**PROTEIN SEQUENCING**

Purification of a protein is usually only a prelude to a detailed biochemical dissection of its structure and function. What is it that makes one protein an enzyme, another a hormone, another a structural protein, and still another an antibody? How do they differ chemically? The most obvious distinctions are structural, and these distinctions can be approached at every level of structure. The differences in primary structure can be especially informative. Each protein has a distinctive number and sequence of amino acid residues. The primary structure of a protein determines how it folds up into a unique three-dimensional structure, and this in turn determines the function of the protein. Therefore, understanding the primary structure given an idea about the basic functioning of the protein and one of the key methods to understand the primary structure of the protein is sequencing the proteins.

The bacterium *Escherichia coli* produces more than 3,000 different proteins; a human produces 25,000 to 35,000. In both cases, each type of protein has a unique three-dimensional structure and this structure confers a unique function. Each type of protein also has a unique amino acid sequence. This amino acid sequence must play a fundamental role in determining the three-dimensional structure of the protein, and ultimately its function.

Two major discoveries in 1953 were of crucial importance in the history of biochemistry. In that year James D. Watson and Francis Crick deduced the double-helical structure of DNA and proposed a structural basis for its precise replication. Their proposal illuminated the molecular reality behind the idea of a gene. In that same year, Frederick Sanger worked out the sequence of amino acid residues in the polypeptide chains of the hormone insulin, surprising many researchers who had long thought that elucidation of the amino acid sequence of a polypeptide would be a hopelessly difficult task. It quickly became evident that the nucleotide sequence in DNA and the amino acid sequence in proteins were somehow related. Barely a decade after these discoveries, the role of the nucleotide sequence of DNA in determining the amino acid sequence of protein molecules was revealed. An enormous number of protein sequences can now be derived indirectly from the DNA sequences in the rapidly growing genome databases. However, many are still deduced by traditional methods of polypeptide sequencing.



Frederick Sanger

The amino acid sequences of thousands of different proteins from many species have been determined using principles first developed by Sanger. These methods are still in use, although with many variations and improvements in detail. Chemical protein sequencing now complements a growing list of newer methods, providing multiple avenues to obtain amino acid sequence data. Such data are now critical to every area of biochemical investigation.

**Short Polypeptides Are Sequenced Using Automated Procedures**

Various procedures are used to analyze protein primary structure. Several protocols are available to label and identify the amino-terminal amino acid residue. Sanger developed the reagent 1-fluoro-2,4-dinitrobenzene (FDNB) for this purpose; other reagents used to label the amino-terminal residue, dansyl chloride and dabsyl chloride, yield derivatives that are more easily detectable than the dinitrophenyl derivatives. After the amino-terminal residue is labelled with one of these reagents, the polypeptide is hydrolyzed to its constituent amino acids and the labelled amino acid is identified. Because the hydrolysis stage destroys the polypeptide, this procedure cannot be used to sequence a polypeptide beyond its amino-terminal residue. However, it can help determine the number of chemically distinct polypeptides in a protein, provided each has a different amino-terminal residue.

To sequence an entire polypeptide, a chemical method devised by Pehr Edman is usually employed. The **Edman degradation** procedure labels and removes only the amino-terminal residue from a peptide, leaving all other peptide bonds intact. The peptide is reacted with phenylisothiocyanate under mildly alkaline conditions, which converts the aminoterminal amino acid to a phenylthiocarbamoyl (PTC) adduct. The peptide bond next to the PTC adduct is then cleaved in a step carried out in anhydrous trifluoroacetic acid, with removal of the amino-terminal amino acid as an anilinothiazolinone derivative. The derivatized amino acid is extracted with organic solvents, converted to the more stable phenylthiohydantoin derivative by treatment with aqueous acid, and then identified. The use of sequential reactions carried out under first basic and then acidic conditions provides control over the entire process. Each reaction with the aminoterminal amino acid can go essentially to completion without affecting any of the other peptide bonds in the peptide. After removal and identification of the aminoterminal residue, the *new* amino-terminal residue so exposed can be labeled, removed, and identified through the same series of reactions. This procedure is repeated until the entire sequence is determined. The Edman degradation is carried out on a machine, called a **sequenator,** that mixes reagents in the proper proportions, separates the products, identifies them, and records the results. These methods are extremely sensitive. Often, the complete amino acid sequence can be determined starting with only a few micrograms of protein.

