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

Defining the Human Heart Proteoform Landscape with Top-down Proteomics

By

Trisha Tucholski

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: June 10th, 2020

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

Ying Ge, PhD, Professor, Cell & Regenerative Biology and Chemistry

Lloyd Smith, PhD, Professor, Chemistry

Lingjun Li, PhD, Professor, Pharmaceutical Sciences and Chemistry

Helen Blackwell, PhD, Professor, Chemistry

CHAPTER 2:

Tipping the Scale Toward Measurement of Larger Proteoforms in the Human Heart with Mass Spectrometry*

**The creation of this chapter was sponsored and supported by the Wisconsin Initiative for Science Literacy Graduate Thesis Award for Communicating Research to the Public*

This chapter is publicly available at: http://www.scifun.org/Thesis_Awards/thesis_awards.html

2.1 Communicating Science with the Public

Science communication is a fundamental component of our duty as scientists. Since scientific research is often funded by the public and performed for their benefit, it is essential that we share our findings with the non-scientists in our community. Now, more than ever, with a public health crisis underway and scientific discovery unfolding before us, it is critical that we communicate the importance of scientific research with a broad, non-scientific audience. I have written this chapter in celebration of science, my own research, and its impact on public. I would like to thank to the Wisconsin Initiative for Science Literacy at UW-Madison for providing this platform, and for sponsoring and supporting the creation of this chapter.

2.2 Proteins, From Table to Cells

The organs and tissues in our bodies are made up of cells, which each contain functional and structural biomolecules called *proteins*. Proteins are the molecules in our cells that carry out important biological processes and allow us to live. When one hears "protein," they usually think of *dietary* protein- coming from steak, tofu, or egg whites- that we consume to meet our nutritional needs. Aside from eating steak and eggs for taste, we eat protein because we need to provide our bodies with the building blocks to make new proteins that our cells can use. Our body breaks down, or digests, the protein molecules we eat into their original building blocks, called *amino acids*, and sends these to be assembled into new protein molecules. In this way, our body can "recycle" the amino acids from the food we ingest to rebuild new proteins (**Figure 2.1**). Proteins are simply chains of amino acids strung together. However, unlike a string of pearls, amino acids are not connected by a string, but by chemical connections called *covalent bonds*. There are 20 different amino acids that our cells know how to use, and we call these the natural amino acids. Each amino acid has its own name, letter code, and properties. Amino acids in a protein "backbone" interact

with each other, and this allows the protein to fold into 3-dimensional shapes which allow them to perform a variety of structural and functional roles in cells.

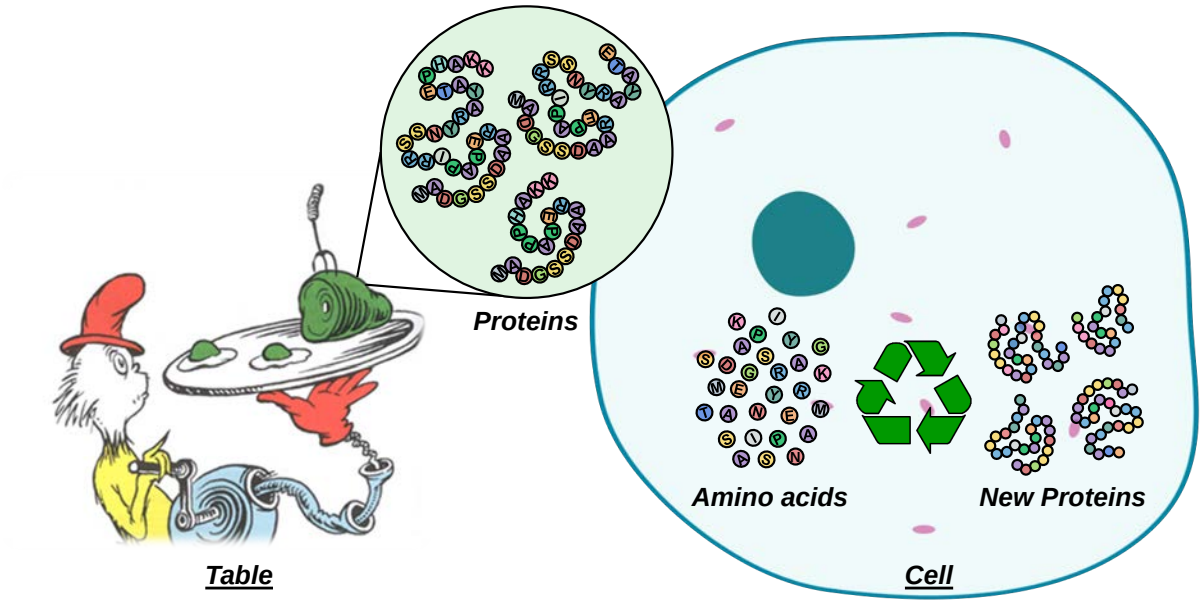

Figure 2.1. Ingested protein molecules are recycled to build new proteins in cells.

