**Title: Chemical coupling to the N-termini of proteins**

Polypeptides are synthesized naturally in cells by ribosomes in a process where amino acids are covalently coupled to each other through a covalent bond. The bond between amino acids occurs between the amine attached to the alpha carbon of one amino acid and the carboxylic acid on the alpha carbon of another amino acid, to form an amide, referred to as the peptide bond. This bonding can be repeated and gives rise to a polymer (polypeptide/protein) that has an amine on one end of the chain (N-terminus) and a carboxylic acid on the opposite end (C-terminus). In some proteins, the ends can be modified by for example acetylation of the N-terminus. However, the N-terminal to C-terminal directionality (and nomenclature) persists despite such modifications. There are 20 canonical amino acids that are genetically encoded. There are additional amino acids that can be incorporated in naturally occurring proteins (e.g., selenocysteine), as well as a wide range of modifications to proteins that are well known (e.g., phosphorylation, acetylation, disulfide bonds, methylation, glycosylation). Methods have also been developed to chemically synthesize polypeptides. Such synthetic polypeptides can be chemically or functionally indistinguishable from naturally produced molecules. However, synthetic polypeptides can also be made using an almost endless range of amino acids that do not occur naturally. In particular, amino acids, with the exception of proline, are defined by a so-called R group that is attached to the alpha carbon of the molecule. It is variations in the R group that define different types of amino acids, and in synthetic peptides, there is a nearly endless variation of non-natural R groups that can be used. Synthetic polypeptides can also be produced using isomers of amino acids, as well as amino acids that have isotopic ratios of one or more atoms that differ from the naturally occurring ratios. Proteins can also be produced using methods that effectively a combination of natural and chemical methods, such as *in vitro* translation. *In vitro* translation utilizes a naturally derived component, such as a cell extract that contains some or all of the translation machinery in combination with chemicals and/or proteins that facilitate or enable the synthesis of proteins from mRNA molecules introduced into the system (*ex vivo*). mRNA molecules used for *in vitro* translation can be isolated from living organisms or cells. These mRNA molecules can also be produced by *in vitro* transcription, using natural or synthetic DNA templates. By introducing tRNAs charged with non-natural amino acids, proteins with non-natural amino acids can be synthesized. For our purposes, the terms polypeptide and protein are equivalent and refer to any polymer of amino acids that is at least two amino acids long, that are connected via a peptide bond, no matter how they are synthesized and whether or not any of the monomers in the polymer occur naturally. The term peptide refers to a subset of proteins, specifically proteins or parts of proteins that are relatively short (<50 amino acids long). By protein, we also mean an intact (full-length) protein or fragments or portions of a protein.

Outside of the natural, co-translational, post-translational, co-synthetic, or post-synthetic modifications that can occur, there are many useful reasons to chemically modify polypeptides. This includes attaching various sorts of labels or tags that allow the presence and/or location of a protein to be determined. Such labels include fluorophores, chromogenic compounds, molecules detectable by electron paramagnetic resonance, enzymes that can produce detectable products (e.g., luciferase and horseradish peroxidase), etc. Modifications are also used to probe structure, activity and functions of proteins using for example Förster Resonance Energy Transfer (FRET). Modifications are also frequently used to introduce tags or handles that can be used to capture molecules (Massa and Devoogdt, 2019). For example, biotinylation of a protein is often performed so that the protein can be captured by avidin, or streptavidin connected to another molecule or structure. One central theme in many modification methods is that the modification does not destroy, reduce or significantly change the structure, function or activity of a protein. Hence, the attachment of a fluorophore to an antibody should not reduce the ability of that antibody to bind to its ligand or reduce the ability of secondary antibodies to bind to the first. Enzymes immobilized on a surface or in a gel using a biotin-streptavidin linkage should not have compromised activity because of the modification used to facilitate the coupling. As a consequence, most methods developed for the modification of proteins seek to retain structure, function, and/or activity of the modified protein. There is a much smaller group of methods for more specialized applications that are to varying degrees destructive, in which structure, function, or activity is compromised or lost.

One particularly difficult form of protein modification is the coupling of molecules to one of the ends, N-terminal or C-terminal, of a protein, without altering or modifying another part or parts of the protein. There are many useful reasons to modify one or both ends of a protein. Specific end modification allows for the concatenation of proteins and the post-translational assembly of smaller proteins into larger ones, and the construction of proteins with novel structure/function/activity by for example fusing two proteins with different functions. Further, in general, the modification of the end of a protein is less invasive than for example modifying all the primary amine or carboxylic acid-containing molecules. A number of approaches have been developed to couple molecules specifically to the N-terminus of proteins (Koniev and Wagner, 2015). These include exploiting small differences in the pKa of the N-terminal amine and the primary amine on the lysine side change, as well as chemistries that depend on specific side chains. An example of the latter is that there are chemistries by which N-terminal cysteines can be specifically modified. However, methods that require a specific amino acid or peptide sequence that is part of the target protein to begin with significantly limit the utility of the method to a small subset of proteins or to proteins that are modified by genetic engineering. Exploiting differences in the pKa of amines is limited because “controlled reaction conditions and, the preference for terminal amino groups achieved by control of pH is rather limited, mainly owing to the fickleness of the pKa of the amino group depending on the microenvironment and reaction conditions” proteins (Koniev and Wagner, 2015). Other approaches have been developed, but chemical coupling to the N-terminal amine of proteins remains an important and unsolved problem.

Further, such methods are not useful in complex mixtures where there are many amines in addition to the N-terminal amine and lysines that are part of proteins.