**Large Proteins Must Be Sequenced in Smaller Segments**

The overall accuracy of amino acid sequencing generally declines as the length of the polypeptide increases. The very large polypeptides found in proteins must be broken down into smaller pieces to be sequenced efficiently. There are several steps in this process. First, the protein is cleaved into a set of specific fragments by chemical or enzymatic methods. If any disulfide bonds are present, they must be broken. Each fragment is purified, then sequenced by the Edman procedure. Finally, the order in which the fragments appear in the original protein is determined and disulfide bonds (if any) are located.

***Breaking Disulfide Bonds*** Disulfide bonds interfere with the sequencing procedure. A cystine residue that has one of its peptide bonds cleaved by the Edman procedure may remain attached to another polypeptide strand via its disulfide bond. Disulfide bonds also interfere with the enzymatic or chemical cleavage of the polypeptide. Two approaches to irreversible breakage of disulfide bonds are outlined in Figure below.





Figure: **Breaking disulfide bonds in proteins.** Two common methods are illustrated. Oxidation of a cystine residue with performic acid produces two cysteic acid residues. Reduction by dithiothreitol to form Cys residues must be followed by further modification of the reactive OSH groups to prevent re-formation of the disulfide bond. Acetylation by iodoacetate serves this purpose.

***Cleaving the Polypeptide Chain*** Several methods can be used for fragmenting the polypeptide chain. Enzymes called **proteases** catalyze the hydrolytic cleavage of peptide bonds. Some proteases cleave only the peptide bond adjacent to particular amino acid residues and thus fragment a polypeptide chain in a predictable and reproducible way. A number of chemical reagents also cleave the peptide bond adjacent to specific residues.

***Sequencing of Peptides*** Each peptide fragment resulting from the action of trypsin is sequenced separately by the Edman procedure.

***Ordering Peptide Fragments*** The order of the “trypsin fragments” in the original polypeptide chain must now be determined. Another sample of the intact polypeptide is cleaved into fragments using a different enzyme or reagent, one that cleaves peptide bonds at points other than those cleaved by trypsin. For example, cyanogen bromide cleaves only those peptide bonds in which the carbonyl group is contributed by Met. The fragments resulting from this second procedure are then separated and sequenced as before.

The amino acid sequences of each fragment obtained by the two cleavage procedures are examined, with the objective of finding peptides from the second procedure whose sequences establish continuity, because of overlaps, between the fragments obtained by the first cleavage procedure. Overlapping peptides obtained from the second fragmentation yield the correct order of the peptide fragments produced in the first. If the amino-terminal amino acid has been identified before the original cleavage of the protein, this information can be used to establish which fragment is derived from the amino terminus. The two sets of fragments can be compared for possible errors in determining the amino acid sequence of each fragment. If the second cleavage procedure fails to establish continuity between all peptides from the first cleavage, a third or even a fourth cleavage method must be used to obtain a set of peptides that can provide the necessary overlap(s).

***Locating Disulfide Bonds*** If the primary structure includes disulfide bonds, their locations are determined in an additional step after sequencing is completed. A sample of the protein is again cleaved with a reagent such as trypsin, this time without first breaking the disulfide bonds. The resulting peptides are separated by electrophoresis and compared with the original set of peptides generated by trypsin. For each disulfide bond, two of the original peptides will be missing and a new, larger peptide will appear. The two missing peptides represent the regions of the intact polypeptide that are linked by the disulfide bond.



