Section 1

Proteins and Proteomics

Learning Objectives

At the end of this assignment, you should be able to:

1. Draw the chemical structure of an amino acid and small peptide.
2. Describe the difference between free and residue amino acid mass.
3. Calculate the monoisotopic and average mass of a peptide.
4. Define proteomics and interpret a scientific paper on the application of proteomics in cancer research.

Section 1A. What is a protein?

Proteins are biological macromolecules consisting of long chains of amino acids. A shorter chain of amino acids is called a **peptide**. Proteins and peptides are biological polymers formed from amino acid monomers. Each amino acid is composed of an amine group, a side chain (R), and a carboxylic acid functionality (Figure 1).

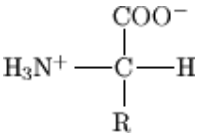


Figure 1. An amino acid.

There are 22 naturally occurring amino acids, 20 of which are encoded by the genome (Figure 2). These amino acids are usually grouped according to the character of their side chains, which may be acidic, basic, neutral, or hydrophobic. In a protein, portions of the linear sequence of amino acids may take on a secondary structure (such as a helix or a sheet) based on intermolecular forces between amino acid residues. The full-length protein will form a tertiary structure or overall shape, and the structure of the protein is intimately linked to the protein's function.

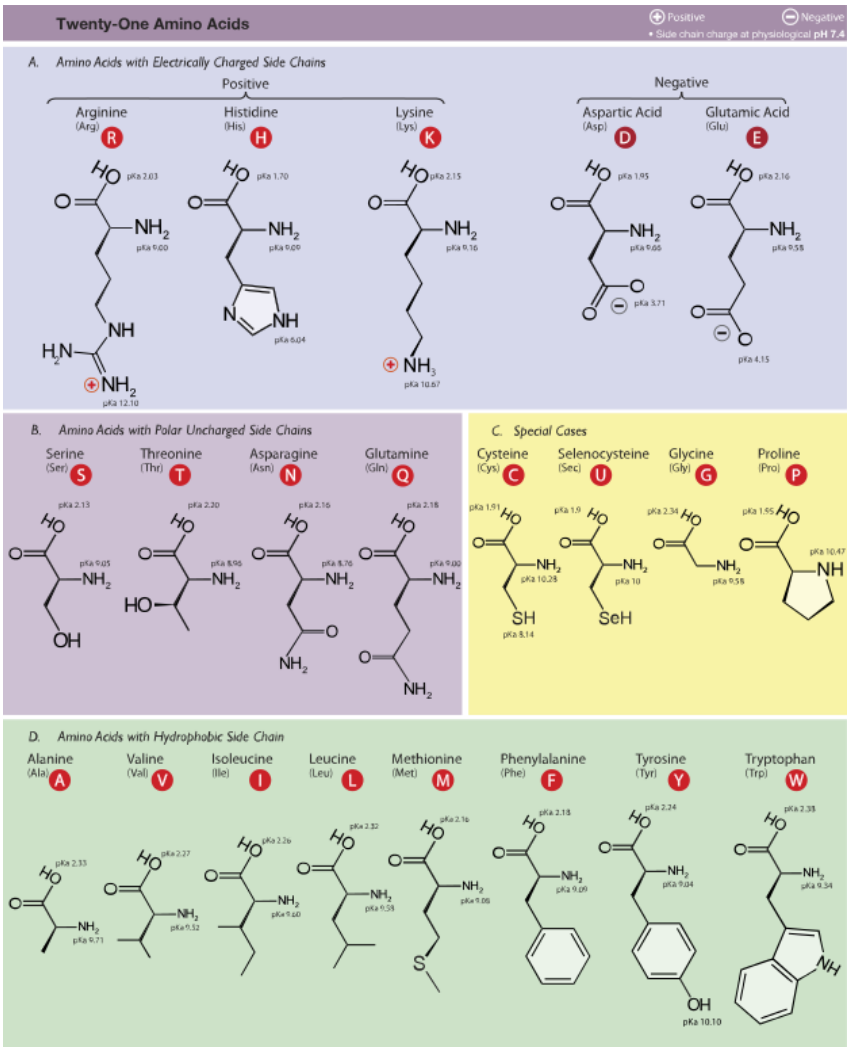


Figure 2. Table of naturally occurring amino acids sorted by side chain and with their three-letter and one-letter codes. (Reproduced under a Creative Commons license from Dancojocari.)

Reading Question

1. In Figure 1 above, circle the amine group, draw a triangle around the side chain, and draw a square around the carboxylic acid group of the amino acid.

The linear chains of amino acids that form proteins are connected by amide (or peptide) bonds that form by a dehydration reaction between two amino acids (Figure 3). By convention, we write and draw peptides from the N-terminus (the side with the free amine group) to the C-terminus (the side with the free carboxylic acid).

