# 5

# Determination of the Primary Structure of Proteins and Peptides

## *INTRODUCTION*

The determination of the primary structure (sequencing) of a polypeptide chain by chemical means represents a major investment of time for all except relatively short polypeptide chains. With small proteins it is possible that the complete sequence can be established at least semiautomatically using a "sequenator": With larger proteins, proteins containing disulfides (either internal or between subunits), or proteins with covalent substituents such as glycoproteins, the straightforward approach is not always possible. In such instances it is necessary to fragment the polypeptide chain in some defined and controllable way, purify the fragments, and characterize them in terms of their amino acid sequences. Once the primary structure of each fragment is obtained the fragments must be aligned in the correct order to give the linear sequence of the protein. The cysteine residues involved in disulfide bonds must be matched and a number of other assignments, such as the position of carbohydrate or lipid substituents, must be made, as well as the positions of amides determined. Each one of these aspects represents separate challenges to the protein chemist.

Although many proteins have been sequenced using the approaches described in this chapter, the main aim is not to prepare the protein chemist to sequence a polypeptide chain completely since, as outlined briefly at the end of the chapter, there are faster approaches available using the tools of molecular biology. The approaches described are aimed at another problem, that of the isolation and identification of a particular peptide in a protein. A frequently encountered problem is identifying a peptide containing a residue that may have been specifically modified by a chemical reagent or which might contain a disulfide bond. If the primary sequence of the

protein is known (which increasingly *is* the case), the peptide can be identified by establishing the sequence of a relatively small number of amino acids in the fragment (usually four or five residues are sufficient). Of course, prior to sequencing the fragment, it must be (1) generated and (2) purified, hence the need for an understanding of the topics covered in this chapter. Some of the techniques discussed here involving the chemical modification of particular amino acid side chains are discussed in detail in Chap. 7.

Before proceeding, we must digress briefly to discuss the problem of the purity of fragments, which must be achieved prior to attempting sequence work. If we consider a hypothetical situation where the yields of peptides obtained in purification may vary between 5 and 90%, we can discuss the problem of contaminating proteins. During peptide purification it is a natural assumption that the peptides present in the highest amounts are those that arise from the intended protein rather than from impurities. If, however, we have a 10% contaminant in the original protein preparation (which could easily escape detection), and after fragmentation we isolate a peptide from this contaminant with a 90% yield, we may overlook a peptide from the real protein which was isolated in only 5% yield and consider it as a minor contaminant, or isolate both peptides, sequence them, and assume that both arise from the real protein. Obviously, under such conditions small amounts of contaminating protein can be a real problem. If we are attempting to isolate a specific peptide that has been labeled with a site-specific modification reagent, it is less likely that such a problem will be encountered unless there is a significant amount of nonspecific labeling. Even in such favorable circumstances, however, contaminating proteins can hinder the purification of the desired peptide: Simply by being there they will increase the work (and probably decrease the yield) of purifying the wanted peptide.

## FRAGMENTATION METHODS

There are two different approaches that can be used to cleave a polypeptide chain at fairly defined positions. These involve chemical and biochemical methods using specific proteolytic enzymes. The phrase "fairly defined" implies that the cleavage specificities of both approaches may vary with the particular protein. In this section we discuss the generalities of the specificity. In the context of using proteolytic enzymes, the question of whether to perform the cleavage under conditions where the protein is in its native state or has first been denatured must be addressed. In some cases limited proteolysis of a native protein can give considerable useful information, as emphasized in Chap. 10.

Prior to attempting cleavage with any of these various reagents, another question must be addressed which pertains to disulfide bonds. There are two problems encountered: (1) Do you want to leave them intact during fragmentation, and (2) can you prevent spurious disulfide bonds forming as a result of oxidation events occurring either during fragmentation or subsequent peptide purification? Usually, sulfide bonds that exist in the protein are reduced and alkylated prior to fragmentation—overcoming problem 2. Of course, when one is attempting to establish the location

of disulfide bonds, cleavage is performed without first reducing the disulfides. Any free sulfhydryls are usually blocked from reaction by prior alkylation of the denatured protein.

### *Chemical Methods*

Although a number of quite esoteric chemical methods of cleaving polypeptide chains at specific sites have been described, the ones mentioned here are those most commonly encountered.

*Cyanogen Bromide Cleavage.* Cyanogen bromide cleaves polypeptide chains at methionyl-X peptide bonds via the mechanism shown in Fig. 5-1. The reaction is usually performed in a 70% formic acid or trifluoroacetic acid to water (v/v) mix with a 50- to 100-fold molar excess of cyanogen bromide to methionine. The total reaction time is usually 24 hours and in many cases the cyanogen bromide is added in two batches during the incubation.

While cyanogen bromide cleaves at methionine-X peptides, the nature of X can influence the rate of cleavage. In particular, the hydroxyl-containing side chains of serine and threonine interfere by attacking the iminolactone intermediate of Fig. 5-1, leading to the conversion of methionine to homoserine lactone without the cleavage of the polypeptide chain (see Fig. 5-2). However, the use of high CNBr-to-methionine ratios (up to 500:1) overcomes this effect and cleavage of Met-Ser or Met-Thr up to 80% is usually achieved under the experimental conditions described.

The acid conditions employed in CNBr cleavage can cause cleavage of other peptide bonds in the protein, especially Asp-Pro bonds, and if a particular sequence contains this bond, more fragments than would be predicted based on the number of methionine residues may be obtained.

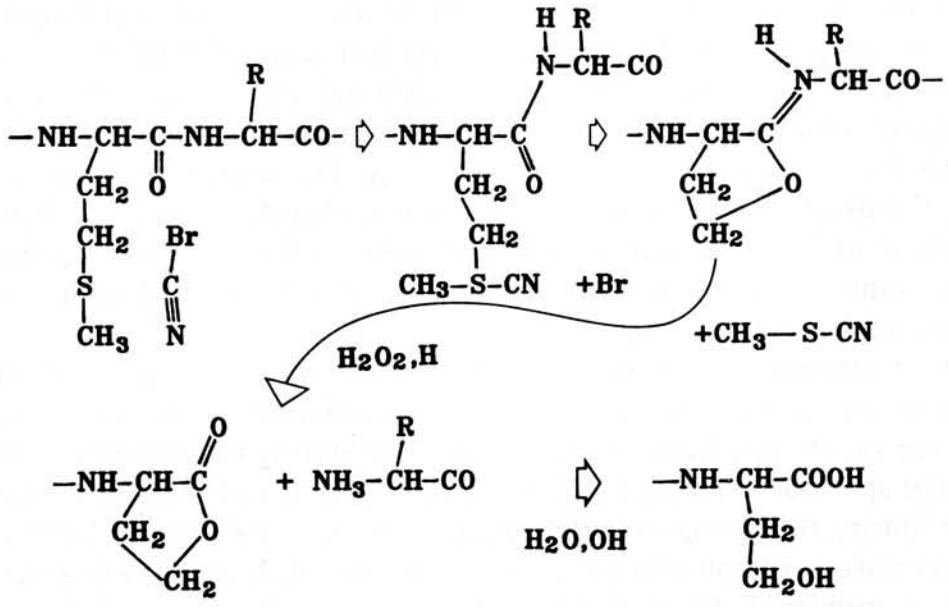



**Figure 5-1** Mechanism of cleavage of methionyl-X bonds by cyanogen bromide.

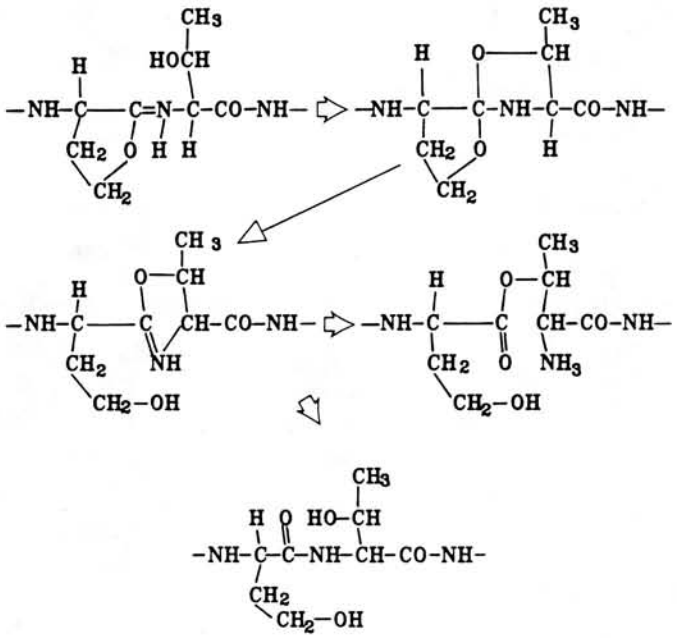


**Figure 5-2** Interference of cleavage of methionyl-serine or methionyl-threonine bonds by hydroxyl attack on the iminolactone intermediate.