**FIGURE**: **Cleaving proteins and sequencing and ordering the peptide fragments.** First, the amino acid composition and aminoterminal residue of an intact sample are determined. Then any disulfide bonds are broken before fragmenting so that sequencing can proceed efficiently. In this example, there are only two Cys (C) residues and thus only one possibility for location of the disulfide bond. In polypeptides with three or more Cys residues, the position of disulfide bonds can be determined as described in the text.

**POST TRANSLATIONAL MODIFICATIONS OF PROTEINS**

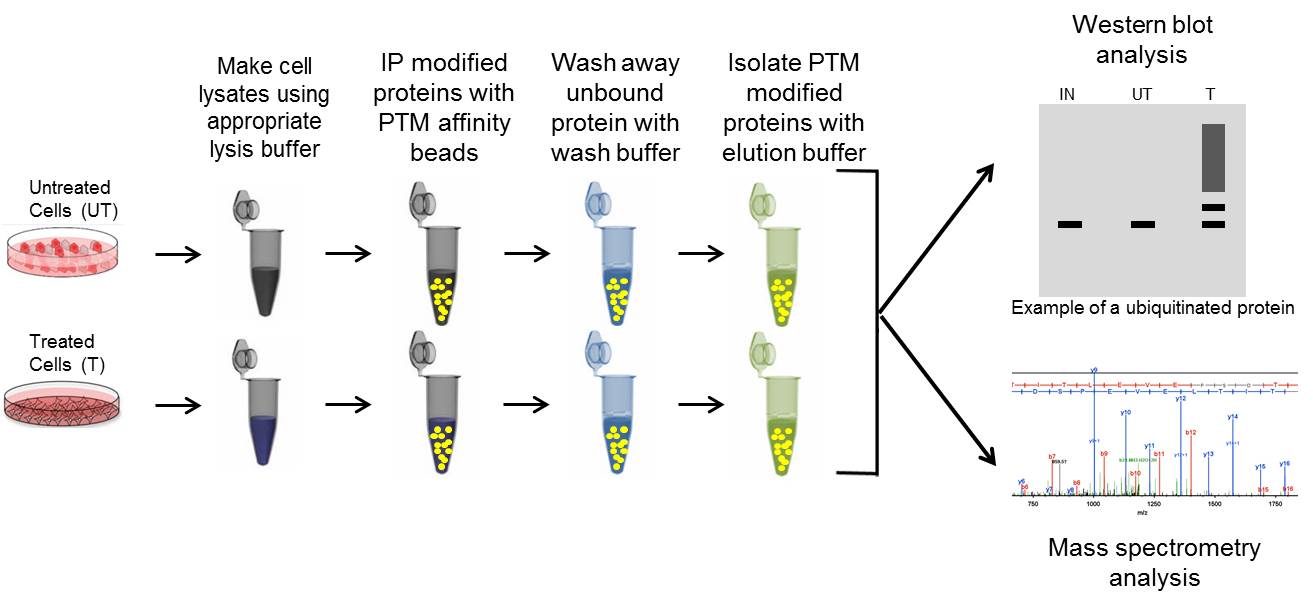
Post-translational modifications (PTMs) like phosphorylation, acetylation, SUMOylation, ubiquitination, and others account for the vast increase in proteome complexity. For any given protein, a variety of PTMs offer a way to facilitate rapid cellular changes by altering the structure and function of the protein. PTMs play a critical role in signal transduction, protein stability and turnover, protein-protein recognition and interaction, as well as spatial localization.

Due to the importance of PTMs in basic biology as well as in disease pathogenesis, there is a significant interest in identifying regulatory PTM mechanisms for every protein of interest (POI). Investigations of PTM modified proteins normally require an enrichment step as these modified proteins are relatively low in abundance. The majority of PTM detection methods have been developed in combination with enrichment strategies to provide the best opportunity to identify, validate, and study the function of regulatory PTMs for a protein. Choosing the appropriate assay, method, and/or protocol will depend on the question you are trying to answer; ultimately, it is recommended to use a combination of these methods for identification, validation, and mechanistic characterization of a PTM for a protein.