Another application for end-modifying polypeptides arises from new methods to sequence polypeptides and proteins. Recent work of Decker and coworkers (Brinkerhoff et al., 2021), and from Oxford Nanopore (presented online by Dr. Clive Brown at London is calling, 2022, see attached screenshot from his presentation), proposes a new approach for sequencing proteins using protein (or non-protein) nanopores. This approach is an extension of the nanopore-based nucleic acid sequencing and nanopore-based polymer characterization to proteins.However, one problem with proteins is the heterogeneity in chemical properties (charge, hydrophobicity, polarity, etc), both overall between different individual molecules and along the length of individual molecules (further complicated by post-translational modifications, such as phosphorylation, that can add charge). Nanopore sequencing of nucleic acids was originally implemented using the relatively uniform (uniformly periodic) negative charge on the backbone of these molecules in aqueous or other polar solvents, which arises from the regular periodicity of phosphates on the backbone, to electrically move the molecules through protein pores. By measuring electrical parameters, such as current fluctuations, as a molecule passes through the pore the sequence of bases in the molecules can be determined. Initially, the ability to determine the sequence of molecules was barely better than random, but that gradually improved and is now well above 90% (often even better than 95%). An important development in improving nanopore-based nucleic acid sequencing was the addition of a molecular motor to the process. This motor is attached near the opening of a nanopore. A nucleic acid is captured by the motor, and the motor facilitates the translocation of the molecule through the pore at a controlled speed. Some motors of this type belong to a class of molecules known as helicases (and their various subtypes). The motor in some cases utilizes ATP to move the molecule across the membrane. In other instances, the motor does not utilize ATP, but it nonetheless controls the speed of electrically based conformational changes in the proteins during translocation of a nucleic acid. This can be viewed as a dead motor, one without its own power source, but it still facilitates nanopore sequencing by controlling translocation speed. As such, the translocation of nucleic acids through a nanopore depends on an electrical potential, or a combination of an electrical potential with a molecular motor. However, even if the electrical potential is not required for translocating a molecule through a pore, an electrical potential can increase the speed at which a molecule to be sequenced finds/binds to/locates a pore, or otherwise improve performance.

In comparison to nucleic acids, the charge distribution along a protein is highly heterogenous. There is no periodic charge associated with the backbone of a polypeptide. Instead, charge comes from amino acids such as, for example, glutamic acid and lysine, which are typically non-uniformly distributed along the length of the polypeptide. Although there can be stretches or segments of specific proteins where the charge is uniform. The amino acid histidine has a pKa of ~6. As such the charge of histidine is sensitive to pH variations in the range 5-7 (and less so in the ranges 4-5 and7-8), and thus can contribute varying amounts of charge depending on the experimental conditions (which are often in the range pH 6-8, or pH 5-10). The net charge of a protein as a whole can be positive, negative or neutral, depending on the combination of positively and negatively charged present in the molecule, and the structure of the molecule), at the relevant pH, while linear nucleic acids are negatively charged at relevant pHs. Although folding of nucleic acids into 3-dimensional structures can introduce some non-uniformity to the charge distribution. This “charge problem”, with respect to the distribution of charge in proteins, seriously complicates the use of an electrical potential to translocate proteins. Since for any given polarity and potential, some proteins will move toward and through the pore, move away from a pore, or not move appreciably in either direction. That is exploited as a basis for separation in isoelectric focusing. However, it is a problem for the use of nanopores to sequence proteins as well as other analytical or preparative methods that employ electrical potentials (such as electrophoresis).

One solution to this problem has been to bind a charged amphiphile, commonly, sodium dodecyl sulfate (SDS), to one or more (different types of) proteins in a sample to be analyzed. SDS is negatively charged at pH’s above ~2. SDS binds to most proteins at an average density of 1.4 g SDS per gram of protein. There can be many fold variations in this binding density of SDS to proteins, hydrophobic proteins (or parts of proteins) tend to bind more, and hydrophilic proteins (or parts of proteins) tend to bind less. But the binding is sufficiently general and uniform that an SDS bound protein subjected to an electric field will migrate through a polyacrylamide gel in a fashion that is approximately proportional to its molecular weight. As a result, SDS-polyacrylamide gel electrophoresis is one of the most widely used and important analytical methods available for characterizing proteins. It should be noted that SDS also tends to disrupt the 3D structure of proteins (denature) and make them more linear, which is important to the correlation between the electrophoretic mobility of proteins and their molecular weight. Thus, SDS has at least two functions that are important for protein analysis, the disruption of the 3D structure of proteins to make them behave more like a linear polymer, i.e., closer to a polymer in a theta solvent or a good solvent (well-known terms from colloidal science) and conferring a relatively uniform charge density that is correlated with the mass of the molecule.

In the context of protein sequencing with nanopores, a problem with SDS and other amphiphiles, is that the membranes used to suspend the pores, the pores themselves, and any proteins used to facilitate translocation through the pore, or any protein in the solution with the amphiphile whose activity is required for the optimal sequencing, will tend to bind SDS. Thus, SDS can disrupt the membranes (this includes non-biological membranes, such as membranes made from organic polymers), making them structurally unable to sustain the voltage needed, as well as disrupt/denature/inactivate pore and other proteins whose structure is required for the process. One solution to this limitation of SDS is to use a membrane and pore-based system that is devoid of proteins or membrane components that are disrupted by amphiphiles. Non-protein nanopores for such applications are under active development and some successes have been reported, although nucleic acid or protein sequencing based on such pores is not yet sufficiently reproducible to be practical or commercially viable/available. Nanopore-based sequencing of SDS (or another amphiphile)-protein complexes would depend in part on the strength and specificity of the SDS binding to the protein, and the way in which SDS binding influenced the electrical signal as the protein (with or without the SDS, as the SDS may be stripped off along the way) is translocated through the pore.

Another approach for ameliorating or overcoming the “charge problem” in proteins, for the purpose of nanopore sequencing, is to attach a charged polymeric leader to one and/or both ends of the proteins (Figure 1). Both Decker and colleagues and Oxford Nanopore have used oligonucleotides coupled to the end of proteins to facilitate protein translocation through pores. This leader (whether oligonucleotide or another polymer) must be negatively charged in excess of the charge on the peptide, such that the net charge on the leader plus the polypeptide is negative. When the net charge on a polypeptide is negative without the leader, a leader is still needed to cause the molecule to enter a pore from one end of the polypeptides. Such a leader should have greater than negative charges in excess of the polypeptide. Further, the leader should have a charge density that exceeds that of any part of the polypeptide by at least 0.01 charge units per nanometer. As such, when an electrical potential is applied to the molecule the leader effectively controls the motion of the peptide. The electromotive force on the leader then causes the leader to effectively pull the polypeptide in the direction of the positive side of the potential applied. In the presence of a nanopore of appropriate dimensions, the leader will effectively pull the polypeptide through the pore. When doing so, there are electrical signals/signatures similar to when a nucleic acid is translocated, when the polypeptide traverses the pore. That signature depends on the sequence of amino acids in the polypeptides, and by using machine learning methods algorithms, artificial intelligence and other computational tools, that identify amino acids in a polypeptide chain can be trained and used to predict sequences.

The leader can also be a positively charged polymer, in which case the positive and negative charges used in the preceding paragraph are inverted. The leader must be sufficiently positive to make the net charge of the molecule positive, etc.