*BNPS-Skatole.* For a number of years brominating reagents in acidic media have been used to cleave polypeptide chains. Reagents such as *N*-bromosuccinimide will cleave polypeptides at a variety of sites, including tryptophan, tyrosine, and histidine, but often give side reactions which lead to insoluble products. BNPS-skatole [2-(2-nitrophenylsulfenyl)-3-methylindole] is a mild oxidant and brominating reagent that leads to polypeptide cleavage on the C-terminal side of tryptophan residues, as shown in Fig. 5-3.

Although reaction with tyrosine and histidine can occur, these side reactions can be considerably reduced by including tyrosine in the reaction mix. Typically, protein at about 10 mg/ml is dissolved in 75% acetic acid and a mixture of BNPS-skatole and tyrosine (to give 100-fold excess over tryptophan and protein tyrosine, respectively) is added and incubated for 18 hours. The peptide-containing supernatant is obtained by centrifugation.

Apart from the problem of mild acid cleavage of Asp-Pro bonds, which is also encountered under the conditions of BNPS-skatole treatment, the only other potential problem is the fact that any methionine residues are converted to methionine-sulfoxide, which cannot then be cleaved by cyanogen bromide. If CNBr cleavage of peptides obtained from BNPS-skatole cleavage is necessary, the methionine residues can be regenerated by incubation with 15% mercaptoethanol at 30°C for 72 hours.

*Cleavage with o-Iodosobenzoic Acid.* *o*-Iodosobenzoic acid cleaves tryptophan-X bonds under quite mild conditions. Protein, in 80% acetic acid containing 4 *M* guanidine hydrochloride, is incubated with iodobenzoic acid (approximately 2 mg/ml of protein) that has been preincubated with *p*-cresol for 24 hours in the dark at room temperature. The reaction (which can be terminated by the addition of dithioerythritol) proceeds via the scheme shown in Fig. 5-4.

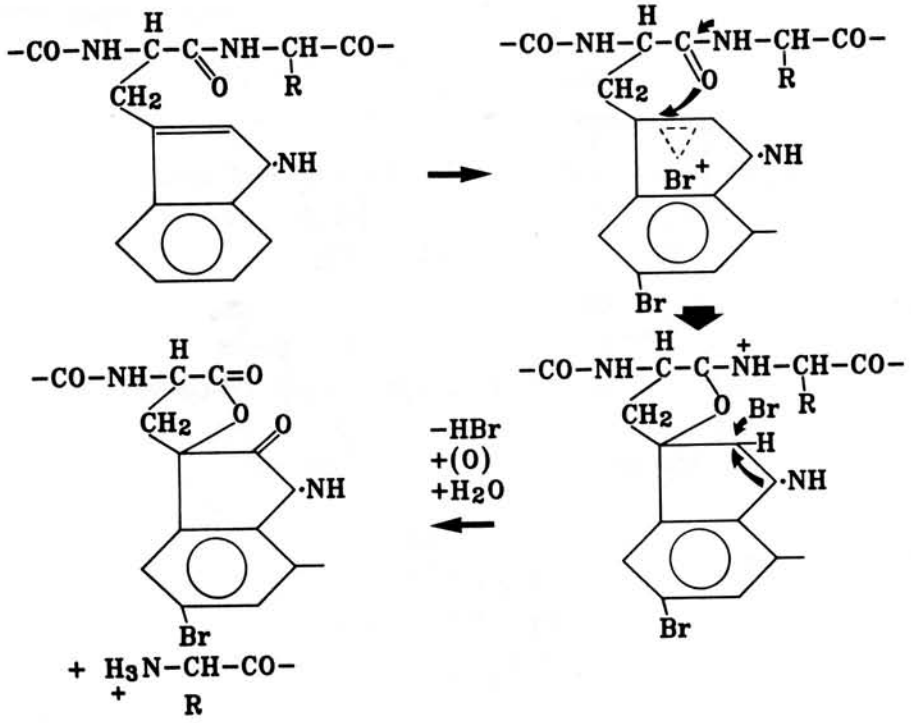



**Figure 5-3** Cleavage of tryptophan-X peptide bond by brominating reagents such as BNPS-Skatole or *N*-bromosuccinimide.

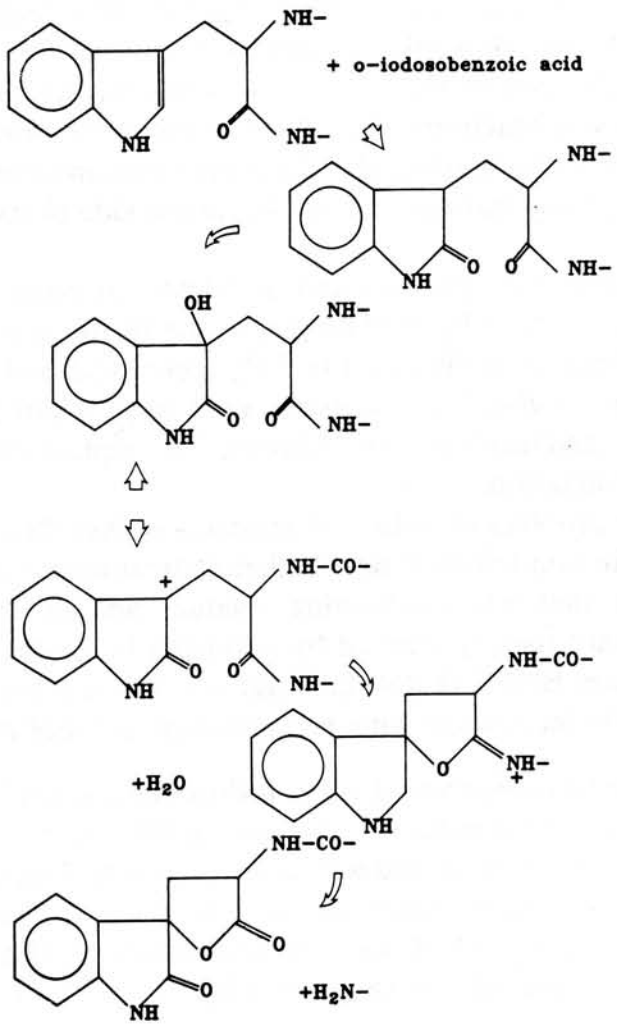



**Figure 5-4** Cleavage of tryptophan-X peptide bonds by *O*-iodosobenzoic acid.

Care must be taken to use purified *o*-iodosobenzoic acid since a contaminant, *o*-iodoxybenzoic acid, will cause cleavage at tyrosine-X bonds and possibly histidine-X bonds. The function of *p*-cresol in the reaction mix is to act as a scavenging agent for residual *o*-iodoxybenzoic acid and to improve the selectivity of cleavage.

*Cleavage of X-Cysteinyl Bonds.* Two reagents are available that produce cleavage of peptides containing cysteine residues. In both cases cleavage occurs on the amino-terminal side of the cysteine. The mechanisms of these reagents, (2-methyl)-*N*-1-benzenesulfonyl-*N*-4-(bromoacetyl)quinone diimide (otherwise known as Cyssor, for "cysteine-specific scission by organic reagent") and 2-nitro-5-thiocyanobenzoic acid (NTCB), are shown in Fig. 5-5.

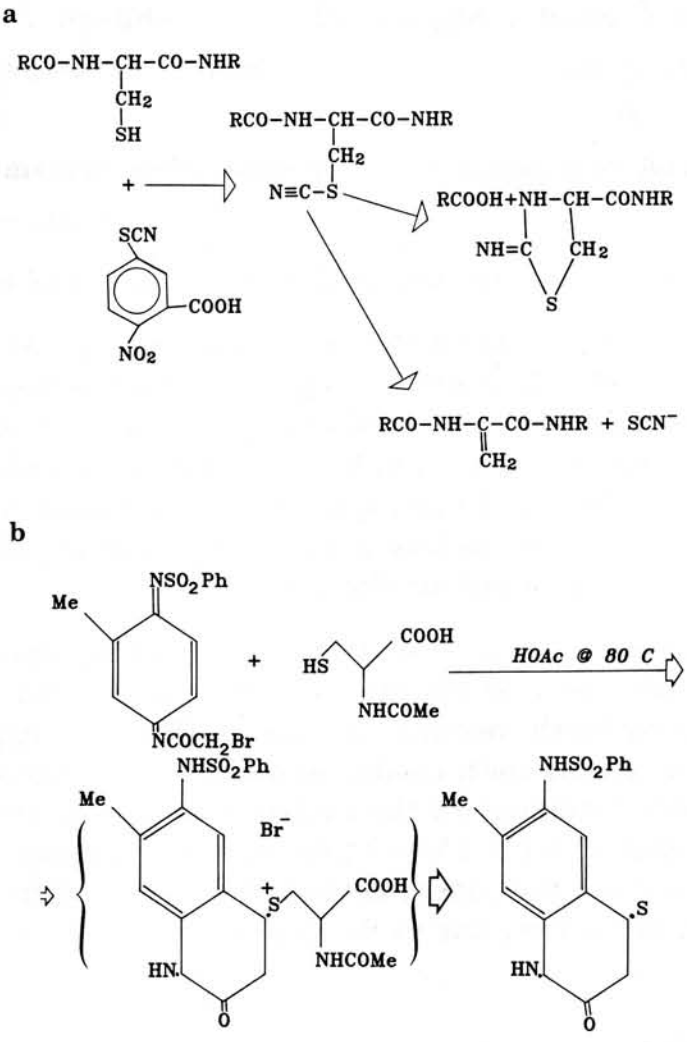



**Figure 5-5.** Cleavage of X-cysteinyl peptide bonds by reagents such as NTCB (a) or "Cyssor" reagents such as Cyssol-1 (b).