As you can imagine, the possibilities for different combinations of the 20 natural amino acids are endless. However, amino acids are not randomly strung together. Our cells have a plan and know specifically which amino acids to assemble and in what order. There are sets of instructions to assemble the proteins we need, and these are called *genes*. A gene is a set of instructions for assembling a particular protein. Genes are passed down to you from your parents through chromosomes, which house tightly wound spools of another type of biomolecule, DNA. From one single gene, many different forms of a protein can be produced. The cells in our body can carry out an amazing number of biological processes *because* more than one form of a protein can be produced from each set of instructions we have. Scientists have determined that we have about 20,300 protein-coding genes. Given the complexity of humans, this number was surprising

to scientists who had previously estimated that humans had 100,000 genes. Today, we understand that our biological complexity can be explained by the many *proteoforms* -or protein forms - that are produced from a single gene.[1] We will revisit the concept of proteoforms later.

2.3 Well-rounded Proteins and Their Role in a Cellular Society

For now, let's circle back to the concept of proteins and the roles they play in our cells. Consider our society, and all the people performing the different jobs that help it operate. Like people in our society, proteins perform diverse jobs in each of our cells. Like individual people, proteins assembled from different sets of instructions, or as discussed in the section above, even from the same set of instructions, perform different functional roles. We have essential workers in our community, such as health care professionals and mail delivery persons. Essential jobs performed by these individuals can be compared to the essential jobs which are carried out by proteins assembled in our cells. These proteins have jobs like repairing broken DNA molecules; delivering messages to the "headquarters," or *nucleus*, of the cell; and generating energy that is needed for the cell to operate. All cells in our body contain these types of proteins, commonly considered "housekeeping proteins," which carry out essential operations. In our society, there are also individuals who carry out more specialized roles, like wedding photographers or orthodontists. In our cells, there are also proteins that have specialized roles. There are highly specialized proteins which facilitate the contraction and relaxation of our heart cells, allowing the whole heart muscle to beat. All the tissues and organs which make up *us* contain specialized cells, and some are so specialized that they are named for the organ in which they exist. For example, "cardiomyocyte" translates exactly to "heart cell." Our brain, liver, and heart all consist of specialized cells because these organs do specific things, and in turn, consist of specialized proteins that carry out specific functions.

However, also consider that the mail delivery person who brings your Amazon packages plays other roles in their community when they are off the job. They may coach a little league baseball team or sing in a local church choir. Similarly, *you* may list different skills and attributes on your resume. Proteins, like individual people, are well-rounded and can have more than one function in a cell, even if they were produced from the same set of instructions.

A few more important notes about proteins: Like people, proteins are collaborative and work together to carry out jobs. Just like you might recruit help from your roommate to carry a heavy box to the second floor of your apartment, two proteins might collaborate to send a signal to the nucleus of a cell. Similarly, it often takes an entire team of people to execute jobs, like the team of nurses and doctors in a hospital operating room. It takes a team of specialized proteins- experts in their field, if you will- to make your heart muscle beat (**Figure 2.2**). These proteins make up a network called the *sarcomere* in heart cells; they communicate through messages and signals to control the contraction and relaxation of your heart on the molecular level. If one expert involved in an operation is missing, the whole operation can go awry. This concept applies for proteins in our cells, too. It is important for us to know when a protein was mis-assembled or mis-folded, and it is also important for us to know when too many or too few of one type of protein was present.

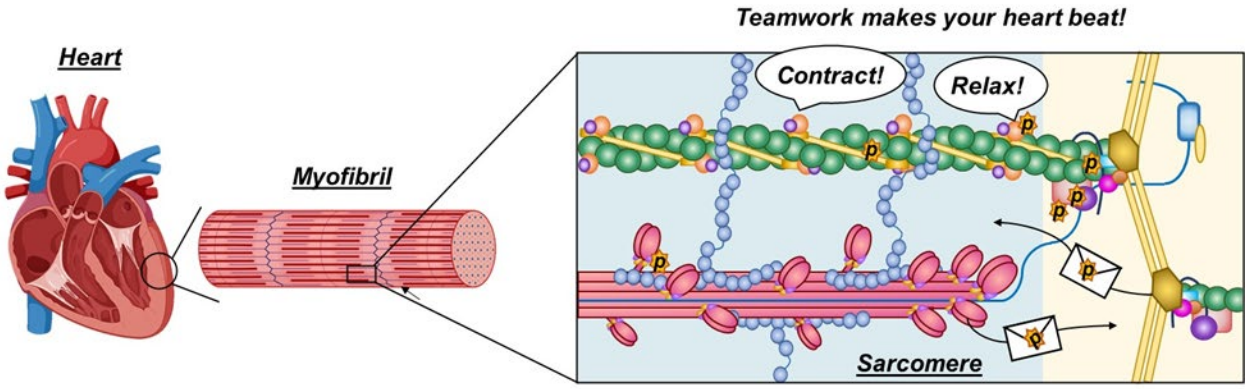

Figure 2.2. Teamwork makes your heart beat: Heart tissue is made up of heart cells which contain myofibrils. Myofibrils contain repeating sarcomere units. The sarcomere is a team of proteins which interact and send messages to control the contraction and relaxation of your heart on the molecular level.