The leader can also be a charged non-polymer, or a non-polymer mixed with monomers of a polymer that are positively charged, negatively charged, or not charged (neutral). Polymer refers to repeating units of a molecule (monomer) connected together in a chain (that may or may not have branches). But chain-like molecules can also be constructed from molecules in a way that they are not repeating, but still in many respects behave like a polymer. As such, other chain-like molecules can also be used to serve the same purpose as a polymer leader. Water can be used as a solvent, although non-aqueous solvents that are compatible with the molecules being examined can also be used.

When using a leader, the performance will be better if the leader is attached to the N-terminus or C-terminus of the polypeptide. The attachment of a leader to the middle of a polypeptide will cause the polypeptide to fold and produce a structure where the leader is effectively trying to pull two adjacent polypeptide chains through the pore. The size of two adjacent chains will cause a steric problem and may prevent the polypeptide from being translocated, thereby clogging the pore. If the double-chained polypeptide can still be translocated, the electrical signal will be complicated by the fact that there will be two strands with different sequences translocating at the same time. For a non-end coupling, the closer the coupling point is to one end the shorter this double-chained region will be. Similarly, the coupling of a polymeric leader should be by one end of the leader or the other. The superiority of end attachment to polypeptides of a leader leads to a need for end-specific attachment chemistries. That is chemistries that enable the attachment of the end of a polymer, such as an oligonucleotide, to the C-terminal or N-terminal end of a polypeptide. Such chemistries involve the specific coupling of leaders to the free amine at the N-terminus of a protein, the free carboxylic acid at the C-terminus of a protein, the acetyl group on an N-terminally acetylated protein, or any other modification to either the C-terminus or N-terminus of a protein.

There is also a benefit to attaching two different leaders to a molecule, where one leader has a net negative charge and has a net positive charge. When the net charge and charge density on one end of the molecule is larger than the other, the molecule will move in the direction of the larger net charge. While doing so, the charge on the opposite (trailing) side of the molecule will pull in the opposite direction and tend to straighten the molecule. That will reduce the entropy of the molecule, which for longer molecules can reduce entanglement and problems such as knotting of large molecules during transit through the pore.

In some organisms, or in some samples, a significant fraction of N-termini of proteins is modified. Acetylation is a common modification, and others include methylation, myristoylation, propionylation, and palmitoylation group. End modification with a leader molecule, and end modification in general, can be performed with chemistry specific for one or more of these modifications. Such a leader attachment will complement the attachment of leaders to free N-termini, thus producing a more complete representation of proteins in the sample (when sequencing). When the leader attachment chemistry is specific for an end modification, it also allows the subset of end-modified protein molecules to be sequenced.

It is also possible to deblock or remove N-terminal modifications that prevent access to the primary amine at the N-terminus of a protein.

Leader polymers can also be used when characterizing other types of polymers (in addition to proteins), including but not limited to DNA, RNA, nucleic acids, polyethylene glycol, polyvinyl alcohol, and various block co-polymers.

Another aspect of a leader is the distribution of charge along the length of the leader. A leader molecule with a more negative charge toward the free end, away from where it is connected to the molecule being examined, will cause the leader to re-orient in response to an electric field. A leader molecule with a higher charge density toward the free end will more quickly engage with (begin to translocate through) a pore used for sequencing or polymer characterization. Taking DNA oligonucleotides as an example. Synthetic oligonucleotides are often produced with hydroxyl groups at the 5’ and 3’ ends, and thus have uncharged ribose units at either end. The 5’ of these molecules is commonly modified with a phosphate, which facilitates enzymatic reactions such as ligation. In that situation, there is a negative charge approaching -2 on the 5’ end at neutral pH (pKa values for the 5’ phosphate are ~0.9 and ~6.1). The negative backbone charge is spaced periodically (one negative charge on the phosphate between each base). The 5’ end of oligonucleotides can also be produced with 2 and 3 phosphates. In the case of the diphosphate base, there will be 3 negative charges and in the case of triphosphate base there will be 4 negative charges. Thus, an oligonucleotide leader with a 5’ triphosphate will enter a pore more readily than one with two, one or no phosphates, all other things being equal. Charge can also be added to bases in an oligonucleotide by modification of the 2’ position of the ribose with a charged moiety, such as, for example, carboxyl groups, phosphates or sulphates. This allows an oligonucleotide to have 4 or 5 negative charges on the 5’ monomer, followed by 1-40 residues with 2-3 negative charges, followed by 1-40 residues with 1-2 negative charges. The charge characteristics of a leader can be described by the magnitude of the end charge, the charge density along the leader in terms of units of charger per monomer or a unit of length (such as charge per nanometer), and the charge gradient of along the leader away from the free end (relative to the free end). The gradient of charge density can be linear, exponential, or logarithmic.ATP has pKa values of 0.9, 1.5, 2.3, and 7.7; ADP has pKa values of 0.9, 2.8, and 6.8; and the alkyl phosphate has pKa values of 1.9 and 6.7. The methylphosphonate modification can be used to make the charge along the oligonucleotide backbone close to zero.

While DNA, other nucleic acids, or related molecules are convenient for many reasons, the leader polymer can also be another polymer such as polyacrylic acid, polystyrene sulphonate, or others.

Leader properties can also be modified and used in mixtures, such that molecules with certain leaders are sequenced prior to others in the same mixture. That also allows for multiplexing.

A method for attaching an oligonucleotide or other polymer to the N-terminus of a protein would be useful. Chemical coupling reactions that are specific for primary amines, such as the N-terminal amine of a protein, are well-known and widely used (Koniev and Wagner, 2015). However, for naturally occurring proteins as well as synthetic or non-natural proteins, the presence of lysines (or any other non-N-terminal primary amine) complicates the use of existing chemistries because coupling occurs not only at the N-terminus but also at the free amine (epsilon amine) on lysines. In nanopore sequencing, this would interfere with the translocation of a polypeptide chain by producing bulky attachments at the lysine that would not translocate, translocate more slowly, or signals that are difficult or not possible to interpret.

Further, using current methods of coupling molecules to amines on proteins requires that a sample is presented in a form that is devoid of competing amines outside of the protein. That means that buffers, solvents, or solutes that have amines must be avoided and/or removed before the modification chemistry is performed. It also means that samples that originate from complex fluids, such as cell lysates, serum, urine, or lymph should be purified (primary amines removed) before the coupling chemistry is performed.

