Lecture 1

Introduction

Before coming to class consider the analogy of the cell as being the city. Where in a city are the plans and all of the directions to run and maintain the city come from? What would be the analogous portion of the cell? Think of the other things that a city must have in order to function properly and what in a cell will represent that function. We will discuss this analogy and the function of proteins within the cell to begin our discussion of proteomics.

So what is proteomics? In my thinking it is the process of identifying, characterizing, and quantifying all of the proteins found in an organism under different conditions. The purpose of this is to understand the role of the proteins in the functioning of the cell. The two hardest things are to identify and characterize all of the proteins under a certain condition of growth. In addition how do you define a different protein. Does one modification make it a different protein from another? If it is a regulatory modification the answer is yes. If it something that appears to have not effect on its function then the modification may not indicate a different protein.

Proteomics has followed genomics in its development. Initially genomics was just to sequence the genome of an organism. This aspect is still continuing and many scientists have the favorite organisms to be sequenced. As more and more are completed the comparisons are becoming useful.

Once the genomes were available the development of microarrays has allowed the determination of the expression profile of an organism. This process was problematic due to the many possible artifacts with the process and a long time was spent in determining the best methods to obtain reproducible and reliable data. Now genomics is being used to study many disease and normal states and is leading to the identification of the multiple mutations of base changes that can lead to the disease state.

Proteomics has followed a similar path. Initially proteomics began with the effort to identify and characterize all of the proteins present in a system at a set time under set conditions. This continues today since many of the proteins identified by genomics have not been observed. In many cases proteins that were not identified in genomics have been found through proteomics. There are many difficulties in doing this which we will discuss later. Methods are now being developed and improved to allow comparative studies to understand the functioning of various organisms. We will be discussing these methods. Most interestingly, many companies are working to define protein “markers” to indicate disease conditions. The potential for diagnostic proteins for disease conditions, especially if found in the blood, is driving a very large area of comparison proteomics in industry.

Many of the things you have already studied in genomics such as means of alignment and assigning possible function based on homology have been done. Still most

of the work presently is on identification and characterization of the proteins found in cells. Most of the lectures in this section will deal with the methods used to identify and characterize the proteins. We will also discuss some of the goals of proteomics. In the final analysis I think it will be the ability of proteomics to identify changes in protein function that will be the lasting usefulness of this type of study.

Proteomics

Let's start with the question: "What do we need or want to know about the proteome in a cell?"

1. Structure
2. Composition
3. Function
4. Regulation
5. Location
6. Interactions
7. Quantity

If we were to reduce this to the basics we would want to know the composition, structure and function.

There are several things we need to know about the composition. First of all would be the amino acids present. There are the normal 20 amino acids found in most eukaryotic organisms. These amino acids can be modified by various reactions which makes them different from the parent compound. Lastly there are unusual amino acids that are found in special systems or enzymes.

In addition to the amino acids, there are other components such as metal centers, hemes, cofactors such a cobalmin and ubiquinone. There are other additional compounds that I have not stated. One thing to consider is whether the item is covalently bound or not. If it isn't it could be lost during processing for analysis.

Many amino acids undergo further processing after incorporation into a protein. These modification are plentiful and influence the function. Some of the modifications are the result of oxidative damage or chemical damage.

Amino Acid Single Letter Designations

There are two common amino acids which come in two different stereo-chemical forms. There are both D and L amino acids. The stereochemistry may not seem important but it can affect the secondary structure and the selectivity of the enzyme. For this class you will need to learn the single letter identifiers for each amino acid. These are fairly easy to learn since most are logical. When devising this code the authors took a look at them and assigned the first letter of the amino acid. When there were two amino acids they used the first letter for one. Examples of the letters coinciding with the code are

Alanine	A
Cysteine	C
Glycine	G
Histidine	H
Isoleycine	I
Leucine	L
Methionine	M
Serine	S
Threonine	T
Valine	V

Other amino acids can be remembered by simple making simple memory aids. For example Tyrosine has the “”Y” sound at its beginning, “Ty” = “Y”. For phenylalanine the “ph” has the “F” sound and so it has the letter “F”. Tryptophan is easy to remember if you think I trip over a “W” and so the code for tryptophan is “W”. For lysine we know that we have already used the “L” for leucine. They authors then chose the next closest letter “K”.