2.4 A Plethora of Proteoforms From a Single Set of Instructions

As discussed earlier, the different forms of a protein produced from a single gene, or set of instructions, are called *proteoforms*. There are a variety of ways proteoforms can be produced. For instance, copies of the original instructions, called *transcripts*, are sometimes altered before proteins are assembled. As a result, only a portion of the complete set of instructions is used, resulting in shorter or slightly rearranged versions of the amino acid chain that was described in the original plan. This process is called *alternative splicing*, and gives rise to the formation of different proteoforms (**Figure 2.3**). Proteoforms can also result from modifications after the amino-acid assembly, or *translation*. These are called *post-translational modifications* (PTMs) and like amino acids, there are many different types of these. PTMs can be thought of as functional accessories, and they often change the function of the protein to which they were added by changing its structure or its properties. PTMs diversify protein function, and this leads to production of a variety of different proteoforms (from a single gene) that can carry out different jobs (**Figure 2.3**). Proteoforms can also be modified by more than one type of PTM. A protein can

wear different types of functional accessories; just like a person can wear a hat, a scarf, or *both* a hat and a scarf.

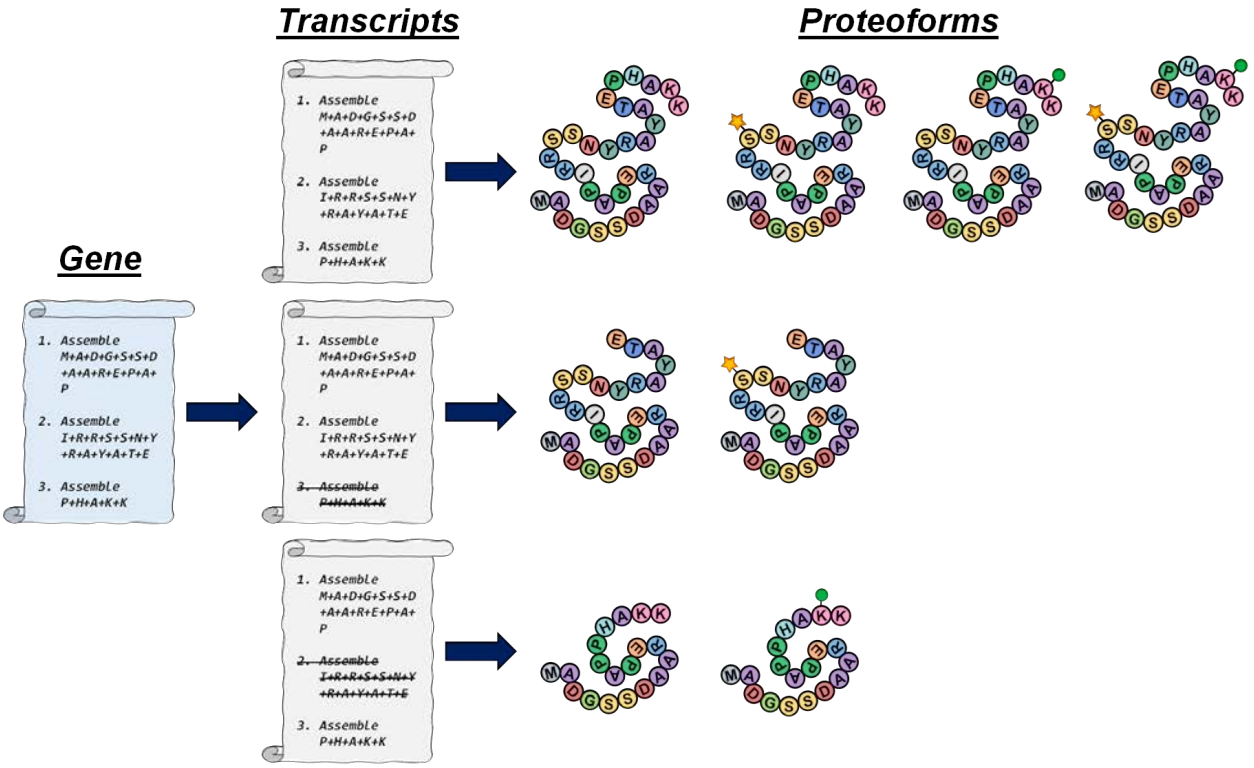

Figure 2.3. Several proteoforms produced from a single set of instructions: A gene, or a set of instructions, for assembly of a protein (left)**.** Transcripts, or copies of the original gene that have been altered or edited (middle). Proteoforms, or different forms of a protein that are made from the same gene following alternative splicing and addition of post-translational modifications (right).