The characterization of proteins is a cornerstone of modern biomedical research. The most basic characteristic of a protein is its identity, which in the simplest terms is defined by which gene it was derived. That is to say, all proteins derived from the same gene have the same identity. However, because of alternative splicing, co-translational and post-translational modifications, including non-enzymatic modifications, there is a large diversity of structurally unique versions of proteins from any given gene. The structurally unique version of proteins from any given gene have been given the name proteoforms (Smith et al., 2013). Thus, the proteome of an organism or biological system (e.g., cultured cells, organ, organism, a tissue explant, the products of an in vitro translation, etc.) can be viewed as the sum of all proteoforms in that organism or system at any given point in time. Alternatively, given that proteoforms vary over time in an organism or biological system, the proteome can be considered the sum all the proteoforms that are present in a population over time (longer than the lifetime of the organism, and under all conditions to which the organism might be exposed). The size of the proteome is not well-defined, but a million is likely on the very low side and it is more likely much larger than billions (depending on how abundance is taken into account). As such, an exhaustive accounting of the proteome is at best a daunting task, and likely will not be practical for many decades if ever. The size of the proteome expands even further when one takes into account proteins that are made synthetically, and which can have non-natural sidechains or modifications. Modern proteomic methods seek to determine the protein composition (proteome) of an organism of some biological system. The quantities determined can be absolute, but are more typically relative to other proteins in the sample. The relative concentrations can be determined by determining the ratio of one protein to some reference protein, where the reference protein is part of the original sample or is added to the sample for the purpose of quantitation.

The core defining feature of a protein is the primary sequence of amino acids, which is determined by the sequence of bases in gene from which it was derived. Analytically, mRNAs are sometimes used as a surrogate for proteins present in a living system (e.g., animal, plant, fungi, etc) or some part of a living system (e.g., cell, tissue, etc). This is sometimes referred to as RNASeq and is typically performed by making cDNA copies of mRNA molecules and then sequencing those cDNAs. In principle the idea is to determine the number or amount of each type of mRNA, if that can be done directly. When conversion to cDNA is employed, compositional biases can arise. There is often a significant correlation between mRNA levels and amounts of certain proteins. mRNA levels can capture protein variants that arise from alternative splicing, but do not capture any of the co-translational or post-translational modifications of proteins. Further, different proteins turn over at different rates, which severely limits the utility of using mRNA as a surrogate for protein levels.

Another common approach to determine the identity of proteins is the use of affinity reagents, such as antibodies, aptamers (including aptamers with modified bases), affibodies, imprinted polymers (molecular imprinting), etc. Affinity reagents can in some situations distinguish between splice variants. However, it is very difficult to develop affinity reagents for splice variants and there are many splice variants of proteins for which affinity reagents do not exist. The difficulty here is tracking the combinations of junctions between parts of a protein that arise from different exons in the gene that codes for the protein. Affinity reagents can be developed against many co-translationally or post-translationally modified proteins. Although the number of variations of a protein is so large, currently a very small fraction (<10%, <1%, <0.1% or <0.01%) of all protein variations have been identified, that doing so is enormously time-consuming, costly and is likely not practically possible. Current affinity-based methods also by their nature require relatively large amounts (much more than single molecules) of material.

Mass-spectrometry has emerged as an important method for identifying proteins and their variants in a variety of samples. This, for example, includes identifying a protein extracted from a band on polyacrylamide, the product(s) of an enzymatic modification of a protein (such as phosphorylation), and the composition of proteins in a sample with many different proteins. While it is possible to measure masses for whole proteins in some situations, doing so is technically difficult and at present has limited utility. Instead, using mass-spectrometry to characterize proteins is based on fragmenting proteins chemically or enzymatically, and then determining the masses of the resulting peptides. Those masses are then compared to a library of previously studied proteins, and/or predicted peptide masses based on protein sequences, to establish the presence of and approximate amounts of peptides in the sample. This peptide profile/composition is then used to establish which proteins were present in the sample. However, mass-spectrometry is a bulk method that has relatively poor sensitivity, and variants of proteins in low abundance may not be detected. Further, because the proteins are fragmented not all splice variants are detected. For complex samples, such as those derived from whole cells or tissues, the large number of peptides produced requires that the samples are fractionated prior to mass-spectrometry. Liquid chromatography of various types is often used in combination with mass-spectrometry (LC-MS).

To achieve a full and accurate accounting of the proteome, it is likely that single-molecule characterization of whole or intact protein molecules will be required. It is against this background of analytical methods that nanopore-based sequencing of proteins is being developed. Nanopores is the current technology that has the potential to achieve the deepest/largest/most complete coverage of the proteome.

The order of amino acids in a polypeptide chain can be determined using the Edman method. The details of this method are well known. But briefly, a protein with a free primary amine at the N-terminus, is reacted with phenylisothiocyante (PITC, a.k.a. the Edman reagent) under mildly basic conditions (often pH 8-10). The sample is then exposed to low pH (often using TFA or HCl) and heat (often ~50 °C), which releases a thiazolinone carrying the sidechain of the N-terminal amino acid. The thiazolinone is separated from the remaining peptide sample and converted into the more stable thiohydantoin derivative of the amino acid in question by further incubation at low pH and under heat. The thiohydantoin derivative is then identified by liquid chromatography, mass spectrometry, or other analytical method. The identification of the thiohydantoin derivative in question, and thus the amino acid that was at the amino terminus, can be facilitated using appropriate standards for the thiohydantoin derivatives of the various amino acids.

For primary sequence analysis a protein sample should be as pure as possible. That is to say that to the greatest extent possible it must be composed of a single protein (i.e., a protein with a particular amino acid sequence) and not contain any molecules that interfere with the Edman degradation chemistry or subsequent analytical methods. The more pure the protein in the sample is the more sequence can be obtained for any given situation. The presence of other proteins will produce thiohydantoin derivatives that effectively act as noise and make identification of the amino acid in the protein of interest more difficult. Additional proteins will shorten the length of amino acid sequence that can reliably be obtained from the protein of interest. A second or contaminating protein at sufficiently high concentrations (in some cases approaching 50% of the protein) makes unique identification of the first amino acid in the protein of interest not possible. The greater the quantity and number of proteins in addition to the protein of interest, the more complicated interpretation of the results and in general the less information about the amino acid sequence of the protein of interest can be obtained. Thus, sample purity is emphasized in sequencing by Edman degradation.

