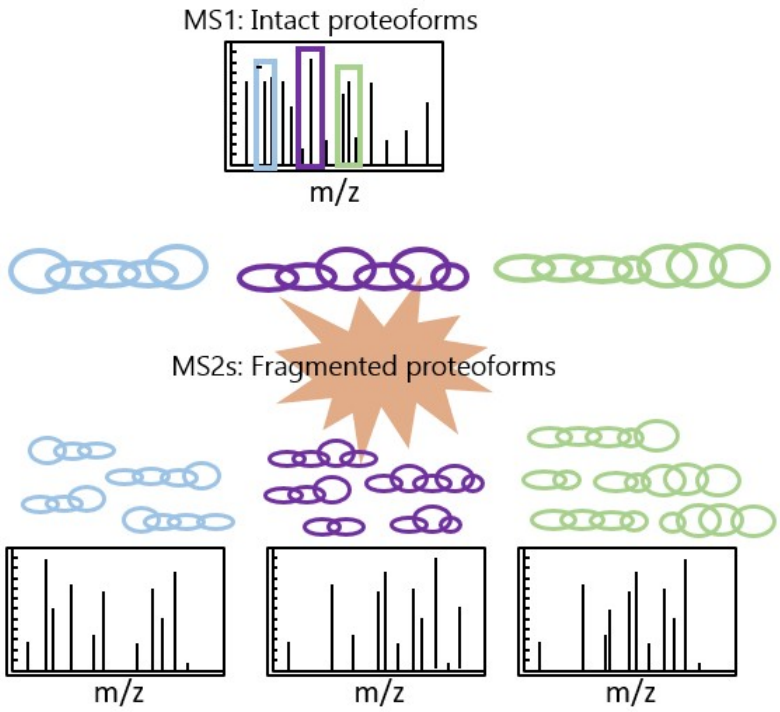

Figure 7.2. Mass spectrometry analyzes proteoforms by taking a mass spectrum (intensity of different mass-to-charge ratios) of intact proteoforms, and the most abundant proteoform peaks are then selected to be fragmented. Data analysis software identifies proteoforms based on their intact mass and the fragmentation masses.

four medium links; these two possible chains weigh almost the same while intact. If we break the unknown chain between the links and weigh the fragmented chain links (MS2), we can figure out which of the two chains it is. If we determine that one of the fragments weighs the same as a small link plus a big link and another fragment weighs the same as a big link, then we can determine that our unknown chain is the one known to have two big links and one small link. In this example, we successfully identified our chain with MS2 fragmentation data.

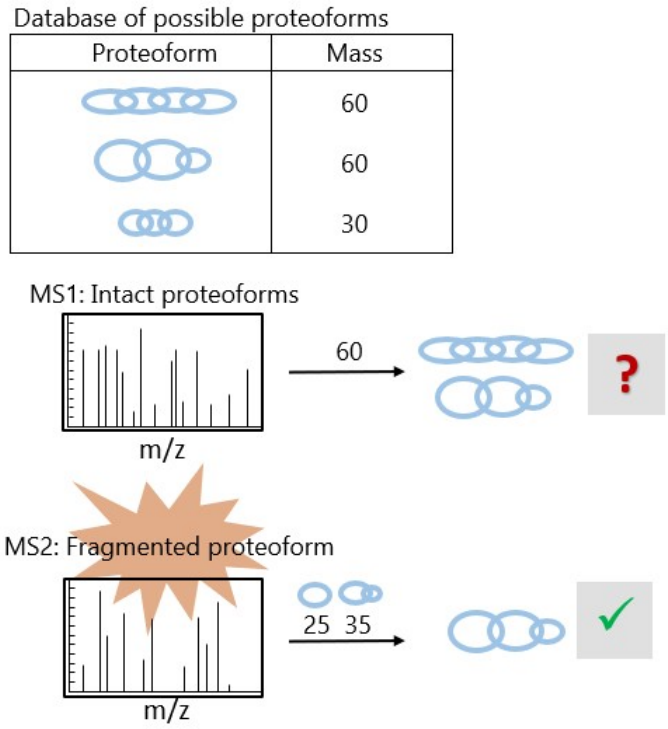

Figure 7.3. We typically identify a proteoform by analyzing its intact mass and fragment masses.

7.4 Increasing the number of proteoform identifications

It takes a lot of instrument time to obtain high-quality mass spectra of a proteoform and its fragments. As a result, many proteoforms finish leaving the column before the instrument has time to fragment them. Additionally, some proteoforms do not fragment as easily as others. The different software programs that we use in these proteoform analyses require fragmentation data to identify proteoforms, so we are unable to identify any proteoforms that weren't fragmented or with low-quality fragmentation data with the previously available software tools. As described above, identifying proteoforms is vital to understanding biological systems; the field needed a data analysis tool that could analyze all of these observed, unidentified proteoform masses and try to identify them. That is where my research enters the scene.