Unlike genes, counting the number of proteoforms that exist in nature is not straightforward.[2] Even though we have ~20,300 protein-coding genes, we know that there are processes occurring after a set of instructions is unraveled (*e.g.* alternative splicing and PTMs) which give rise to many proteoform possibilities. As discussed, proteoforms can have different functional roles in a cell, despite originating from the same gene. Studying proteoforms in our cells helps us understand how our body works, at truly the most basic level. Importantly, proteoforms which don't function properly can cause disease. Think of genetic mutations, or type-o's in the

instructions- these can result in the mis-assembly of amino acids, which lead to incorrect protein-folding. As a result, the protein cannot perform its job. Mutant proteins that are caused by these errors are also types of proteoforms. Changes to regularly programmed alternative splicing can also lead to disease. Imagine if a factory received an order for 20 full-size refrigerators, but instead they produced 20 mini refrigerators. Both can store food, but one to a lesser extent. You see, it is important that we know which proteoforms our cells are producing and how many because it affects the processes our cells can carry out.

2.5 Individualizing Proteoforms with Mass Spectrometry

One way that scientists study proteoforms is by weighing them and determining their abundance in cells with a technique called *mass spectrometry* (MS). A mass spectrometer is like a molecular scale, which measures the weight (or *mass*) of molecules indirectly by measuring their mass-to-charge ratio (*m/z*). It is important to note that while we will use "weight" and "mass" interchangeably for simplicity, these terms refer to different measurements, with weight considering the pull of gravity. Modern-day mass spectrometers can measure the mass of intact proteoforms extremely accurately. The mass spectrometers available today can also differentiate between two proteoforms that are very closely related in structure, and therefore mass. Because we want to measure proteoforms that have been modified with PTMs, which make up only a small percentage of the proteoform weight, this feature of a mass spectrometer is extremely important.

The analysis of biomolecules (like proteins) was made possible by the invention of electrospray ionization (ESI) in the 1980s. ESI allowed proteins to be taken from solution-phase to gas-phase. ESI was so significant in enabling the analysis of biomolecules (e.g. peptides and proteins) by mass spectrometry, that its developer, Dr. John Fenn, won the Nobel Prize in Chemistry in 2002. In ESI, a high voltage is applied to a solution of proteins, which simultaneously

aerosolizes and charges them. Thus, the proteins are now present as ionized (charged) gas-phase molecules. Ions are molecules which carry an overall positive or negative charge, which is why we can manipulate and filter them inside of a mass spectrometer and measure their mass-to-charge ratio, or *m/z*. ESI is called "soft" ionization because the structure of the protein molecules is kept intact during the ionization process. After ESI, protein molecules (now protein *ions*) can carry different charges. The mass spectrometer measures the mass-to-charge ratio (*m/z*) of these protein ions, and a group of protein ions with the same mass, *m*, will be detected at a different *m/z* based on the charge, *z*, of the ion.

To illustrate this concept, let's consider that we have three copies of a proteoform with mass 3,521 Daltons (a unit of mass) (**Figure 2.4**). Imagine that, after electrospray ionization, all three proteoform ions carry five positive charges (5+). This means that we would detect a peak at 704 *m/z* (mass/charge → 3521/5=704) in our mass spectrum (**Figure 2.4A**). For this case, all three proteoform copies contribute to the intensity of the peak detected at 704 *m/z*. This would be similar if you had a pitcher of lemonade, and poured its entire contents into one glass. Now consider the same three proteoforms, but after electrospray ionization, each proteoform ion carries a different number of positive charges: 5+, 4+, and 3+. We would detect peaks at 704 *m/z*, 880 *m/z*, and 1174 *m/z*, respectively, in our mass spectrum (**Figure 2.4B**). As a result, the intensity of each peak will be 1/3 the intensity of the single peak detected at 704 *m/z* in the first case. This can be compared to dividing the pitcher of lemonade among three glasses, instead of one. As a result of dividing the lemonade three-ways, each glass is less full. The full-ness of a glass can be compared to the intensity of an ion. Note how the MS signal is spread across an increased number of charge states in **Figure 2.4B** compared to **Figure 2.4A,** resulting in a less intense signal at each *m/z* peak. This signal-spreading is amplified for large proteoforms which can carry a higher number of charges.

We will come back to this concept later when we talk about what this means for the detection of very large proteoforms.