The Edman degradation can also be used to determine N-terminal amino acid composition of a sample with one or more proteins. For this application, more than one protein may be present in the sample. However, the sample remaining after the N-terminal amino acid is generally not of interest. For both sequence determination and for amino acid composition determination, non-protein molecules containing primary amines are avoided, since they compete with the proteins for the Edman reagent and create reaction products that complicate subsequent analysis of the sample.

The original Edman method is generally described as a solution-based method, although it involves drying the sample into a film at the bottom of a test tube at several points. The details of the chemistry are well known, and some examples are given in Chapter 6 of Shivley (1986).

Since its initial development, the Edman degradation has undergone many improvements. Some of these improvements were relatively speaking incremental and involved things like reduction of sample volume, improvement of solvent and reagent quality used, etc. However, there were two major changes in the method. One was the development of a solid-state method and the associated automation of the chemistry, initially using the so-called spinning cup method (Chapter 7 of Shivley, 1986). This involved using a matrix for immobilizing proteins (which is what made the method “solid state”). Such immobilization permitted the implementation of a spinning cup design for exchanging liquids, where liquids could be passed through a matrix on which the sample was immobilized. Later a vapor phase approach was introduced, in which some steps of the process utilized treatment with a vapor, which reduced sample washout and significantly improved the sensitivity (Hunkapiller et al., 1981). The fluidic design also changed from the spinning cup to a cartridge through which liquids were pumped (as opposed to centrifuged). Protein sequencing by Edman degradation has continued to improve and is now highly automated and widely available. This chemistry and technology are well known.

The Edman degradation is limited in how many consecutive amino acids, the length of a protein or part of a protein, that can be sequenced. Early in its development sequences were often limited to 10-20 amino acids. As the method has improved, the lengths have increased such that 50 amino acids is not uncommon and even longer has been achieved. However, to sequence entire proteins, which are typically >100 amino acids long, it is necessary to break the protein into parts. Thus, a protein can be fragmented using for example a protease or a chemical method (e.g., CNBr), and the peptides produced from the fragmentation can be isolated and sequenced individually. If all the peptides produced from a protein are captured, then the entire sequence can be represented in the data. However, the order in which the peptides are connected is lost. To obtain that order, the same protein can be fragmented using a different protease or chemical method that breaks the protein in different places than the first, the peptides are again isolated, and sequences of individual peptides are determined. Some of these peptides may overlap the junctions between the peptides from the first fragmentation and can therefore be used to order those peptides by so-called tiling. The second set of peptides can also be used to fill in gaps in the sequence that arise from peptides that may have been missing from the first fragmentation. If needed, this process can be repeated again until a sufficiently complete sequence is achieved.

The Edman degradation can be used to determine the presence of some modified amino acids, such as phosphoserine.

While the original Edman degradation method was performed with PITC, a number of Edman-like molecules have since been developed. In particular the 4 positions of the phenyl ring in PITC can be modified in many ways, while still retaining the ability to perform an N-terminal amino acid cleavage from a protein. One important way in which PITC has been modified is to make detection of the phenylthiohydantoin, of the amino acid release, easier to detect (for example fluorescein-5-isothiocyanate, FITC). Thereby improving the sensitivity of the method and reducing the amount of material required for sequencing. The collection of PITC and all of these related molecules are here referred to as Edman like reagents (ELRs).

Here we describe a method to covalently attach a molecule (make a modification) to the N-terminus of proteins, and not other primary amines in a sample. The approach combines the blocking of all primary amines (including the N-termini of proteins) using an Edman or Edman-like reagent, followed by the sacrifice (removal) of the first amino acid in the proteins by completing the cleavage step of the Edman reaction, to expose a primary amine. Modified amines that are not bonded to another amino acid through a peptide bond are effectively blocked (rendered unreactive to amine chemistry). The exposed primary amine of what initially was the second amino acid in the protein, now the only (or majority) primary amine in the sample, can then be coupled to any desired molecule through well-established amine chemistries. This approach is particularly useful because it can be used to couple molecules to the N-termini of proteins in samples that have multiple other primary amines, such as buffer components, metabolites, carbohydrates, or nucleic acids. One useful type of modification is an affinity tag or an oligonucleotide that can act as a handle or targeting molecule by which to isolate, purify, localize or otherwise control the spatiotemporal dynamics, of the molecule. This includes attaching an oligonucleotide that can be used to facilitate the analytical determination of the sequence of amino acids in the protein using nanopores.

The sample may be composed of at least one (a first) protein with a primary amine at N-terminus and at least one additional (+1) primary amine. The N-terminal amine may be naturally occurring or may have been formed by deblocking an acetylated or otherwise modified N-terminus. The +1 primary amine may be attached to the protein. The +1 primary amine can be a naturally occurring primary amine, such as the amine on lysine or the amine on a post-translational modification. A +1 amine attached to the protein may also be from a synthetic modification made to the protein, for example coupling a fluorophore or other molecule to the protein. The +1 amine may also be on a molecule separate from the protein, including on another protein (as the N-terminal amine, or another amine), a component of the buffer, a non-protein solute, a metabolite, an unidentified molecule, a contaminant, a naturally occurring molecule, a synthetic molecule, or any other molecule. The sample can be composed of:

1. One or more protein(s) with a primary amine at the N-terminus and a lysine with a primary amine anywhere in the sequence.
2. Two or more proteins with primary amines at their N-termini.
3. One or more proteins with a primary amine at the N-terminus in a solution that contains one or more non-protein molecules with a primary amine. The non-protein molecule can be a buffer component or other molecule that is added to the sample such as trishydroxyaminomethane or a molecule that was present in the sample as a result of how the sample was produced or isolated (e.g., thyroxine, epinephrine, phosphatidyl ethanolamines, urea, glutamine, to name just a few). The non-protein can also be an individual amino acid.
4. Two or more proteins with primary amines at their N-termini in a solution that contains one or more non-protein molecules with a primary amine.
5. Two or more proteins with primary amines at their N-terminus in a solution that contains two or more non-protein molecules with a primary amine.
6. Three or more proteins with primary amines at their N-terminus in a solution that contains one or more non-protein molecules with a primary amine.
7. Three or more proteins with primary amines at their N-terminus in a solution that contains two or more non-protein molecules with a primary amine.
8. Three or more proteins with primary amines at their N-terminus in a solution that contains three or more non-protein molecules with a primary amine.
9. A biological fluid such as urine, serum, lymph, cerebral spinal fluid, cell culture medium (eukaryotic or prokaryotic) before or after it has been used for culture, lysates or extracts from cells, tissues, plants, fungi, bacteria, animals, eukaryotes, prokaryotes, organ or tissue biopsies, tumors, cancers, and others.
10. A sample as in 9, with one or more (two or more, three or more) amine-containing molecules present in the sample was added during the processing of the sample.
11. A sample that has been derived by fraction or isolation from a sample as in 9 or 10, by chromatography, electrophoresis, centrifugation, dialysis, or another of the many well known fraction techniques. The degree of purity for the desired product is above 1%.
12. A sample as in 1-12, where all or a portion of the sample is immobilized on a solid support, such as a membrane, glass, silicon, metal, a bead a coated surface, or a material that is insoluble in the solvent used.