*Hydroxylamine Cleavage.* Hydroxylaminolysis slowly leads to cleavage of a number of peptide bonds; however, it has been found that the asparaginyl–glycine bond is particularly susceptible and cleavage can be achieved by the reaction outlined in Fig. 5-6. The reaction occurs by incubating protein, at a concentration of about 4 to 5 mg/ml, in 6 *M* guanidine hydrochloride, 20 m*M* sodium acetate + 1%

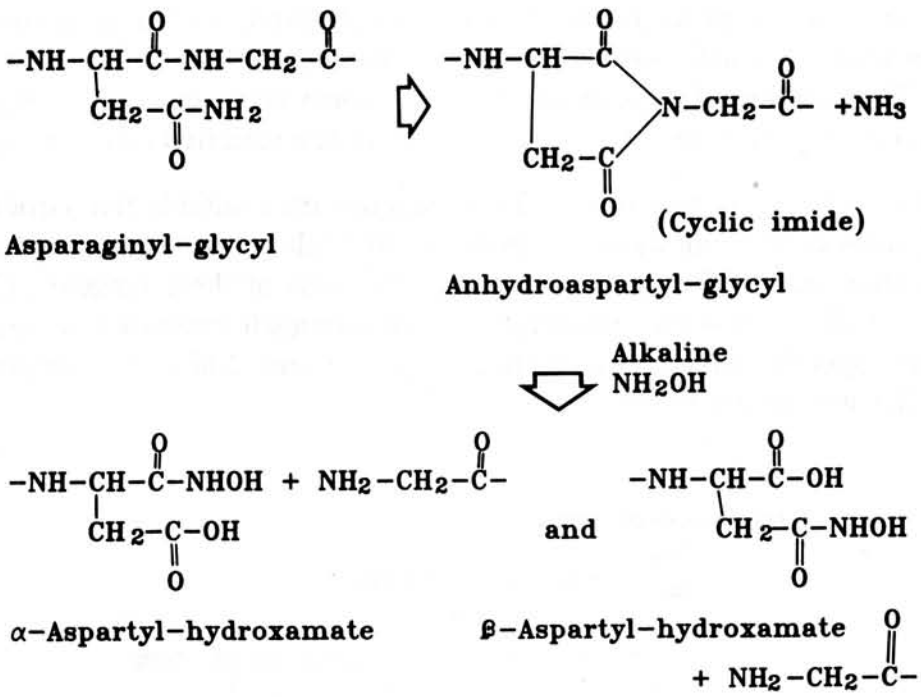



**Figure 5-6** Hydroxylamine cleavage of asparaginyl–glycine peptide bonds.

mercaptoethanol at pH 5.4, and adding an equal volume of 2 *M* hydroxylamine in 6 *M* guanidine hydrochloride at pH 9.0. The pH of the resultant reaction mixture is kept at 9.0 by the addition of 0.1 *N* NaOH and the reaction allowed to proceed at 45°C for various time intervals; it can be terminated by the addition of 0.1 volume of acetic acid. In the absence of hydroxylamine, a base-catalyzed rearrangement of the cyclic imide intermediate can take place, giving a mixture of α-aspartylglycine and β-aspartylglycine *without* peptide cleavage.

*Cleavage of Asp-Pro Bonds.* Peptide bonds containing aspartate residues are particularly susceptible to acid cleavage on either side of the aspartate residue, although usually quite harsh conditions are needed (Fig. 5-7). Asp-Pro bonds have been found to be susceptible under conditions where other Asp-containing bonds are quite stable. Suitable conditions are the incubation of protein (at about 5 mg/ml) in 10% acetic acid, adjusted to pH 2.5 with pyridine, for 2 to 5 days at 40°C. The enhanced reactivity of Asp-Pro bonds relative to other Asp-containing peptides under these conditions is presumably due to the greater basicity of the imino ring in the proline.

### *Enzymatic Methods*

A wide variety of different endopeptides have been described and purified, and many have found use in the enzymatic fragmentation of polypeptide chains. Some of these proteases are specific for peptide bonds containing certain defined side chains on adjacent amino acids, and the more common are given in Table 5-1 together with their specificities.

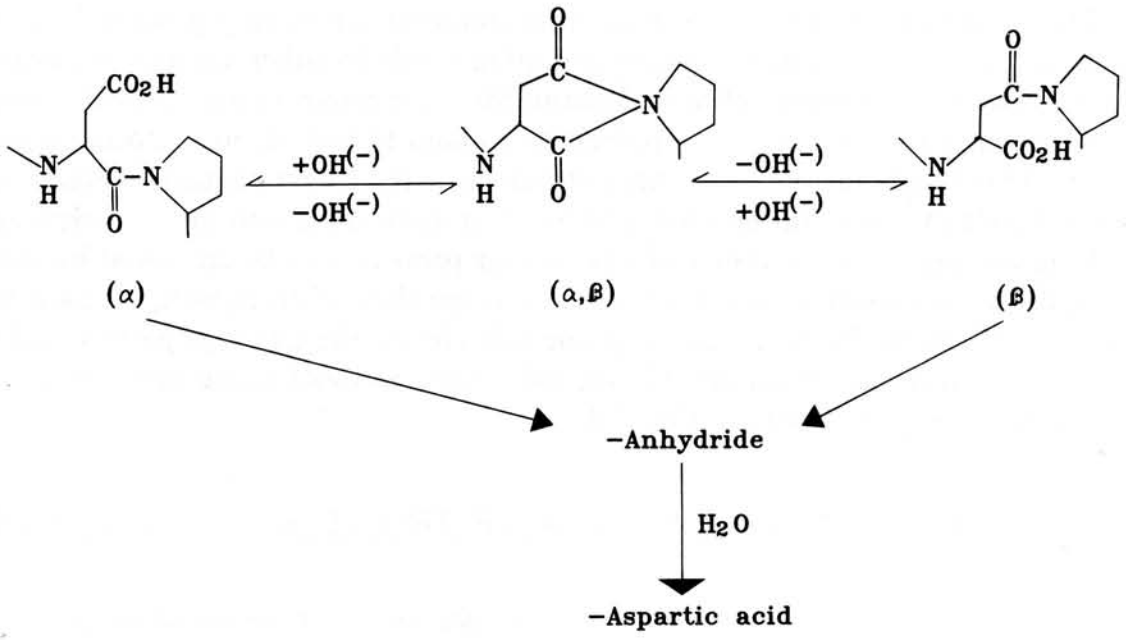



**Figure 5-7** Hydrolysis of aspartyl–proline peptide bonds at acidic pH.

**TABLE 5-1** Specificity of endopeptides used in fragmentation

| Enzyme | B Site–C site specificity[a] | Other requirements |
|---|---|---|
| Trypsin | B = Lys or Arg | No hydrolysis if C = Pro; hydrolysis slowed if A or C are acidic |
| Chymotrypsin | Preferentially, B = Trp, Tyr, Phe; some hydrolysis if B = Met, Leu, or His | As with trypsin |
| Pancreatic elastase | B = Ala, Val, Gly, or Ser | |
| Thermolysin | Preferentially, C = Phe, Leu, Val, Tyr, Ile, Met, or Trp; some hydrolysis if C = Ala, Asn, Thr, or His | No hydrolysis with C = Gly or with D = Pro |
| Pepsin | B or C = Phe, Tyr, Trp, Leu (cleaves others slowly) | No cleavage if B = Pro |
| Subtilisin | B or C = hydrophobic | |
| Papain | Broad specificity | |
| Streptococcal proteinase | A, B = bulky (e.g., Phe, Tyr, Leu, His) | No cleavage if B = Gly |
| Staphylococcal protease | B = Glu or Asp | Hydrolysis slowed if C = hydrophobic; C = Pro does *not* prevent cleavage |

[a] Specificities given in terms of a tetra-peptide A-B-C-D with *site* of cleavage between B and C.

The conditions used for enzymatic fragmentation are usually governed by the optimal activity of the protease. Many are quite stable to either somewhat elevated temperatures or the presence of mild denaturants. It is common practice when using proteolytic enzymes to fragment a polypeptide chain to include low concentrations of SDS. This breaks the native tertiary structure of the target protein, exposes susceptible bonds in a uniform manner, and leads to more rapid and uniform cleavage.

In many instances, the utility of a particular protease can be enhanced by using specific chemical modifications to alter the cleavage sites: With trypsin, for example, which has specificity for lysine and arginine side chains, the cleavage pattern can be altered by chemical modification of lysine side chains to block those sites. This process is outlined schematically in Fig. 5-8.