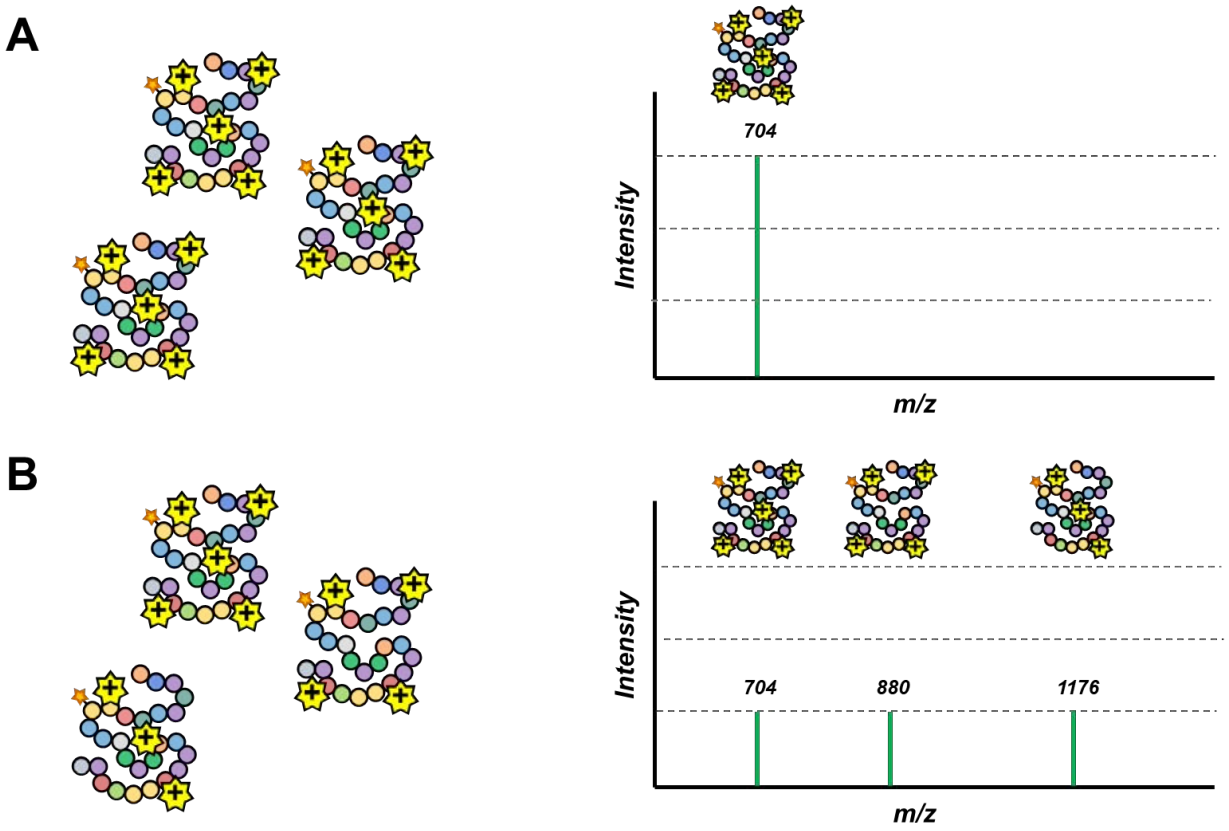

Figure 2.4. Signal spreading across multiple proteoform ion charge states: A) A group of proteoform ions with 5+ charge and resulting ESI mass spectrum **B)** A group of proteoform ions with 3+, 4+, 5+ charge and resulting ESI mass spectrum. Note how the signal is spread across an increased number of charge states.

Knowing the weight of an intact proteoform helps us to learn its identity, which is one reason we use mass spectrometry as a tool to study these molecules. However, proteoform mass, like any single metric, can only tell us so much about a proteoform. While knowing a person's weight, eye color, or hair color can help us identify groups of similar people- these characteristics, alone, are not *individualizing*. To individualize a proteoform, and distinguish it from another proteoform with the same mass, we need more information about it. For example, from what set of instructions (gene) was its structure based and what is its exact amino acid sequence? Does it contain PTMs? If so, *what* type and *where* are they located along the amino-acid backbone? This type of information can be obtained by breaking an intact proteoform into smaller pieces and measuring the mass of those pieces. This process is called tandem mass spectrometry (MS/MS). Fragmenting the proteoform into smaller pieces helps us to determine its identity more confidently and distinguish it from another proteoform that has the same mass. Proteoform ions are broken into smaller pieces by colliding them with gas molecules or exposing them to electrons. The fragments produced in this process create a unique "signature" and we can use this information together with the intact mass to *individualize* the proteoform.

Say that you are trying to identify an individual person from a group of people, and all you know about the individual is their weight. As you can imagine, knowing that a person weighs 180 lbs. helps to eliminate people who weigh more or less than 180 lbs., but this still leaves you with all the candidates who weigh 180 lbs. If you happen to have a very precise scale, you can know that the person of interest weighs 180.145 lbs. This helps you eliminate people who weigh more or less than 180.145 lbs., and certainly, there are fewer people weighing exactly 180.145 lbs. Even still, to help distinguish one 180.145-lb. person from another, we must examine unique features of the people. Fingerprints, for example, are unique from person to person. Collecting an MS/MS

spectrum for a proteoform is like collecting its fingerprints, and knowing the intact proteoform mass together with information about its amino acid backbone and PTMs can help us to identify and individualize the proteoform more confidently.

All of this seems simple enough if your goal is to measure the mass and abundance of a single proteoform. But, as we will soon explore, this process becomes more complicated when your goal is to measure the mass and abundance of the multiple proteoforms that exist in cells. As you can imagine, trying to weigh many proteoforms at once could overwhelm the molecular scale, and compromise the quality of the information, or mass spectrum, that is obtained.