The sample is dissolved in a solvent that may contain buffers or other solutes that influence solubility or stability of the sample. The solvent may be an aqueous (water-based) solution or a non-aqueous liquid such as list A of solvents (see below)**,** or other well-known solvents. The sample solvent may also be a mixture of two or more aqueous and/or non-aqueous solutions. The sample can be dried and resuspended in a new solvent for coupling, or the sample can be used as is. The sample is subjected to the first steps of the Edman degradation using PITC or another Edman-like reagent (ELR, see list). The specific ELR is selected in part based on the solvent of the sample and the solubility of the ELR that solvent. Conversely, the solvent for the sample can also be selected such as to make it more appropriate for one ELR or another. Taking PITC as an example, the sample is reacted in EtOH:TEA:water:PITC 7:1:1:1 under mildly basic conditions (pH 8-10) at 50°C. The concentration of PITC is ~0.8 M, but the reaction is not highly sensitive to the exact concentration of PITC. In an aqueous solution sulfo-PITC or other ELRs that are more water-soluble than PITC can be used.

This initial Edman step produces a thiocarbamide modification of most of the primary amines in the sample. In the case of PITC, the thiocarbamide forms a bridge between the phenyl group from the PITC and the alpha-carbon of the N-terminal amino acid in the protein (Figure 2). For other ELRs such as sulfo-PITC a corresponding structure is formed.

The unreacted ELR is inactivated by the addition of a primary amine or primary amine containing molecule immobilized on a solid support. The solid support can form chromatographic matrix such that inactivation can be performed by passing the sample through a column (including spin column). The solid support can also be in the form of a bead that can be removed by gravity or centrifugal sedimentation. The bead can also be magnetic or paramagnetic and removed from the solution using a magnet. Soluble primary amines can also be used to quench unreacted ELR by adding a stochiometric amount of the molecule. The quenching reagent can also be titrated into the sample until all of the ELR is consumed or it is sufficiently low so as to not interfere with subsequent chemical processing. It is also possible to not quench or remove the ELR from the reaction if it will not interfere with subsequent coupling chemistry or subsequent uses of the sample.

The sample is then acidified and heated. The acidification may be performed by the treatment with hydrochloric acid or trifluoracetic acid to the sample, and heating is to 50°C for 2-3 minutes. In this step, the thiocarbamide at the N-termini of proteins and a reaction wherein a thiazolinone that carries the side chain from the first amino acid in the protein is released. That release exposes a primary amine from what was originally the second amino acid in the protein. Other reaction products of the PITC/ELR do not produce new primary amines.

In the conventional Edman degradation, the thiazolinone is at this stage separated from the mixture, and further incubated/treated to form a thiohydantoin. It is this thiohydantoin that is used to determine the identity of the amino acid at the relevant position in the protein by virtue of which side chain it carries. However, in the (N-1) method, the removal of the thiazolinone is not required. The sample can at this stage be purified if desired, or it can be used without further purification. The sample is brought to a neutral pH. At this point the sample is composed of one or more proteins with a primary amine at the N-terminus and one or more amines that have been blocked and rendered unreactive to amine chemistry by the reaction with the Edman reagent or ELR.

Further incubation of the sample at low pH and with heat causes the thiazolinone to rearrange into a thiohydantoin.

The newly created primary amine at the N-terminus of a protein can be used for any of the common amine coupling chemistries and used to attach a leader polymer or leader molecule (Figure 2). There are many chemistries for coupling molecules to primary amines that are well known. One particularly widely used chemistry for coupling molecules to primary amines is based on N-hydroxysuccinimide (NHS). NHS is commonly conjugated to fluorophores, tags, peptides, oligonucleotides and other functionalities, such that those molecules can be coupled to another molecule with a primary amine. If those functional groups are termed **X1**, then the reaction of **X1**-NHS with a primary amine on molecule **X2** (NH2- **X2)** produces **X1** coupled to **X2** via an amide bond. When **X1**-NHS is an oligonucleotide with the NHS on the 3’ or 5’ end of the oligo, and NH2-X2 is a protein with a primary amine at the N-terminus, then the reaction yields an oligonucleotide coupled via one end to the N-terminus of a protein via an amide bond.

The sample for some chemistries the sample can be used directly, in others some changes to the solvent or the solutes in the sample will facilitate the specific chemistry. For example, the sample can be subjected to chromatography as to remove interfering compounds, including so-called spin column chromatography (a form of size exclusion chromatography performed using centrifugal forces to speed the process). Dialysis can also be used to remove molecules below a certain threshold determined by the cutoff size of the dialysis membrane used. Salting in/out can also be used for purification. The sample can also be dried or lyophilized and resuspended in a new solvent. The product of the coupling reaction can be a library (one or more) of proteins, lacking the amino acid that was first in the sequence prior to processing, with a leader molecule.

Other primary amines will also react with an isothiocyanate group, and generally form a thiocarbamide. Thus, the reaction of PITC a molecule of the general structure NH2-X will form a molecule with the phenyl group of PITC linked to X via a thiocarbamide bridge. This structure makes the original NH2 group inaccessible for reactions of primary amines. Thus, a reaction with PITC effectively blocks the amine group of NH2-X.

The (N-1) method is described with a single Edman cycle, where the amine group from the second amino acid in the protein is exposed (as a primary amine). However, the method described can also be used with 2 or more cycles of chemistry, where the N-termini for the subsequent amino acid in the protein is exposed and used to attach another molecule.

Notably, the N-termini of proteins are commonly “blocked” – modified in a way as to make the primary amine at the N-terminus unavailable. Acetylation is particularly common, but there are a number of different possible N-terminal modifications. Such N-terminal modifications prevent the reaction of PITC or other ELRs with the N-terminal primary amine, and also prevent sequence determination by the Edman degradation. In that case, (N-1) method can be applied in two ways – after treatment of the proteins to deblock the N-termini and expose a primary amine, or prior to fragmentation of a protein to expose new N-termini that are otherwise internal to the protein (but leaving the original N-terminus blocked).