The last few Aspatate, Glutamate, Asparagine, and Glutamine are probably the most difficult. If we start with Aspartate it starts with an “A” and so a letter close to “A” was chosen. The letters “A” and “C” were already chosen. Therefore they chose D. Since glutamate is close in structure to aspartate they chose “E” for glutamate. Asparagine was given the letter “N” since it has a “N” sound towards the end. That left Glutamine. I reason this to have a “Q” like sound at the beginning “Glu” = “Q”.

These are the ways I remember the letters. As you use them more you will get to where you know what they are just like knowing the notes in music.

Unusual and Modified Amino acids

There are several unusual amino acids that can be found in some proteins. These are usually found in proteins, such as toxins, that are not synthesized by robosomes. First of all, D amino acids are found in bacterial systems. Often this difference determines whether a certain compound which affects a common enzyme is active against bacteria but not humans. Ornithine is another amino acid that does not often appear in proteins but is a critical compound in the urea cycle and in the regulation of development through the formation of polyamines such as putriscine, spermine and spermidine. Other examples of unusual amino acids are homocysteine and sarcosine. Homocysteine if found in the synthetic pathway for methionine and for cysteine.

Several protein modifications are performed enzymatically. In some cases we understand why the modification is made and in others we are not certain. In the protein modification database there are some 450 modification documented. Almost all of the amino acids can be modified. http://dbptm.mbc.nctu.edu.tw/browse.php

Some examples are hydroxylation which is the addition of an OH group to the amino acid. For example hydroxyl proline is a required modification for collagen formation since it provides an important hydrogen bond as well as the forced turn. As you can see from the list there are many known hydroxylation processes.

Glycolsylation is adding sugars to different amino acids. This can occur through either an OH group as in the case of S, T, and (OH)P. In many cases the sugar is attached to the side chain N of asparagine.

Sulfination is basically the esterification of sulfuric acid to the alcohol amino acids such as tyrosine, serine and threonine.

Phosphorylation is the most known modification since this is a common control mechanism. This is usually the esterification of phosphoric acid to the alcohol of the amino acids S, T, and Y.

Several amino acids can be oxidized. Methionine can have one or two oxygens added to it to form methionine sulfoxide or methionine sulfone. Cysteine can also be oxidized. In some cases this occurs through time without an enzyme and in others it is enzymatically produced.

An acetyl group [CH3COOH] can be added to any primary amine such as the amino terminal, which it is used to promote stability, or K. It can also react with the sulfur in cysteine.

The addition of fatty acids is again an esterification reaction. Here the acid group reacts with primary amines, the sulfur of cysteine, and sometimes the carboxyl group of the C-terminus.

ADP ribosylation is essentially adding a sugar, ribose, and connecting the ribose to ADP. This sugar will then react with the side chain nitrogen of asparagine, lysine, arginine and histidine.

Disulfide bond formation is obvious for maintaining the structure but many different disulfide bonds can be made in a protein and so which ones exist is important.

Lastly is the addition of a methyl group to the side chains. As you can see from the list there are many amino acids that can be modified by the methyl group. There are 753 known examples of methylation of different amino acids.

Protein Modification

I will mention only two protein modification. I call these protein modifications since the actual modification is to tag a protein with another completely independent protein. Ubiquitin is the protein that is used by a system to mark proteins for degredation. SUMO is a protein to actually mark a protein for keeping it around.

Lastly is proteolytic processing. Many proteins are produced that are not functional until a portion of the protein is removed. This is true for insulin and or ornithine decarboxylase. In developing embryos ornithine decarboxylase is synthesized and attached to the cell membrane where it is inactive. When polyamines are needed the enzyme is cleaved from the membrane and then is rapidly degraded. In flies the total time the enzyme is active for development is only about half an hour. Insulin is formed a one long peptide that folds making two disulfide bonds but is inactive. When a 31 amino acid portion is cleaved from the two ends it becomes active. The active insulin consists of two fragments, A and B, that are held together by disulfide bonds. The C-peptide was originally though to have not function but recent work has shown that it is involved in reducing neuropathy associated with insulin dependent diabities.

Cofactors

There are several different cofactors found in proteins. The problem is that they can be lost during processing or fortuitously bind to the proteins. Many metals are found in proteins. Other metals are actually the substrate for some reactions or processes. Iron is found in iron sulfur proteins and hemes. Copper is found in cytochrome oxidase. Molybdimum is found in nitrogenase and Zinc is found in many exymes and zinc fingers. Manganese (Mn) is found in the water oxidizing enzyme while magnesium is found in chlorophyll and chlorines.