2.6 Weighing Proteoforms From Mixtures

So, you want to measure the mass of proteoforms in cells? Congratulations! You're among many other interested parties who seek to study proteoforms in living systems by measuring their mass and abundance. First things first, you must break the cells, and release the proteoforms. At this point, you have a mixture of proteoforms in a solution of mostly salt and water. It is not easy to simply select a proteoform from the mixture to weigh it. First, we need to separate the proteoforms in the mixture from one another so that they are easier to weigh on our molecular scale. We can separate one proteoform from another by their different physical properties. In this way, we can measure proteoform masses in a certain order. To better picture this, imagine a doctor asked a group of her patients to line up by the color of the shirt they were wearing. If they organized themselves by the colors of the rainbow, patients wearing red shirts would be weighed first, and those wearing purple shirts would be weighed last. There are different ways we can separate mixtures of proteoforms, but one of the most common ways is to separate by *hydrophobicity*, or how "afraid" the proteoforms are of water. The proteoforms least afraid of water rush into the mass spectrometer first, and the most afraid of water enter the mass spectrometer last.

Unlike the example used above, proteoforms from complex mixtures do not simply enter the mass spectrometer in a single-file line. Typically, crowds of proteoforms with the same hydrophobicity enter the mass spectrometer at the same time. Often, this means that not all proteoforms make it inside the mass spectrometer to be detected or that only the most abundant proteins are detected by the mass spectrometer. Our proteome, which refers to our body's collection of proteins, is complex, containing thousands of proteoforms. Realistically, by using only one property to sort proteoforms before weighing them, we only detect the most abundant proteoforms in the mixture. Additionally, smaller proteoforms are easier to detect with mass spectrometry. So, if small and large proteoforms with similar hydrophobicity enter the mass spectrometer at the same time, only the small proteoforms will be detected. We observed this phenomenon when we separated a mixture of proteoforms extracted from human heart tissue by hydrophobicity. Heart tissue is made up primarily of heart cells, or cardiomyocytes. These cells contain a high abundance of highly specialized proteins which help our heart muscle beat, called myofilament proteins. Myofilament proteins are very abundant compared to other proteins in heart cells, and mostly very small. For these reasons, we could only detect these small, high-abundance myofilament proteins and their proteoforms with our mass spectrometer. However, we know that there are many more proteoforms present in our mixture of heart proteoforms than just the myofilament proteoforms- many much larger- that are not detected by our mass spectrometer.

Why are large proteoforms more difficult to "see," or detect, with a mass spectrometer than small proteoforms? Naturally, you might think that large proteoforms would be *easier* to see than small proteoforms, just as an elephant is easier to see than a mouse. However, this is not the case with MS, and it has a lot to do with the way that proteins are ionized in ESI. Recall that ESI creates multiply-charged protein ions. Large proteoforms typically have longer amino-acid backbones,

which means they can “carry” more charge. So, what does this mean for the detection of a group of large proteoform ions (of the same *m*, but many different *z’s*) from complex mixtures in a mass spectrometer? These large proteoforms will exist in several charge states and the signal will be spread across many *m/z*’s. Imagine the example from **Figure 2.4**, but with a much larger amino acid backbone and with tens (**Figure 2.5**) or even hundreds of charge-states. With more charge-states, the signal of one proteoform can be divided many times. Imagine sharing a pie with one other person, versus sharing a pie with your whole family. A proteoform that can hold more charge divides its molecules among different charge states, just like sharing a pie with your many cousins at a family reunion. The more people you must share a pie with, the less pie each individual person will receive. The more charge a proteoform can hold, the more charge-states that will exist, with fewer proteoform molecules existing at each charge-state. This illustrates one major issue associated with MS detection of larger proteoforms. We refer to this issue as a reduction in the MS signal-to-noise (S/N) with increasing proteoform size; as the size of a proteoform increases, its signal gets divided over more *m/z*’s. This reduction in signal-to-noise with increasing proteoform size is exponential.