There are several different methods for deblocking the N-termini of proteins. The methods are widely used in protein sequencing by the Edman degradation and thus well known. They include the use of enzymes such as **Pyroglutamate Aminopeptidase and Methionine Aminopeptidase, and chemical methods such as the use of** trifluoroacetic acid together with methanol to remove N-terminal acetyl groups.

The (N-1) method can be applied to proteins prior to or after chemical or enzymatic fragmentation. When applied before, Edman cycle blocks all the primary amines and exposed the primary amine on the second amino acid. The protein is then fragmented using enzymatic or chemical methods, exposing additional new N-termini with primary amines for the fragments. Molecules can then be coupled to the primary amines in the sample. Alternatively, the can first be subjected to enzymatic or chemical fragmentation. This sample is then subjected to a cycle to the Edman reaction as above. This sample now has peptides with primary amines at their N-termini, but for which the amino acid that was initially at the N-terminus after fragmentation has been removed. Other primary amines in the sample, including lysine side chains, have been blocked.

A limitation of the (N-1) method is that it eliminates the N-terminal amino acid of a protein, and thus the identity of the first amino acid is unknown. This can prevent the unique identification of the N-terminal amino acid, including potential modifications of the N-terminal amino acid. However, in the vast majority of cases, the remaining amino acids retain sufficient information to identify the protein.

The (N-1) method can be applied to a sample immobilized on a solid support. A sample, including a complex sample such as serum, cell culture medium, interstitial fluids, plant cell lysate, cell culture lysate, lymph, etc. The sample may be immobilized on a solid support such as PVDF (including derivatized versions), polybrene or other materials used for protein sequencing based on the Edman method, including automated sequencing. The sample is subjected to a cycle of the Edman reaction. From some types of supports the sample can then be removed with an appropriate solvent and/or detergent, and subsequently used in a coupling reaction to the primary amines at the N-termini of the proteins. However, removing proteins adsorbed to a material such as PVDF typically produces very low yields. Instead, after the Edman cycle, such a sample can be subjected to enzymatic or chemical fragmentation. This will release peptides with primary amines at their N-termini to which other molecules can be coupled. This process can be automated by modifying instruments that perform automated Edman sequencing to perform a coupling reaction of the first Edman cycle.

When sequencing proteins with nanopores it is useful to tag and track the individual proteins or groups of proteins. This, for example, allows multiplexing, by permitting different samples or proteins from different sources to be mixed and sequenced at the same time. For example, proteins from sample 1 can be tagged with leader **I** while proteins from sample 2 can be tagged with leader **II**. Here leader **I** and **II** produce different electronic signatures or different sequences (in the data). That allows two (or more) samples to be mixed together and sequenced. By determining the sequence of the leader, or some other measurable property of the leader, for each protein that is sequenced, the sequences of the proteins in the two samples can be separated from each other. Such multiplexing is common in a variety of biochemical assays. However, when using a polymer/oligonucleotide leader coupled to a protein (Figure, Leader A, A-, B, B-), a pore designed for protein sequencing can fail to accurately determine the sequence of bases/monomers in the leader. To overcome this problem a hybrid leader is used, where one part of the leader is a polymer or oligonucleotide and a second part (attached to the first) is a short peptide/protein with a predetermined sequence (Leaders C, C-, D, D-). The hybrid leader is coupled to the protein of interest via the C-terminal end of the peptide part of the leader (or some other molecule that is attached to the C-terminal end of the peptide part of the leader). The peptide part of the hybrid leader can also be inverted, such that the N-terminus of the peptide part is coupled to the N-terminus of the protein to be sequenced. The first part of the leader serves as described earlier, where it guides the protein to the pore, and it may or may not include an adaptor (commonly used to facilitate engaging the end of the oligonucleotide or polymer with the pore. The second part of the leader, the protein part, is then sequenced as the molecule passes through the pore. The leader peptide encodes information about the source of the sample, and thus permits multiplexing (different samples have different peptide leaders). The leader can have redundant encoding to improve robustness, or the signal produced.

The leader can be a polypeptide with 2 or charged amino acids that carry 2 or more charges of the same sign. The charge on the amino acid leader can be uniformly distributed or non-uniformly distributed. One type of peptide leader is polyglutamic acid, another is polyaspartic acid, another is mixture of polyglutamic and polyaspartic acid. The leader can also be phosphorylated polyserine, phosphorylated polythreonine, phosphorylated polyphospho-tyrosine, or a mixture of these phosphorylated amino acids and glutamic or aspartic acid. Peptide leaders can be added chemically to N-termini in a sample that are made available by the (N-1) method. Peptide leaders can also be encoded in vectors used to express proteins.

The (N-1) method can also be performed on samples in which the proteins in the sample are fragmented. A fragmented sample consists of a sample in which at least one protein is broken into two parts, which each part is at least 2 amino acids long A sample can be treated in a way to fragment proteins prior to (N-1). Such a sample can be subjected to proteolysis by trypsin, chymotrypsin, pepsin, Acylamino-acid-releasing enzyme, Arginylendopeptidase, Asparaginyl endopeptidase, Endoproteinase Asp-N or any other protease that produces a desirable pattern of degradation products. Such cleavage can be performed on a column where the protease is immobilized, and the protein products are eluted. The sample can also be treated chemically, using for example cyanogen bromide, to cleave the proteins into shorter proteins/peptides. There are other methods for cleaving proteins known. The fragmentation exposes new N-termini on the protein fragments produced. The (N-1) method can then be applied to the sample, which reduces the length of all the proteins with an available primary amine at the N-terminus by 1 amino acid and produces a primary amine at the new N-termini. In these samples, non-peptide primary amines are blocked and not available for subsequence amine-specific reactions. The new N-termini on the proteins on the other hand can be reacted with any desired amine chemistry, including amine chemistries by which a leader for nanopore protein sequencing is attached.

A sample can also be subjected to (N-1) chemistry prior to fragmentation. In this case, the first amino acid at N-termini with primary amines is removed, and other primary amines are blocked. The proteins in the sample are then fragmented as described above. Performing the (N-1) chemistry first has the benefit that only the first amino acid on proteins with primary amines at their N-termini is lost. The non-N-terminal proteins/peptides from the fragmentation will have primary amines that are available for coupling, and thus they do not lose their first amino acid. (N-1) treatment prior to fragmentation will also alter the fragmentation pattern in some cases. For example, trypsin normally cleaves adjacent to lysines, but will not cleave next ERL modified lysines.