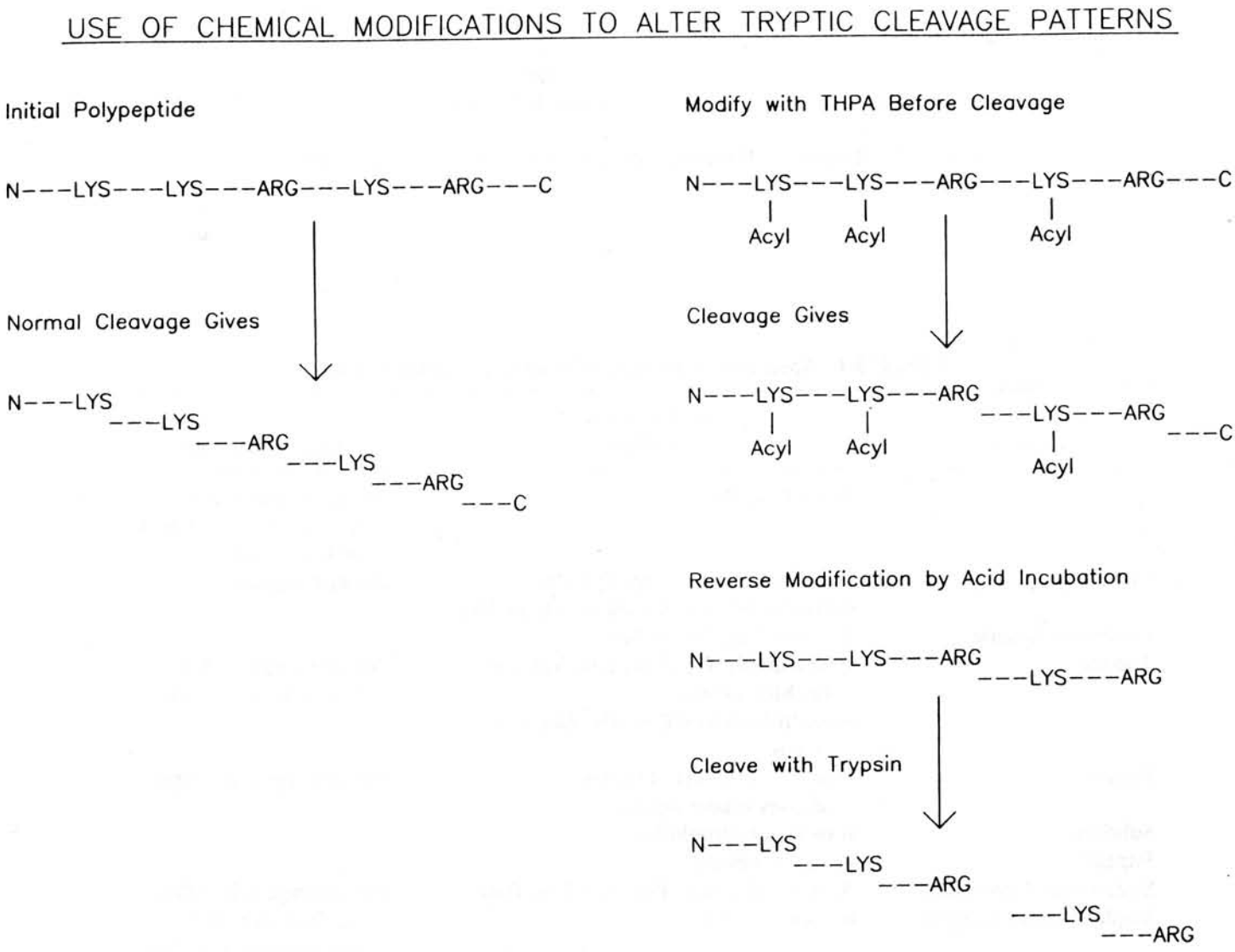



**Figure 5-8** Modification of tryptic cleavage patterns by chemical modification of lysine side chains with tetrahydrophthalic anhydride (THPA).

In addition, reagents can be used that modify arginine and not lysine residues. If reversible modification reagents are selected, the blocked sites can be restored, thus allowing subsequent cleavage. With trypsin it is also possible to create additional cleavage sites if cysteine residues are present.

As shown in Fig. 5-9, the cysteine side chain can be chemically modified with aminoethylating reagents such as *N*-(β-iodoethyl)trifluoroacetamide or ethyleneimine, which produces *s*-(2-aminoethyl)cysteine, which is homologous to the side chain of lysine and is recognized by trypsin. Such manipulations can be quite useful in generating different types of fragments for use in establishing overlaps.

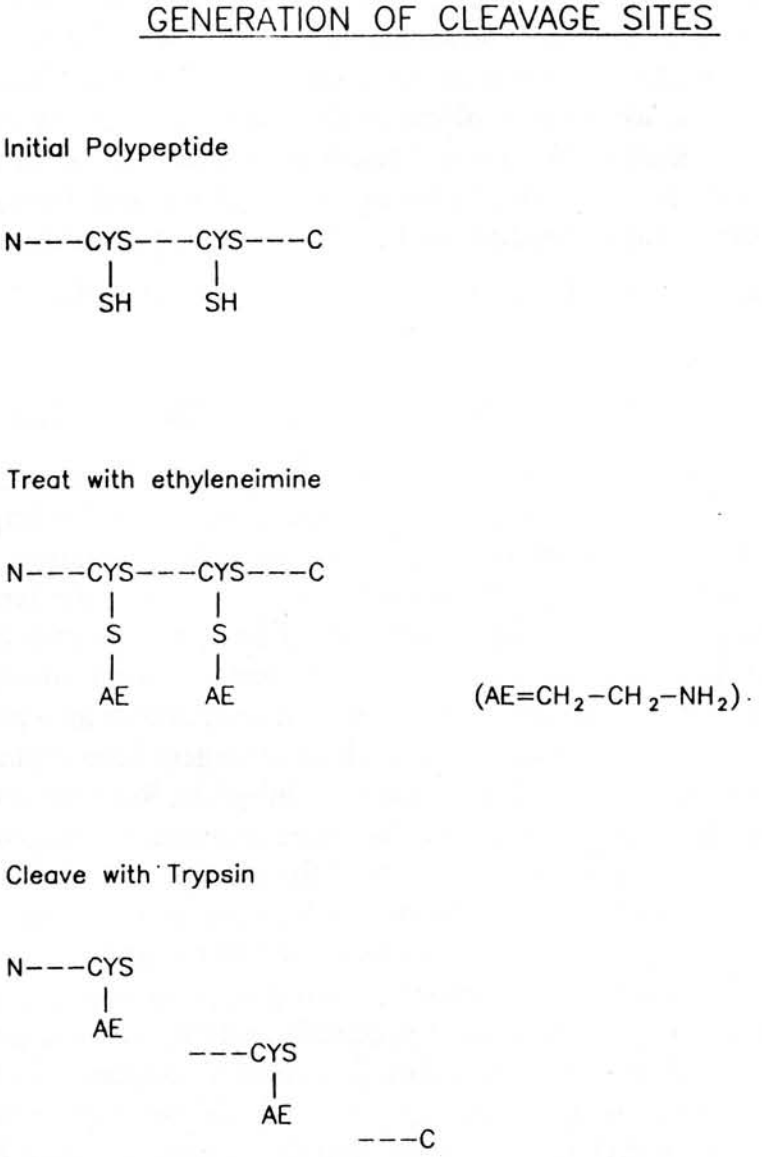



**Figure 5-9** Creation of tryptic cleavage sites by modification of cysteine side chains with ethyleneimine.

## ISOLATION AND CHARACTERIZATION OF FRAGMENTS

In many ways the problems encountered in the isolation of the polypeptide chain fragments are the same as those discussed in earlier chapters on protein purification: gel filtration, ion-exchange chromatography, HPLC (especially RP-HPLC), and even affinity chromatography are all employed to purify fragments. Polyacrylamide gel electrophoresis and paper electrophoresis or chromatography are also widely used. Although all of these approaches have been discussed in Chaps. 1 to 4, several general comments must be made. Preparative work with proteins usually involves buffered aqueous systems: When working with peptides, harsher conditions are often used. With peptides it is not necessary to maintain "native" structure and they (especially those from hydrophobic regions of a protein) may be quite insoluble in aqueous buffers. It is not unusual to use formic acid or acetic acid solvents for column chromatography of peptides or to use solvents with low dielectric constants for hydrophobic peptides. The solubility problem is also overcome in some cases by chemical modification of the peptide. With disulfide-containing peptides, sulfitolysis [Eq. (5-1)] is often used to enhance solubility by introducing a group with high polarity without leading to the oxidation of residues such as tryptophan or methionine.

$$R_1—S—S—R_2 + SO_3^{2-} \longrightarrow R_1—S^- + R_2—S—SO_3^- \qquad (5\text{-}1)$$

then

$$R_1—S^- + SO_3^{2-} + 2Cu^{2+} \qquad R_1—S—SO_3^- + 2Cu^+$$

Another problem, which is somewhat different from those encountered in the protein purification, involves detection and quantitation of the fragments. Absorbance measurements at 280 nm are frequently used to follow elution of proteins from columns. However, with peptide fragments many do not contain tyrosine or tryptophan (the primary contributors to absorbance at 280 nm) and thus are not detected. Measurements taken at wavelengths where the peptide bond absorbs can be used, but quantitation is still a problem. Absorbance measurements also present a problem in terms of sensitivity, and for accurate work have largely been replaced by methods using fluorescence detection or radioactivity. Ninhydrin, long the standard chemical reagent of choice for peptide detection, has been replaced by reagents such as fluorescamine and *o*-phthalaldehyde. Both give fluorescent adducts of amino groups: fluorescamine has an excitation at 390 nm with emission at 475 nm, while *o*-phthalaldehyde gives a derivative with an excitation at 340 nm and emission at 455 nm.