**Immunoprecipitation based techniques**

Immunoprecipitation (IP) is a core technique that is utilized in several different PTM detection assays. Importantly, IP allows researchers to enrich for specific, low-abundance PTM modifications on a target POI. Enrichment is achieved through affinity-based purification technology.  Antibodies (or specific binding molecules) attached to a solid support matrix (such as agarose resin) bind to the POI or PTM, while non-targeted proteins in the complex lysate are not captured and removed through wash steps. The enriched proteins are then removed off of the support matrix using elution buffers and isolated in a concentrated volume. The isolated population is analyzed by downstream methods like western blot or by mass spectrometry to determine if a POI is post-translationally modified.

IP is a critical step in the majority of PTM detection techniques; thus, having optimized, high-quality IP reagents will provide the best likelihood of obtaining meaningful results. These reagents include the appropriate lysis buffer system, antibody conjugated to affinity beads (to minimize antibody contamination), control beads, wash buffer, and elution buffer.



**Figure:Genral workflow for immunoprecipitating PTM modified proteins.**Utilization of optimized IP systems are critical for detection of low abundance PTMs for a target POI.

**Protein antibody specific IP: western blot analysis**

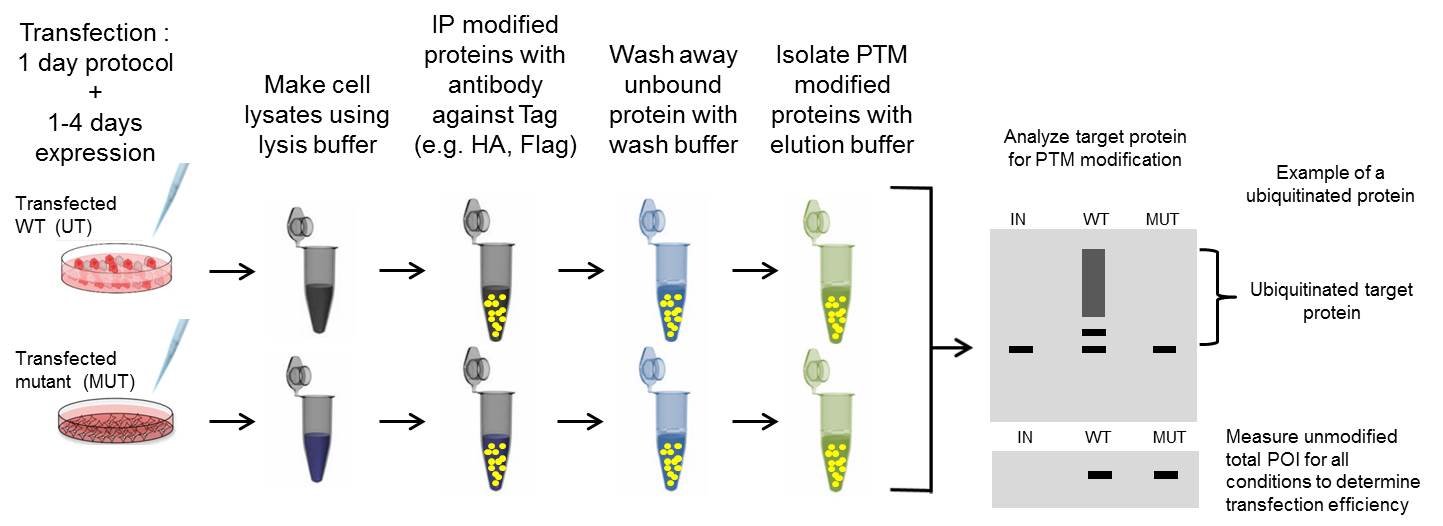
Protein specific IP utilizes an antibody against a POI to immunoprecipitate potentially all species of that protein. The enriched proteins are then separated by SDS-PAGE, transferred to a PVDF membrane, and analyzed *via* western blot, by probing with a target PTM antibody. Target protein antibody immunoprecipitation has traditionally been the first approach used to determine if a POI is modified by a particular PTM, possibly due to the investigator's expertise with the POI specific antibody for western blot applications.