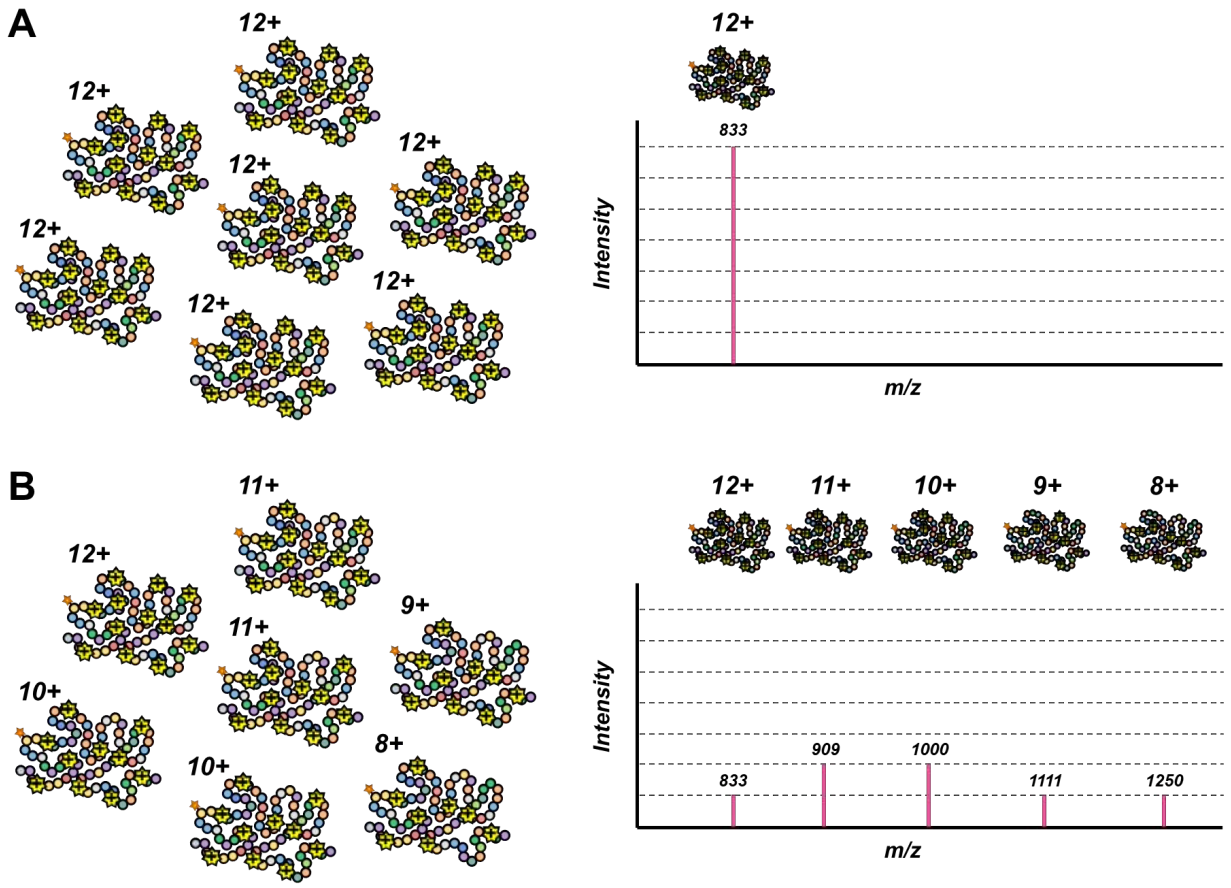

Figure 2.5. Signal spreading across multiple proteoform ion charge states for larger proteoforms: A) A group of proteoform ions with 12+ charge and resulting ESI mass spectrum. **B)** A group of proteoform ions with 8+, 9+, 10+, 11+, and 12+ charge and resulting ESI mass spectrum. Note how the signal is spread across an increased number of charge states. The proteoform is larger, so can carry more charge, and its signal is spread across more charge states.

Now, consider a group of large proteoforms, which are already difficult to detect due to their size, entering the mass spectrometer at the same time as a group of smaller proteoforms. The smaller proteoforms will "mask," or *suppress*, the detection of the large proteoforms. The good news is that there is an easy way to eliminate the issue of small proteoforms entering the mass spectrometer at the same time as large proteoforms. The solution is to separate the proteoform mixture by *size*.

2.7 Sorting by Size Helps with Weighing Larger Proteoforms

Using more than one type of separation can help us measure the mass of different proteoforms that exist in a complex mixture. For instance, instead of separating the mixture of heart proteoforms only by their hydrophobicity, we can first separate the mixture by size. This allows us to separate the most abundant, small myofilament proteoforms from the other heart proteoforms in the mixture that was mentioned earlier, meaning they can no longer interfere with the MS detection of larger proteoforms.

To separate the mixture of heart proteoforms by size before separating them by hydrophobicity, we used a technique called *size exclusion chromatography* (SEC). SEC works by sorting a mixture of molecules by their size, sort of like a molecular sieve. This happens by passing a solution of proteoforms through a porous matrix. Some of the proteoforms will be small enough to pass into and out of the pores in the matrix, and so it will take them more time to pass through the matrix. Large proteoforms, on the other hand, will pass through the matrix quicker than the small proteoforms because they took fewer pit-stops along the way. In this way, proteoforms are separated based on how long it takes them to pass through the porous matrix, and this depends on their size. The largest proteoforms leave the matrix first, and can be collected separately from the small proteoforms, allowing them to be further separated by their hydrophobicity and eventually weighed by MS. Not only did we detect more proteoforms from our mixture, we increased the number of large proteoforms (with mass over 60,000 Daltons) that we could weigh by *15-fold*, compared to the number we could weigh by just separating the mixture by hydrophobicity.[3] We detected a proteoform with a mass of 223,000 Daltons!

Even though separating by both size *and* hydrophobicity increased the number of large proteoforms we detected with mass spectrometry, compared to separation by just hydrophobicity,

we still had trouble identifying them. Remember that MS/MS can be used to obtain more information about a proteoform, and supplement the intact mass information we have. Proteoforms break apart in predictable ways, so the fragments produced can be pieced back together to determine the amino acid sequence and identity/location of PTMs. However, the larger a proteoform, the more fragments that are produced, creating a very complex puzzle to piece together. To overcome these issues and identify large proteoforms in the heart, we used two data analysis tools to help with piecing our puzzle back together.