Fluorescamine reacts with primary amino groups in peptides at pH 7.5 to 9.0 in aqueous solution to give fluorescent products, and the reaction can be carried out to monitor a column effluent. The method is readily adaptable to the detection of peptides on paper or thin-layer chromatograms or on electrophoretograms that are first treated with a triethylamine solution and then sprayed with a fluorescamine in acetone solution. After drying, peptides are detected by their fluorescence under a long-wavelength UV lamp. Peptides with proline as their amino terminus become observable only after heating at 110°C for 3 hr.

*o*-Phthalaldehyde is used in essentially the same way, with the advantage that it is more soluble and more stable in aqueous solutions. The most popular radioactive compound used for peptide detection is phenylisothiocyanate, which is discussed in more detail later. There are several preparative methods which find use in peptide purification that are somewhat unique to peptides.

### *Countercurrent Methods*

Countercurrent methods are based on phase partition of peptides: A mixture in a particular solvent is partitioned using a nonmiscible second solvent. Separation is achieved if one or more peptides has a higher solubility in one of the phases than it has in the other: If the two phases are separated after mixing, peptides with higher solubility in, for example, the top phase are extracted into that phase, while other peptides with higher solubility in the bottom phase remain preferentially in that phase. If the process is now repeated—fresh top phase is added to the bottom phase, and vice versa—further extraction and enrichment takes place. This cycle can be repeated until, for all practical purposes, the peptides have been completely separated. The process is shown diagrammatically in Fig. 5-10 for an example where two

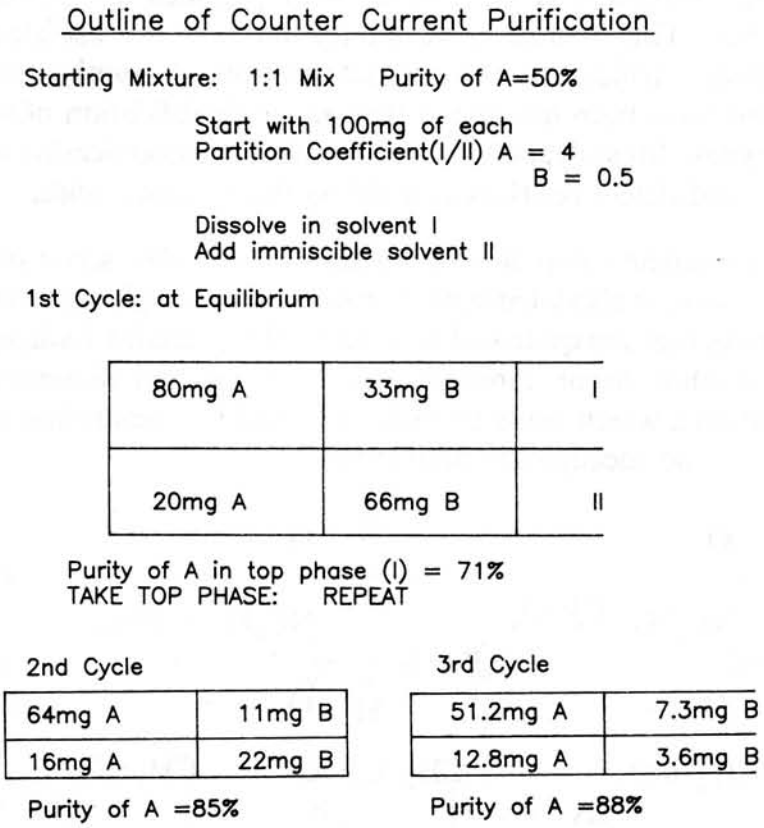



**Figure 5-10** Schematic outline of the separation of peptides by countercurrent purification.

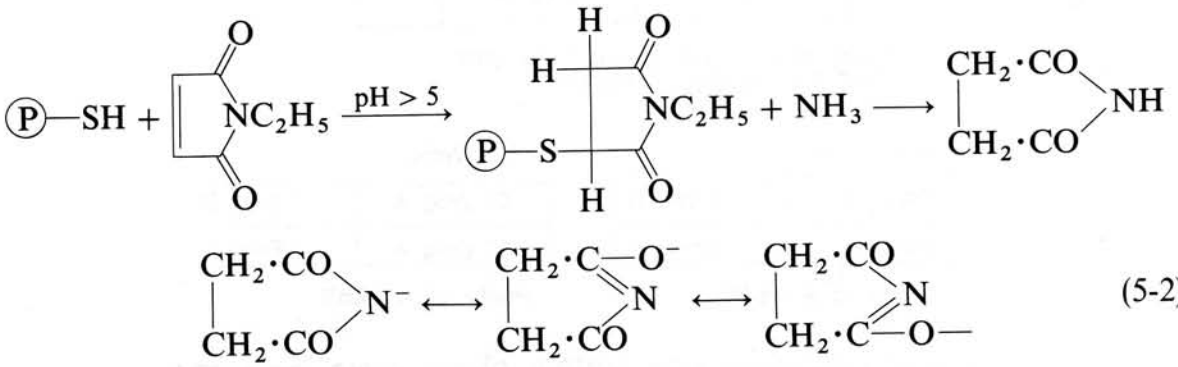
peptides have quite different distribution coefficients between the two solvents. Although this example might be considered to represent an optimal situation, all that is necessary for separation to be eventually achieved is that the peptides to be separated have different distribution coefficients.

### *Diagonal Methods*

The basis of diagonal methods was introduced in Chap. 4. The method depends on a two-dimensional electrophoresis or chromatography with both dimensions carried out under identical conditions except that they are rotated through 90° between the dimensions. If nothing is done to the mixture of peptides separated in the first dimension prior to the second, the peptides behave just as they did during the first dimension, giving a diagonal pattern. However, if between the two dimensions some of the peptides are altered in such a way that their mobility is different in the second dimension than in the first dimension, they do not fall on the diagonal and can thus be identified.

Previously, we discussed the use of diagonal methods for identifying proteins containing inter-subunit disulfide bonds, where reduction was employed between the dimensions. This approach is very useful with peptides for identifying disulfide-bonded fragments. This is not, however, the limit of the usefulness of diagonal methods in peptide purification and characterization. A number of treatments between dimensions have been developed that allow identification of a number of different peptides types. Most depend on specific chemical modification of certain amino acid side chains and detect peptides containing these amino acids.

*Sulfhydryl-Containing Peptides.* A mixture of peptides, some of which contain free sulfhydryl groups, is alkylated with *N*-ethylmaleimide, giving *N*-ethylsuccinimide cysteine derivatives that are quite stable at acid pH but readily hydrolyzed at alkaline pH. After electrophoresis or chromatography in the first dimension the paper is exposed to ammonia, which leads to hydrolysis and the generation of an additional negative charge at the succinimide derivative:




(5-2)

The generated *N*-ethylsuccinamic acid derivative moves differently in the second dimension than in the first and moves off the diagonal. If two-dimensional electrophoresis is used, the sulfhydryl-containing peptides move more toward the positive

in the second dimension. In another approach involving cysteine residues, performic acid oxidation after the first dimension produces a cysteic acid derivative which electrophoreses more to the anodic side of the diagonal in the second dimension.

*Methionine-Containing Peptides.* If after separation in the first dimension the peptides are alkylated with iodoacetamide, sulfonium derivatives are formed from any methionine-containing peptides that migrate to the cathodic side of the diagonal during electrophoresis in the second dimension.

*Lysine-Containing Peptides.* If a protein *prior to fragmentation* is acylated with a reagent such as tetrahydrophthalic anhydride, negative charges are introduced onto lysine side chains and the amino-terminal residue (unless the protein has a blocked amino terminal). The protein is then fragmented and electrophoresed in the first dimension. Prior to the second dimension the pH is lowered to 4–5, which produces deacylation of the modified amino residues and removes negative charge from those peptides, leaving derivatives that electrophorese on the cathodic side of the diagonal in the second dimension. This approach identifies lysine-containing peptides *and* the amino-terminal peptide.

*Identification of the C-Terminal Peptide after Tryptic Digestion.* Unless the protein in question has a C-terminal lysine or arginine residue, all but one of the tryptic peptides of the protein will have a C-terminal lysine or arginine residue. As a result, treatment of these peptides with pancreatic carboxypeptidase B removes a positively charged residue from the C terminal, altering the electrophoretic mobility. If tryptic peptides, after electrophoresis in the first dimension, are subjected to carboxypeptidase B digestion prior to electrophoresis in the second dimension, all except the C-terminal peptide will have altered mobility. As a result, the only peptide remaining on the diagonal will be the C-terminal peptide.