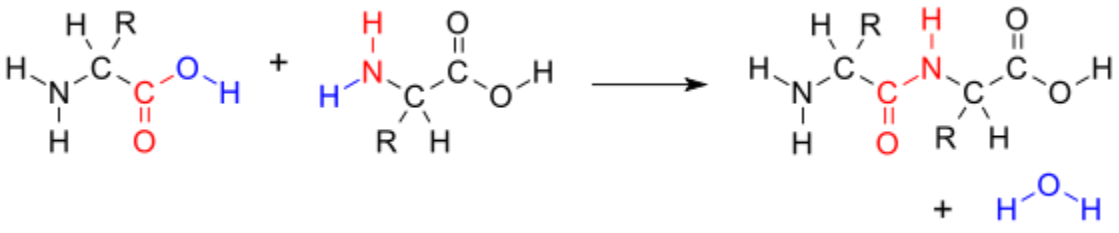


Figure 3. Formation of a peptide bond between two amino acids.

Reading Question

2. In Figure 3 above, label the N-terminus and the C-terminus of the dipeptide product. Circle the amide bond.

While many properties of peptides and proteins can be measured to provide useful information, we will focus on measurements of peptide and protein masses using mass spectrometry. Mass spectrometry measures the mass and charge of molecules in the gas phase. As a result, the mass of individual molecules is important. The **monoisotopic mass** of a molecule is the mass of that compound when it is composed of the most abundant isotope of each atom in the molecule. For example, if we calculate the monoisotopic mass of CO_2, we will use the masses of carbon-12 and oxygen-16 since these are the most abundant isotopes of these elements. This gives us a monoisotopic mass of 12.000 amu + 2 (15.995 amu) = 43.99 amu. This is different than the molar mass we would calculate using average atomic weights from the periodic table; on the periodic table, we find a molar mass for carbon of 12.011 amu. This represents the weighted average of all carbon atoms, which includes 98.9% carbon-12 and 1.1% carbon-13. Table 1 shows the monoisotopic mass of each amino acid and the mass of that amino acid as a residue in a protein or peptide chain.

Table 1. Monoisotopic molecular weight information for all 20 genetically encoded, naturally occurring amino acids.

Amino Acid	Single-Letter Code	Residue MW (amu)	Amino Acid MW (amu)
glycine	G	57.02	75.03
alanine	A	71.04	89.05
serine	S	87.03	105.04
proline	P	97.05	115.06
valine	V	99.07	117.08
threonine	T	101.05	119.06
cysteine	C	103.01	121.02
isoleucine	I	113.08	131.09
leucine	L	113.08	131.09
asparagine	N	114.04	132.05
aspartic acid	D	115.03	133.04
glutamine	Q	128.06	146.07
lysine	K	128.09	146.11
glutamic acid	E	129.04	147.05
methionine	M	131.04	149.05
histidine	H	137.06	155.07
phenylalanine	F	147.07	165.08
arginine	R	156.10	174.11
tyrosine	Y	163.06	181.07
tryptophan	W	186.08	204.09

Discussion Questions

1. Consider the data in Table 1. By what value do the residue molecular weight (MW) and the amino acid MW differ? Why is the MW of an amino acid residue in a peptide chain different from the mass of the full amino acid?

2. Draw the structure for the tetrapeptide G, C, L, W. Refer to Figure 2 for the structure of amino acid side chains.

3. Calculate the monoisotopic molecular weight of the tetrapeptide using the data in Table 1.

4. Calculate the molecular weight of the tetrapeptide using molar mass information from the periodic table. Why is this molecular weight different from the monoisotopic mass you calculate in question 3? Which mass is the mass measured in mass spectrometry?

Section 1B. What is proteomics?

The central dogma of molecular biology, DNA to RNA to protein, has given us an explanation of how information encoded by our DNA is translated and used to make an organism. It describes how a gene made of DNA is transcribed by messenger RNA and then translated into a protein by transfer RNA in a complex series of events utilizing ribosomal RNA and amino acids. Although in essence the central dogma remains true, studies of genes and proteins are revealing a complexity that we had never imagined. For example, distinct genes are expressed in different cell types and the physiological state of the cell alters which proteins are produced and at what level. Furthermore, chemical changes (i.e. phosphorylation) to proteins occur after translation and are critical to a protein's function. The importance and diversity of proteins started a whole new field termed proteomics.

Proteomics is the study of proteins, particularly their structures and functions. This term was coined to make an analogy with genomics. The Human Genome Project, started in 1990 and completed in 2003, sequenced three billion bases in genes (the human genome). The entire set of proteins in existence in an organism throughout its life cycle, or on a smaller scale the entire set of proteins found in a particular cell type under a specific set of conditions is referred to as the **proteome**.