Other small molecules that will bind to proteins include the nucleoside phosphates, coenzymes, vitamins, NADH, NADPH and allosteric regulatory molecules. An example of an allosteric molecule in hemoglobine would be 2,3 bisphosphoglycerate. Binding of this molecule results in the decrease in the binding ability of hemoglobin for oxygen. In cases where there is substrated inhibition it is often due to an allosteric site on the protein. Knowing about these allows us to better understand the control of the protein.

Protein Structure

Besides the proteolytic cleavage to activate a protein many proteins must be exported to a different compartment of the cell. In addition a portion like the C-peptide of insulin may also have to be removed.

The Signal sequence is a sequence that tells the protein where to go. For secretion to the outside of the cell, transport to the mitochondria or chloroplasts, the signal is on the N terminus. To remain in the endoplasmic reticulum the signal is on the c terminus and one example is the sequence KDEL or HDEL. Nuclear localization sequences are found internal to the protein are not well defined as yet. So a protein may have but not required to have a signal sequence, a prosequence or an internal sequence, and then the final mature peptide. The mature peptide may undergo additional modifications.

In protein we call the amino acid sequence the primary structure. The basic types of repetitive protein structure are known as secondary structure and include helices, beta sheets, random coil (somewhat a misnomer since the structure is usually anything but random) and turns of various types. The tertiary structure is how the various secondary structures are assembled. There are regular forms of these structures and many have been named. Examples are the b-barrel the four helix bundle and specialized turns such as the E-F hand. The quarternary structure is the structure formed by multiple subunits.

The next slide shows an alpha-beta structure and a beta barrel. The beta sheets in both of these are antiparallel. Understanding the features of these structures can help in predicting the structure of an unknown protein. For example antiparallel beta sheets are hydrophobic on one side and hydrophilic on the other. Parallel beta sheets are hydrophobic on both sides. Helices that are exposed to water are often amphipathic, one side hydrophobic and one side hydrophilic. This feature can be determined by looking at a helical wheel display of the sequence. There are many combinations of these basic units that are frequently observed including the 4 helix bundle, the triple helix coiled coil, and the EF hand. There are many other structures but we do not have the time to look at them all.

Function

So what do proteins do? My answer, they do all of the useful work and build all of the structures within a cell. In doing this they move from one location to another with the location possibly determining the function of the protein, appear in certain tissues and not others, and they are found in both membranes, in the cytosol, and in the extracellular matrix.

Modifications can alter the function or they can modify other proteins. Proteins will associate with themselves and other proteins to form functional units and lastly their function may be controlled to happen only during specific times of development.

Protein Regulation

Proteins are regulated in a large number of ways. They are regulated at transcription and translation. They are regulated by proteolysis either to activate them or to remove them from the cells. They can be phosphorylated to activate or deactivate and they can bind several different signaling molecules that will either modulate, activate or deactivate their function. Finally the proteins function can be controlled by the concentration present such that there is not enough of the product to have an effect.

Protein Expression

There are basically two types of expression, homologous and heterologous. Heterologolous expression can have many problems. If the protein is toxic to the

expression system its production must be controlled until the very last where lots can be made at the expense of the host cell.

Another problem with heteroglolous expression is that the proteins will not fold properly and float unfunctional in the cytoplasm or can precipitate in inclusion bodies. When eukaryotic proteins are expressed in prokaryotes they are rarely processed correctly and do not have the typical modifications found in the original organism. Finally they may not fold properly because they do not have the appropriate binding partners for the complex.

Inclusion bodies

Inclusion bodies are pellets of protein found in the expression system that are due to coalescence of the proteins into one large, usually non-functional blob. The inclusion bodies can be up to 90% pure protein but can also be a mixture of proteins from the expression host with the expressed protein. If inclusion bodies occur the protein will need to be purified and refolded. Usually refolding is not very successful leaving a large number of non-functional proteins.

Missing modifications

Many times to get a protein product that behaves as the original it will require the same modifications as observed in its original organism. When proteins from eukaryotes are expressed in prokaryotic organisms these modifications do not occur. In a few cases these modification assist in the folding and so the proper folding is not obtained.

Missing interactions

When a protein is part of a complex, the production of the protein without the other components of the complex can result in it being degraded quickly or making an inclusion body.