Although each of the techniques described were designed for use in paper electrophoresis in two dimensions, they can be used with paper chromatography or with native PAGE. With paper chromatography the movement away from the diagonal of the appropriate peptides is not in a consistent direction; the direction away from the diagonal is determined by the solvent system used and the effects produced by the altered charge of the various peptides.

### *Amino Acid Analysis*

The basic principles of amino acid analysis were described earlier and are not reiterated here. Several points need to be made, however, concerning the effects produced by various procedures that may be used during the generation or isolation of peptide fragments.

The sulfur-containing amino acids are oxidized during acid hydrolysis, and performic acid oxidation prior to acid hydrolysis gives stable derivatives of methionine, cysteine, or cystine. Methionine is converted to methionine sulfone, and cysteine or cystine are converted to cysteic acid.

A number of proteins contain derivatives of the normal amino acids, and two problems can arise from acid hydrolysis used to break down the peptide bonds in such fragments. If a fragment contains phosphoserine or $\gamma$-carboxyglutamic acid, the derivative is converted to the parent amino acid during acid hydrolysis. The second problem arises from the fact that some amino acid derivatives, such as 3-methyl histidine and $\varepsilon$-*N*-methyl lysine, although not broken down during acid hydrolysis, are poorly resolved from their parent amino acids. These derivatives behave very similarly to histidine and lysine, respectively, in most systems used to separate amino acids during analysis.

These problems have been overcome in a variety of ways. Analysis of $\gamma$-carboxyglutamate involves alkaline hydrolysis of the protein or peptide (2 *M* KOH, 110°C for 24 hours) prior to ion-exchange chromatography of the hydrolysate. Although $\gamma$-carboxyglutamic acid elutes quite differently from glutamate, it can coelute with other ninhydrin-positive material, making quantitation a problem. This is overcome by independently quantitating the amount of $\gamma$-carboxyglutamic acid: The $\gamma$ proton readily exchanges in titrated water at *slightly* acid pH, allowing specific titration of the $\gamma$-carboxyglutamic acid in a protein. The proton is stable at pH 8 and above, and provided that subsequent steps are performed above pH 8, the specific activity of the titrated water allows quantitation. An alternative method is based on the altered electrophoretic mobility that a $\gamma$-carboxyglutamic acid–containing peptide has after exposure to acid conditions (50 m*M* HCl) and heat, which converts the acid to glutamate.

Resolution of methylated amino acids from their parent amino acids has been achieved by altering the temperature of the ion-exchange columns used in analysis. Figure 5-11 shows the separation of various standards, together with the separations achieved during analysis of hydrolysates of myosin containing mono- and trimethyl lysine and methyl histidine in small amounts.

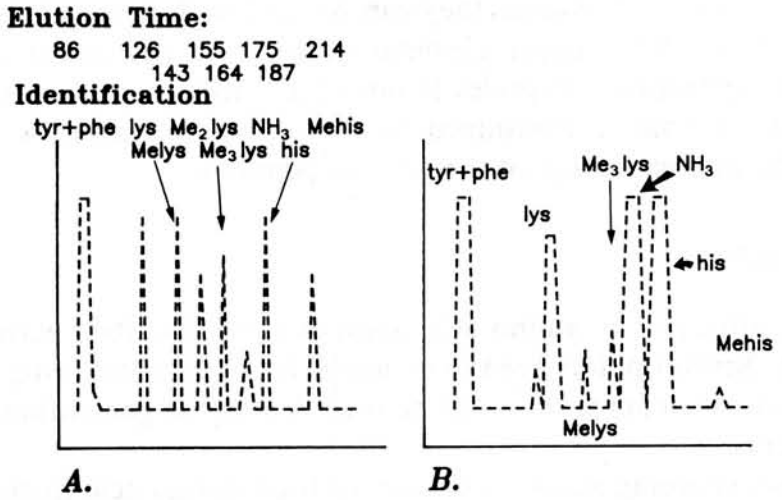



**Figure 5-11** (A) Outline of the results expected for an HPLC separation of various standard amino acids and methylated derivatives of lysine and histidine by ion-exchange chromatography; (B) separation of hydrolyzates of myosin, indicating the presence of methylated lysine and histidine residues.

### *End-Group Analysis*

The final step in the characterization of a peptide prior to sequencing is the determination of the end groups (C terminal and N terminal). In many cases one of the two terminal groups will have been determined by the cleavage method; for example, all tryptic fragments from a protein will have C-terminal lysine or arginine (with the exception of the C-terminal protein fragment). Terminal residue determination serves two functions: (1) it can establish purity, and (2) it can in some instances be sufficient to indicate which fragment has been isolated if the sequence of the protein is shown. Before discussing these two points in more detail, we will first examine some of the procedures commonly used to determine terminal residues.

*Amino Terminus.* There are four commonly used methods; all involve a chemical modification specific for free amino groups followed by hydrolysis and identification of the labeled amino acid. In all four cases some of the derivatives are not particularly stable to acid hydrolysis, and accurate quantitation requires time-course corrections where hydrolysis is carried out for three or four different time periods (usually ranging from 12 to 48 hours) and the analysis extrapolated to $t = 0$. The methods involve labeling with 2,4-dinitrofluorobenzene (DNFB), dansyl chloride, phenylisothiocyanate (PITC), or cyanate, and are shown schematically in Fig. 5-12. With DNFB, dansyl chloride, and PITC, chromophoric derivatives are obtained that are identified by TLC. The sensitivity of these methods can be increased by using radioactive DNFB, dansyl chloride, or PITC. With cyanate labeling, isotopically labeled cyanate is usually used.

With DNFB most amino acid derivatives are quite stable to acid hydrolysis, with the exception of serine and proline, which require time-course corrections. Dansyl proline is somewhat acid labile; however, a number of dansyl derivatives are difficult to resolve by TLC. Dansyl arginine, histidine, and cysteine can present problems. A typical TLC separation of dansyl amino acids is shown in Fig. 5-13. Resolution of some of the closer derivatives can be achieved by using different solvent systems.