Think about if you had a pile of puzzle pieces, that you were trying to put together. It would really help if you knew what the puzzle was supposed to look like at the end. By measuring the masses of the large proteoforms, and comparing them to a list of known proteoform masses, we were able to generate a list of proteoform "candidates." With a few potential proteoform candidates, we could use the fragments to try to piece together the proteoform based on the narrowed possibilities. We did this by comparing the mass of proteoform fragments to the "theoretical" masses of candidate proteoform fragments. Because proteins fragment in a predictable way, we can predict what fragments might be produced. If we found a match, we called this an amino acid "sequence tag," which helped us more confidently identify the large proteoforms using both their intact mass and fragment masses. This data analysis technique allowed us to identify *13 new large heart proteoforms* by simply knowing their intact mass and some potential candidates with a similar mass.[4]

2.8 Proteoforms in Heart Disease

We now know that we can determine the mass of a proteoform and use fragments produced with MS/MS to obtain information about what makes it unique. Using mass spectrometry, we can also measure the amount, or *abundance*, of the proteoforms from cells and tissues, and this helps us learn about health and disease in biological systems. The amount of a proteoform present in a mixture will be proportional to the signal it produces in the mass spectrometer relative to other proteoforms in the same mixture, assuming they were all ionized in the same way.

In my own research, I have used mass spectrometry to measure differences between proteoform levels in healthy heart tissue and diseased tissue. Remember that the basic contractile unit in our heart cells is called a *sarcomere*. These units are repeating, and the sarcomere proteins work as a team to control the contraction and relaxation of our heart muscle. The sarcomere is made up of a complex network of proteoforms, which communicate with each other by sending signals. These signals often come in the form of addition and removal of PTMs. Recall that proteoforms can have different combinations of PTMs which affect their structure and function. If these PTMs are perturbed, it can affect the communication in the sarcomere, the heart cells, and the whole heart muscle. For this reason, we use mass spectrometry to collect the weight of heart proteoforms and identify them, but also measure how many of a specific proteoform are present (relative to the other proteoforms).

We can compare proteoform levels between cells and tissues from different people to better understand the difference between healthy individuals or sick individuals. Recently, I used MS to investigate proteoform changes in a prevalent heart disease, called hypertrophic cardiomyopathy, or HCM. HCM is caused by mutations in the genes which are used to produce sarcomere proteins. In this disease, the heart wall thickens, preventing efficient pumping of blood. Often, patients must

undergo open heart surgery to have the thickened piece of the heart wall removed. I compared proteoforms in this piece of HCM tissue to proteoforms in tissue from healthy donor hearts and found significant differences in proteoforms levels which indicated the addition and removal of a PTM which is important in sarcomere protein communication, or signalling. We compared proteoforms in the tissues of 16 HCM patients (all with different disease-causing gene mutations) with those in the tissues of 16 healthy donors. Interestingly, the proteoform changes were consistent among the 16 HCM tissues. This suggests that similar signals are produced during disease progression, regardless of disease-causing gene mutation. This is just one example of how MS can be used to study proteoforms in our heart to understand disease.

Some scientists who study proteins in cells with MS choose to chop up them into smaller amino acid chains (peptides) before measuring their mass. Then, the peptides are identified by intact mass and MS/MS. In this way, proteins in a mixture can be indirectly identified and quantified from the *bottom-up*. Weighing peptides is much easier than weighing proteoforms, of course, because peptides are smaller than proteins. However, by chopping up proteins, intact proteoforms are never observed, and their mass and abundance cannot be measured by MS. To study relative amounts of proteoforms, we must take a *top-down* approach, which involves measuring the proteoforms intact and then obtaining sequence information directly.

While the top-down approach is a powerful technique for studying proteoforms in living systems, it still faces several technical challenges. One challenge described in this chapter is the detection and identification of large proteoforms from complex mixtures. The separation technique discussed above, which addresses this issue, requires a lot of starting tissue. For many of the diseases we seek to study with top-down MS, only a small amount of tissue or cells are available since they often come from living patients. Therefore, methods amenable to smaller amounts of

sample are required to study larger proteoforms in health and disease. Scientists are working toward this goal such that someday we may be able to measure proteoform levels from single cells. Perhaps in the future, we can use the top-down MS to understand proteoform profiles for each individual person, such that we can personalize treatment for disease.

2.9 Further Reading

- "Proteoform: a single term describing protein complexity" Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P. *Nat Methods* **2013**, *10*, 186-187.
- "How many human proteoforms are there?" Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L., et al. *Nat Chem Biol* **2018**, *14*, 206-214.
- "Top-down proteomics of large proteins up to 223 kDa enabled by serial size exclusion chromatography strategy" Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y. *Anal Chem* **2017**, *89*, 5467-5475.
- "Intact-mass analysis facilitating the identification of large heart proteoforms" Schaffer, L. V.; Tucholski, T.; Shortreed, M. R.; Ge, Y.; Smith, L. M. *Anal Chem* **2019**, *91*, 10937-10942.

REFERENCES

(1) Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P. In *Nat Methods*: United States, 2013, pp 186-187.
(2) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L., et al. *Nat Chem Biol* **2018**, *14*, 206-214.
(3) Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y. *Anal Chem* **2017**, *89*, 5467-5475.
(4) Schaffer, L. V.; Tucholski, T.; Shortreed, M. R.; Ge, Y.; Smith, L. M. *Anal Chem* **2019**, *91*, 10937-10942.