**PTM antibody specific IP: western blot analysis**

PTM specific IP utilizes an antibody against a target PTM to immunoprecipitate potentially all proteins that have been modified by that PTM (dependent on the quality of the PTM affinity reagent). The enriched post-translationally modified population is then separated by SDS-PAGE analysis, transferred to a PVDF membrane, and analyzed *via* western blot, by probing with an antibody targeting a specific POI.

**Overexpression IP: western blot analysis**

Overexpression IP is a well-established system where a plasmid of a tagged version of a POI is transfected into cells.  This usually results in high expression levels of the tagged POI, which can be immunoprecipitated with a well-characterized antibody against the tag. The increased expression and optimized antibody improves the chance of identifying the PTM modified POI. Overexpression IP assays are normally analyzed by western blot.



**Figure: General steps in an overexpression IP assay to detect PTM modified target proteins.**Transfection protocols and efficiencies varies significantly by methodology and an appropriate transfection method should be identified prior to PTM investigation.

**Mass spectrometry after IP enrichment**

Nearly all Mass spectrometry analysis of PTMs are more effective when combined with IP strategies discussed above to enrich for a protein or PTM of interest. Rather than performing a western blot to determine if a specific protein is modified, the sample is analyzed using a mass spectrometer.  The investigator can identify a spectrum of proteins modified by a PTM using bottom-up peptide-based PTM proteomics. Mass spectrometry based PTM identification is a powerful tool, and is a valuable complement to conventional western blot PTM identification; especially for site specific identification.

***in vitro*: Biochemical Assay**

Biochemical assays utilize purified or *in vitro* translated versions of a target protein to determine if it can be modified by a specific PTM. The purified protein is added to a test tube with specific enzymes (e.g. E1, E2, E3 ubiquitin ligase) and the appropriate substrate (e.g. ubiquitin), co-factors, and energy sources. After incubation, the sample is then analyzed by western blot analysis.

**Immunofluorescence for global PTM**

Immunofluorescence techniques may be a fruitful approach to studying global changes in a PTM profile in tissues or cells.  In particular identifying global and spatial changes in response to drug treament (e.g. HDAC inhibitors) or genetic knockout is achievable with this method. The identification of target specific PTM modifications are not possible by this method. However, this approach may be applicable as a biological readout; thus, may be a useful tool as an indicator for development or disease progression

**Proximity Ligation Assay**

Proximity ligation assay (PLA) is a novel immunoassay technology that can be used to study protein interactions and PTMs. PLA is unique in its ability to identify PTMs of a specific POI in fixed tissues and cells. The technology utilizes an amplification process, which allows detection of low abundance molecules and modifications.

The principle of PLA-PTM works by utilizing two antibodies (preferably raised in different species) one against a PTM of interest and one for the target protein of interest. The initial steps in the procedure are similar to standard immunofluorescence staining where the primary antibodies bind the protein of interest and the PTM of interest.

Secondary antibodies are added that recognize the two primary antibodies.  The difference with PLA is that these secondary antibodies have short DNA strands covalently attached to them (these antibody-DNA complexes are called PLA probes). If the two PLA probes are in close enough proximity, presumably because the two antibodies are binding to a single protein that has been PTM modified, they will form circularized DNA.

The next step requires the addition of polymerase, and amplification of circularized DNA. Once amplification is complete, fluorescently labeled complementary DNA probes are added. Due to the significant DNA amplification, which can be up to several hundredfold, the fluorescent signal from very few molecules will be visible by microscopy.