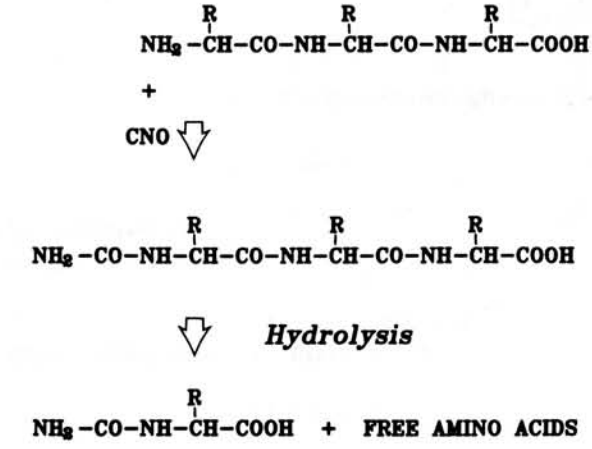



**Figure 5-12A**

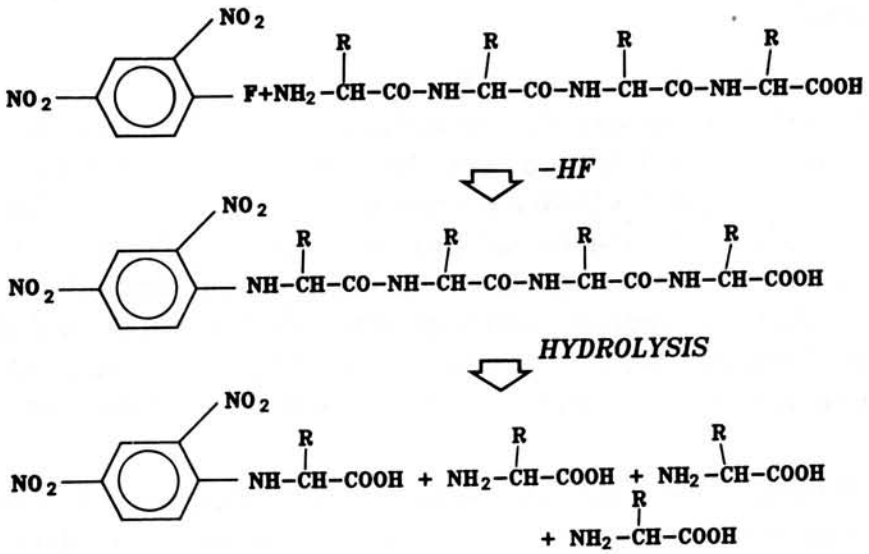



Figure 5-12B

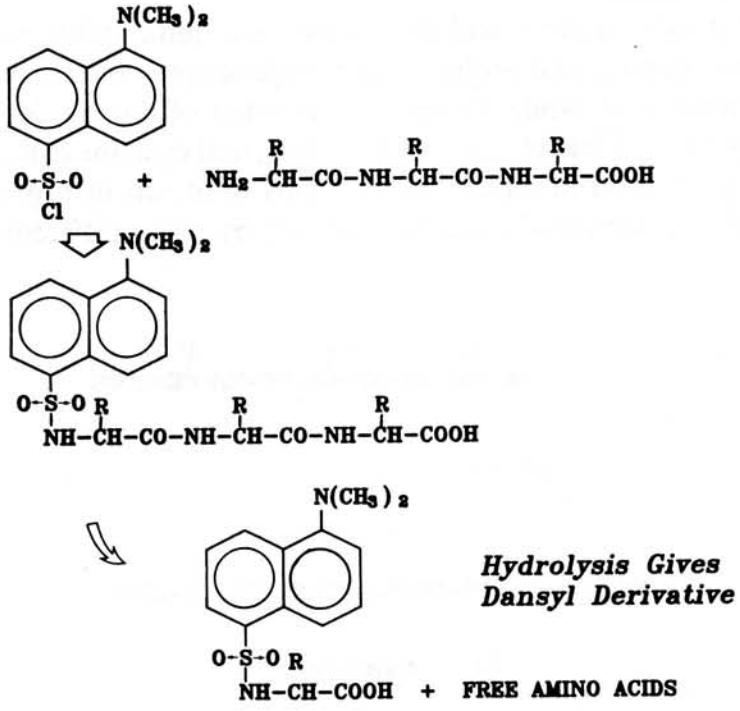



Figure 5-12C

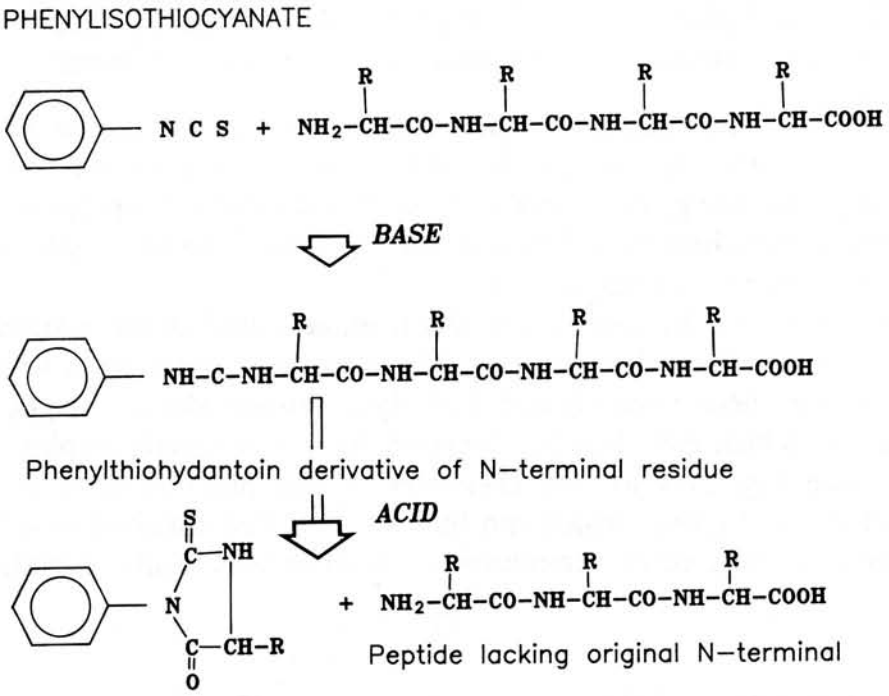



**Figure 5-12D**

**Figure 5-12** Chemistry of amino-terminal labeling by (A) cyanate, (B) 2,4-dinitrofluorobenzene, (C) dansyl chloride, and (D) phenylisothiocyanate.

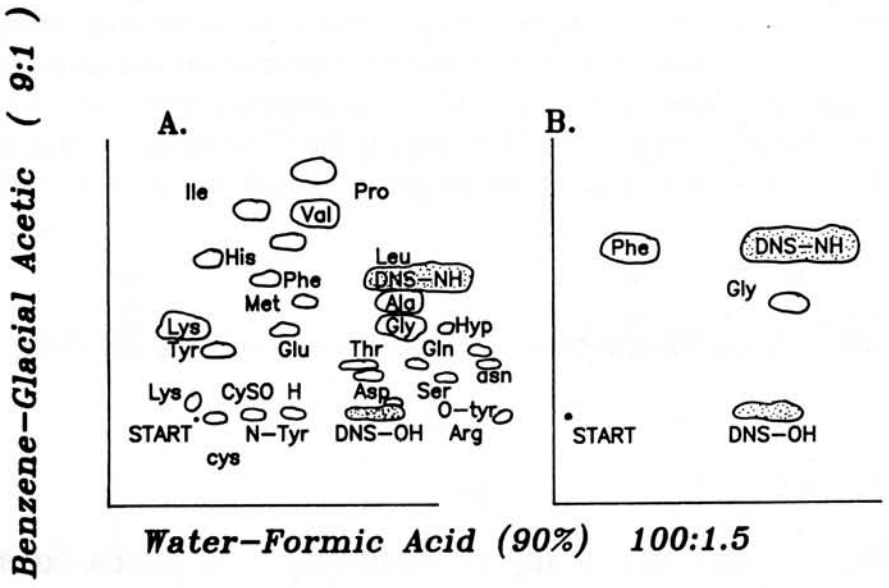


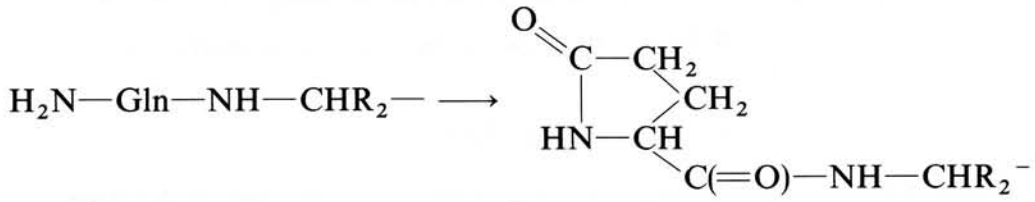

**Figure 5-13** (A) Schematic separation of dansyl amino acids by two-dimensional thin-layer chromatography; (B) identification of the amino-terminal residues of insulin by hydrolysis after dansylation. In each case the dotted areas of the chromatograms indicate the by-products of the dansylation.

The cyanate method gives better accuracy than those using chromophore quantitation but needs large time-course corrections for serine and threonine.

*Blocked Amino Terminals.* In a number of proteins an attempt to use one of these methods to detect the amino-terminal residue of a *protein* will be met by failure. A number of proteins (e.g., cytochrome C, ovalbumin, bovine superoxide dismutase, various immunoglobulins) have blocked amino-terminal residues, where the group is blocked by some type of derivatization.

In many instances the blocking results from acylation of the α-amino group to give either acetyl or formyl derivatives. There are two basic ways to identify the blocking group in these cases: (1) acid hydrolysis releases the carboxylic acid of the blocking group, which can then be identified by gas chromatography, and (2) hydrazinolysis (see Fig. 5-11 for the chemistry of this process) gives the hydrazide derivative of the acyl group, which can then be identified chromatographically.

In some cases (e.g., certain immunoglobulins) intramolecular acylation occurs,




(5-3)

giving a 2-pyrrolidone-5-carboxylic acid (PCA) derivative that can be demonstrated by alkaline cleavage of the PCA ring to give glutamic acid.

*Carboxyl Terminus.* Two chemical approaches to identifying the C-terminal residue of a peptide are often used: hydrazinolysis and tritium labeling. Hydrazinolysis, outlined in Fig. 5-14, leads to the hydrazide derivative of all the carboxyl functions originally in amide linkage: only the C-terminal residue remains as the free amino acid, which is thus identified. Unfortunately, if the C-terminal residue contains an amide in its side chain (i.e., asparagine or glutamine), it is not converted to its free