Modification/blocking of the lysine side chain(s) by ELRs will in many cases inactivate or prevent the normal functioning of the modified protein, both when the N-terminal amino acid is removed or when it is not removed (because it is blocked). Such a change to the function of a protein is analogous to replacing an amino acid by site-directed mutagenesis. The change in function can be to the enzymatic activity, how enzymatic activity depends on solvent composition, the affinity to molecules that bind to the protein, the sensitivity to drugs or modulators or the protein, to the structural or mechanical properties of the molecule, or any other aspect of the protein function that depends on the structure of the protein. Side-chain modifications may also activate or improve such functions or provide for new functions or activities.

Samples can be purified to varying extents, using for example PAGE, 2D PAGE, Capillary Gel Electrophoresis, column chromatography (size exclusion, ion exchange), FPLC, HPLC, differential centrifugation, density centrifugation, and other purification methods. (N-1) can be applied to fractions from an isolation method, pools of the fractions, or the entire purified product.

Proteins can be modified with monosaccharides, disaccharides and/or polysachharides (>2 saccharide groups). Many polysaccharides are too large to pass through the pores used for nanopore sequencing, and thus would interfere with the process. Many such polysaccharides can be removed enzymatically, using for example Peptide-N-Glycosidase F and Endoglycosidase H. Other methods for removing polysaccharides from proteins are well known. If glycosylated proteins present a problem during nanopore sequencing, by for example clogging pores, they can also be removed from the sample using – for example – lectins (on solid supports).

In addition to adding a leader to proteins, the (N-1) method can be used to add fluorophores, spectroscopic labels, binding labels, tags, affinity labels, enzymes, catalytic groups and other functionalities to a protein.

(N-1) can be used to couple two proteins together, by so-called native chemical ligation. Such coupling can be used to produce libraries of NCL produced that produce combinations of natural proteins and/or synthetic proteins.

One example: A sample of serum from a human is collected by conventional clinical techniques. 100 ul of the sample is subjected to a spin column based on Sephadex G-25 or equivalent Sephacryl or similar size exclusion media that removes small molecules and introduces a suitable buffered solution (e.g., 50 mM CHES, N-Cyclohexyl-2-aminoethanesulfonic acid). Sulfo-PITC is added to the sample to a concentration of 1.5 M, and the pH of the sample is adjusted to ~9. The sample is then incubated at 50 C. The sample is then acidified with TFA and incubated at 50 C. This sample is then subjected to another spin column as above, but with a 50 mM sodium phosphate buffer at pH~7.2. If needed, the pH is adjusted to 7.2 or the sample is subjected to a second spin column with sodium phosphate. The sample is now composed of proteins and other molecules from the serum sample that are above the cutoff size of the size exclusion media. Proteins that initially had a free primary amine, have had the first amino acid removed and a primary amine on the second amino acid exposed. The majority of molecules that had primary amines in the original serum sample have either been removed by the size exclusion chromatography or been blocked by the sulfo-PITC. The sample can now be reacted with an oligo-NHS molecule to produce a sample where the majority of proteins with free primary amines at their N-termini are coupled to the oligo. These oligo couple proteins can then be subjected to nanopore sequencing.

**References**

Methods in Molecular Biology, vol. 211: Protein Sequencing Protocols, 2nd ed. Edited by: B. J. Smith © Humana Press Inc., Totowa, NJ

Hewick, R. M., Hunkapiller, M. W., Hood, L. E., and Dryer, W. J. (1981) A gas-liquid solid phase peptide and protein sequenator. *J. Biol. Chem.* **256**, 7990–7997.

Massa, S., Devoogdt, N. (Eds.), 2019. Bioconjugation: Methods and Protocols, 1st ed. 2019 edition. ed. Springer, New York, NY.

Koniev, O., Wagner, A., 2015. Developments and recent advancements in the field of endogenous amino acid selective bond forming reactions for bioconjugation. Chem. Soc. Rev. 44, 5495–5551. <https://doi.org/10.1039/C5CS00048C>

Smith, L.M., Kelleher, N.L., and The Consortium for Top Down Proteomics, 2013. Proteoform: a single term describing protein complexity. Nat Methods 10, 186–187. <https://doi.org/10.1038/nmeth.2369>

Shively, J.E. (Ed.), 1986. Methods of protein microcharacterization: a practical handbook, biological methods. Humana Press, Clifton, N.J.

**List of Edman Like Reagents:**

PITC (including tritium or other isotopically labeled PITC)

FITC

1,3-bis-(Trifluoromethyl)-Phenylisothiocyanate

Any of the Edman-like reagents isotopically labeled with for example tritium or P32.

Sulfo-PITC

Edman Like Reagents also include:

R-N=C=S

In which R represents:

A linear, branched or cyclic C-C, alkyl group, optionally having one or more oxygen atoms,

a C-C haloalkyl group,

a linear or branched C-C alkenyl group,

a linear or branched C-C alkynyl group,

a C-C hydroxyalkyl group,

a protected C-C aminoalkyl group

a C-C cyanoalkyl group,

a linear or branched C-C alkyl group, which is optionally substituted with one or more **R** Substituents, or

an aromatic ring which is non-substituted or Substituted with one or more atoms or groups of atoms Selected from halogens, linear or branched C-C alkoxy, linear, branched or cyclic C-C, alkyl, linear or branched C-C alkylthio, cyano, hydroxy, nitro, trifluoromethyl, trifluoromethoxy, methylenedioxy, ethylenedioxy, difluoromethyl enedioxy, aminoSulphonyl, dimethylamino, C-C, hydroxyalkyl, carboxylic acid, C-C alkyl ester, methane Sulphonylamino, benzene Sulphonyl amino, t-butoxycarbonyl amino, or



groups.

**List A of Solvents:**

Acetonitrile

` Ethanol

Methanol

Tetrahydrofuran

Triethylamine (TEA)

Dimethylformamide (DMF)

Water

Propanol

Acetic acid

Formic acid

---------------

Top of Form

Bottom of Form

**Figure 1 – Leader structure(s)**

Diagram

Description automatically generated

**Figure 2. Edman and (N-1) chemical reaction pathways.**

Diagram

Description automatically generated