Our laboratory developed the software program Proteoform Suite to analyze the intact masses of proteoforms observed in the MS1 spectra. First, we create a list of proteoforms that were successfully identified with fragmentation data. Then, we generate a list of proteoform masses observed in the MS1 spectra, many of which were not identified by fragmentation. We input both of these lists and a database with known proteoform sequences and modifications into Proteoform Suite. Proteoform Suite first compares the experimental masses to the masses in the database to determine "perfect matches," meaning an experimental mass that is very close to the mass of a proteoform in the database. To identify additional proteoforms, Proteoform Suite then compares experimental proteoform masses to each other. It groups together proteoforms with mass differences that correspond to a known modification or amino acid difference and makes additional intact-mass proteoform identifications based on these mass differences.

Back to the chain-link analogy: let's say we identify one chain by fragmenting it and we determine it to be a chain consisting of four medium links; we might observe that another chain weighs the same as this chain minus the weight of one of the medium links; we could conclude that this new chain is probably three medium links, without even having to fragment it. We call this approach intact-mass analysis (**Figure 7.4**). For some genes, no proteoforms are identified by fragmentation, requiring us to rely on intact-mass alone to try and identify them.

We were able to increase the number of proteoform identifications by 40% in an analysis of yeast proteoforms using this intact-mass analysis approach. We also used Proteoform Suite to increase the number of identifications and quantify proteoform abundance changes in mitochondria from mouse cells. Mitochondria are subunits in the cell that generate most of the cell's energy. We found mitochondrial proteoforms are more abundant in muscle fibers than in immature muscle cells, which makes

sense because muscle fibers rely more on mitochondria for energy. In another study, we used Proteoform Suite to help identify proteoforms from human heart tissue. We especially focused on identifying larger proteoforms because they are even more difficult to identify with fragmentation.

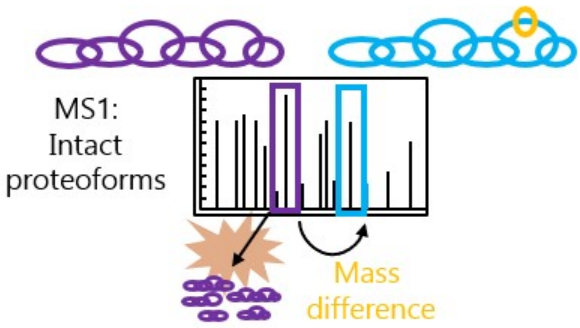

Figure 7.4. The intact-mass analysis we developed identifies additional proteoforms based on the intact mass of observed proteoforms. In this example, one proteoform was identified by fragmentation (purple) and intact-mass analysis enabled an additional observed proteoform to be identified (blue) based on its mass difference from the identified proteoform. The mass difference corresponds to a chemical modification (yellow) added to the proteoform sequence.

7.5 Conclusions and Future Directions

The research described in this dissertation has involved developing strategies to increase the number of proteoform identifications towards the long-term goal of comprehensive proteoform analysis that would enable better understanding of biological systems and diseases. I mainly focused on the development of freely available and open source software that can identify proteoforms that were observed but unidentified in a typical proteoform analysis using mass spectrometry. I expanded this analysis from yeast to more complex biological systems, including mice, a human cancer cell line, and large proteoforms from human heart tissue samples.

Looking forward, I'm interested in harnessing the power of intact-mass analysis to drive follow-up targeted fragmentation experiments. In a targeted analysis, the mass spectrometer is given a list of masses to "target," and if a mass on this list is observed in an MS1 spectrum, it will be selected for fragmentation even if it is not the most abundant peak in the spectrum. Without a target list, the mass spectrometer only has time to fragment the most abundant proteoforms observed in the MS1 spectra. We can give the instrument a target list of masses that we are interested in to gather fragmentation spectra of proteoforms that are less abundant and therefore not normally selected for fragmentation.

As described above, sometimes different proteoforms have different sequences and profiles of modifications but weigh the same when they're intact (such as the chain with two big links and one small one and the chain with four medium links that weighed the same while intact, **Figure 7.3**). These proteoforms can't be identified by their intact mass alone; they require the fragmentation data to help differentiate them. However, if they are not very abundant, they might not be selected to be fragmented by the instrument. We could do a targeted experiment and give the mass spectrometer a list containing the intact mass of these proteoforms. This list would guarantee that if these proteoforms were observed in the MS1 spectra, the instrument would fragment them, enabling their identification with the collected fragmentation data.

In Proteoform Suite, we quantify all proteoform observations, even those that aren't identified, using the intensities of the m/z peaks in the mass spectra. If a proteoform was determined to be changing in abundance across different biological conditions (such as cancer vs. normal tissue), it could also be marked as a target and we could put its mass on a target list in future experiments to try to identify it using fragmentation.

I hope I have conveyed to the public what proteoforms are and why they are

important, how scientists currently typically identify proteoforms using a technique called mass spectrometry, and how my research has involved developing software strategies to increase the number of proteoform identifications. I had never written a computer program before graduate school, but I quickly realized that all of the data being generated from mass spectrometry analyses presents an exciting opportunity if you're able to write programs to analyze it. Writing software programs to analyze mass spectrometry data enables exciting biological discoveries! As scientists are able to identify an increasing number of proteoforms with fragmentation through improvements to both instrument technology and data analysis software programs, I hope that the intact-mass analysis strategy described here will prove to be even more useful to identify currently unidentified observed masses as well as save precious instrument time to target interesting proteoforms that are yet to be discovered.