Proteomics is much more complicated than genomics for several reasons. The genome is a rather constant entity while the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment. Consequently, the proteome reflects the particular stage of development or the current environmental condition of the cell or organism. One organism will have radically different protein expression in different parts of the body, in different stages of its life cycle, and in different environmental conditions. For example, when *E. coli* cells are grown under conditions of elevated temperature a class of proteins known as heat shock proteins are upregulated. Many members of this group perform a chaperone function by stabilizing new proteins to ensure correct folding or by helping to refold proteins that were damaged by the cell stress. Ultimately, the comparison of proteomes of healthy and diseased tissues may identify the molecular nature of a disease and provide potential new targets for drug development. The field of proteomics also presents many analytical challenges when compared to genomics. In DNA there are only four nucleotide bases with similar molecular weights and properties. In a proteome there are thousands of different proteins with a wide range of concentrations, molecular weights, and properties.

Proteomics was initially defined as the effort to catalog all the proteins expressed in all cells at all stages of development. That definition has now been expanded to include the study of protein functions, protein-protein interactions, cellular locations, expression levels, and post-translational modifications of all proteins within all

cells and tissues at all stages of development. It is hypothesized that a large amount of the non-coding DNA in the human genome functions to regulate protein production, expression levels, and post-translational modifications. It is regulation of our complex proteomes, rather than our genes, that makes us different from simpler organisms with a similar number of genes. An international collaboration of scientists in the human Proteome Project (HPP) is working to characterize all 20,300 genes of the known genome and generate a map of the protein based molecular architecture of the human body. Completion of this project will enhance understanding of human biology at the cellular level and lay a foundation for development of diagnostic, prognostic, therapeutic, and preventive medical applications.

Reading Questions

1. Define the term **proteome.**

2. Define the term **proteomics**.

3. Why is the analysis of proteins in a cell more difficult than sequencing DNA?

4. What types of questions can be answered by studying the proteome?

Read the following research paper to learn how the field of proteomics can be useful in the treatment of cancer.

Comparative proteomics of oral cancer cell lines: identification of cancer associated proteins, Karsani *et al,* Proteome Science, 2014, **12:**3. doi:10.1186/1477-5956-12-3 (Open access journal)

Discussion Questions

1. What was goal of the scientific study reported in the paper?

2. Why would a study of the change in oral cancer proteins add to our understanding of the disease?

3. Refer to the results and discussion section and Figure 1 to answer the following questions.

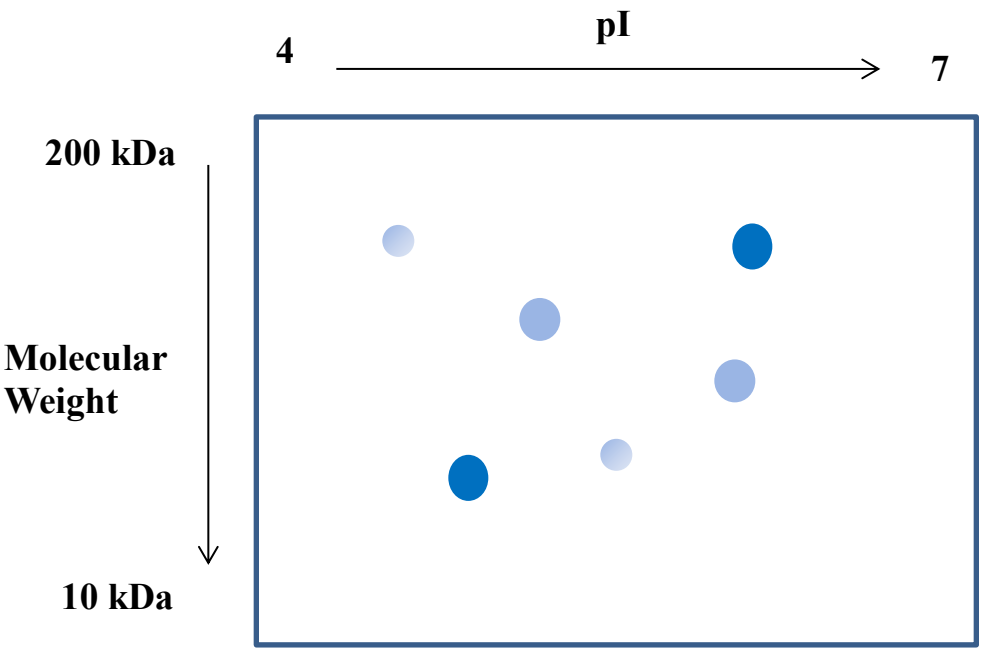
a. How were the proteins in the healthy and cancerous cells separated and detected?

b. How many individual protein spots were resolved on the silver stained gels?

c. How many protein spots exhibited a significant difference in abundance from normal cells to cancerous cells? *A. 24*

4. Table 1 is a list of proteins with different abundances in the cancer cell line.
(The section in this module on peptide mass mapping describes how the identity of the protein in the gel spot was determined.)
Examine the data for two structural proteins: Stathmin (STMN1) and myosin regulatory light chain-2 (ML12A). What is the change in abundance for each protein? Can this change be visualized from the image of the spot? How is the change quantified?

5. The simplified 2D gel shown represents the proteins from a healthy cell line.



Draw a new 2D gel which could represent the changes in protein expression that occur in a cancerous cell line.