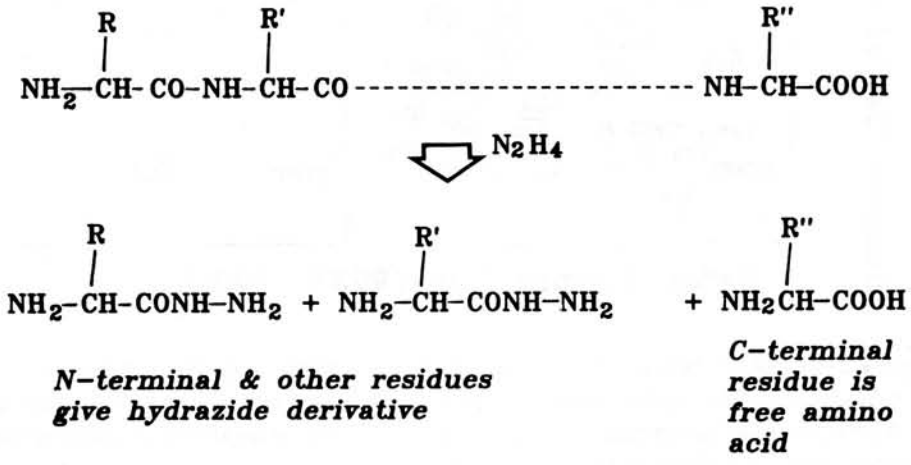



**Figure 5-14** Determination of carbonyl-terminal residue by hydrazinolysis.

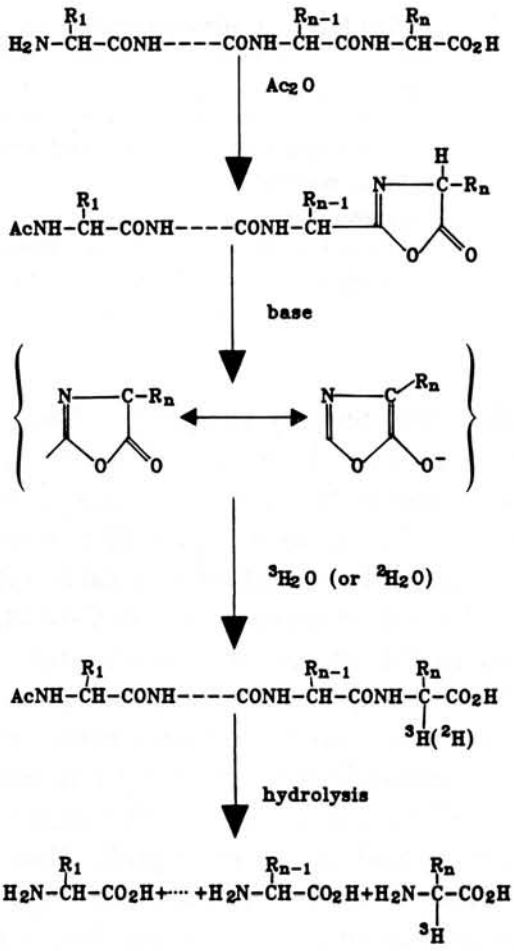



**Figure 5-15** Tritium labeling of the carboxyl-terminal residue of a peptide or protein.

amino acid during hydrazinolysis. In addition, C-terminal arginine or cysteine are fairly unstable under hydrazinolysis conditions.

Tritium labeling of the C-terminal residue is a result of the lability of the hydrogen on the C-terminal residue α-carbon. This hydrogen is made particularly labile by the formation, via intramolecular cyclization of the free COOH with the C=O of the adjacent peptide bond, of an oxazolone (see Fig. 5-15). As a result, a base-catalyzed exchange of this proton occurs in the presence of tritiated water, and after acid hydrolysis the C-terminal residue is identified by its tritium content. Because of its lack of a proton on the α-carbon, C-terminal proline is not labeled in this manner.

In addition to these chemical methods, the C-terminal residue can also be identified enzymatically using carboxypeptidase. There are several types of carboxypeptidase available (see Table 5-2), and the ideal enzyme for use in C-terminal identification would hydrolyze *all* terminal residues with equal facility. Unfortunately, as shown in Table 5-2, this is not the case. Several steps can be taken to help remedy this: The enzymes can be used in combination or denaturants can be employed. The carboxypeptidases are active in denaturants such as dilute SDS or 6 *M* urea, and their presence makes the rate of release of different carboxy terminal residues more

**TABLE 5-2** Carboxypeptidase specificities

| Enzyme | Specificity | Comment |
|---|---|---|
| A | Preference for aromatics and aliphatic hydrophobes | Pro and Arg not released |
| B | Best with Lys and Arg | |
| C | Less specific than A or B | Fairly uniform rate of release |
| Y | Will hydrolyze Pro | Gly and Asp released slowly |

even by breaking secondary and tertiary structures. The pH of the incubation can also affect the rate of release; when lowered sufficiently to suppress the charge on the carboxyl, the rate of release is enhanced. Carboxypeptidase Y, although it releases glycine and aspartate slowly, hydrolyzes all C-terminal residues, including proline. In addition, it is possible to immobilize carboxypeptidase Y in a stable, active form that considerably enhances its utility in C-terminus determination.

Finally, let us return to the somewhat less obvious uses of end-group determination referred to earlier. If, upon end-group determination of a peptide more than one end group is detected, two possibilities must be considered: (1) there is impurity present, and (2) a disulfide-bonded entity has been isolated which in fact contains two peptides joined by a disulfide. In this case reduction and alkylation, followed by repurification, should give two peptides, each with a single terminal residue. Because of the yield variability in many of these detection methods, it is not always easy to distinguish an end group arising from a minor contaminant from that of the "real" peptide.

Where the primary sequence of a protein is known and peptides are being characterized solely for the purpose of identifying which peptide has been isolated, an amino acid analysis and end-group determination are often sufficient to allow identification of a particular peptide. For example, when the peptides have been obtained by a specific fragmentation procedure of a known sequence, the amino acid composition and N-terminal and C-terminal residues of all possible fragments are known. With the smaller fragments it is likely that this information will be unique to a particular peptide. Even with larger peptides, knowledge of the terminal residues and which amino acids may be *missing* from a peptide's composition may allow identification of the peptide.

## DETERMINATION OF THE AMINO ACID SEQUENCE OF FRAGMENTS

The phenylisothiocyanate method of determining the amino terminal of a peptide is the basis of the most popular method of sequencing a peptide. Reexamination of Fig. 5-12 will show that once the amino acid has been released by mild acid hydrolysis,

the remaining peptide is left intact. It is thus a simple procedure to rederivatize the remaining peptide with PITC and repeat the process used for identifying the original N-terminal residue. With each cycle a new N terminal is generated that can subsequently be identified; this is the basis of sequencing a peptide by *Edman degradation.* Because of the yields at each step, the cycle cannot be carried on indefinitely. Using manual procedures with reasonable amounts of starting material, it is possible to go through up to about 15 cycles. With the automated methods available, using immobilized peptides, it is not unusual to sequence 50 to 100 amino acids.

A particularly useful alternative to the Edman system utilizes the same general approach to PITC but involves dimethylaminoazobenzene isothiocyanate (DABITC), which reacts with amino groups to give a highly fluorescent reagent that is readily identified by TLC after release from the peptide. The release involves anhydrous trifluoracetic acid, which does not cleave remaining peptide bonds. After the extraction, the residual peptide (in the aqueous phase) is dried and can be subjected to a further cycle of derivatization. The whole process requires no specialized equipment and is sensitive enough to permit several rounds of identification on nanomole amounts of material.

When the amino acid sequence of prospective peptides is known from the primary structure of the protein, the DABITC method allows for rapid sequence determination of short regions of a peptide and thus identification of the peptide.

An alternative approach involves the use of dipeptidyl peptidases. A variety of these peptidases are available, which split dipeptides either from the amino terminus (DAPs) or from the carboxyl terminus (DCPs). The principle of the method is quite straightforward:

*Step 1.* Digestion of the peptide with the dipeptidyl peptidase is performed (DAP or DCP).

*Step 2.* The dipeptides produced are identified: usually by derivatization to their trimethylsilyl derivatives and identification by gas chromatography or mass spectrometry.

*Step 3.* The original peptide is modified, either by the addition of one amino acid (usually with DAP) or by the removal of an amino acid (usually with DCP).

*Step 4.* Steps 1 and 2 are repeated.

*Step 5.* The original C-terminal and N-terminal residues of the peptide are determined.

*Step 6.* The dipeptides obtained from the native peptide and the modified peptide are listed separately, and the dipeptides are aligned by alternately picking from one set and then the other.

The N-terminal dipeptide is known and is used as the starting point. First let us consider the sequence

Leu-Lys-Cys-Met-Arg-Glu-Thr-Leu-Phe-Val-Ala-Leu

Aminodipeptidase cleavage of this peptide gives the following dipeptides:

Group A:

Leu-Lys
Cys-Met
Arg-Glu
Thr-Leu
Phe-Val
Ala-Leu

After modification of the N-terminal by addition of radioactive glycine and subsequent aminodipeptidase cleavage, we get:

Group B:

| | |
|---|---|
| 14C-Gly-Leu | (therefore, N terminal) |
| Lys-Cys | |
| Met-Arg | |
| Glu-Thr | |
| Leu-Phe | |
| Val-Ala | |
| Leu | (therefore, C terminal) |

Starting with the N-terminal dipeptide from group B,

Gly-Leu

we alternate picks between groups A and B. The next dipeptide must be Leu-Lys, giving (Gly-Lys) (Leu-Lys). With this particular peptide the picks from each table are unequivocal, and the sequence of the starting peptide is easily obtained.

However, consider the related sequence

Leu-Lys-Cys-Met-Arg-Glu-Leu-Ala-Val-Ala-Leu

The original dipeptidase (DAP) gives:

Group A:

Leu-Lys
Cys-Met
Arg-Glu
Leu-Ala
Val-Ala
Leu (must be C terminal)

Modification by addition of an amino acid to N-terminal end (e.g., glycine) gives

Gly-Leu-Lys-Cys-Met-Arg-Glu-Leu-Ala-Val-Ala-Leu

Subsequent DAP treatment gives:

Group B:

Gly-Leu
Lys-Cys
Met-Arg
Glu-Leu
Ala-Val
Ala-Leu

Knowing that the Gly-Leu dipeptide from group B is the N-terminal dipeptide, we start from there, as shown in Fig. 5-16.

After all the dipeptides have been used, we find that two sequences satisfy the pattern and a unique sequence cannot be assigned. If, however, we use a C-terminal

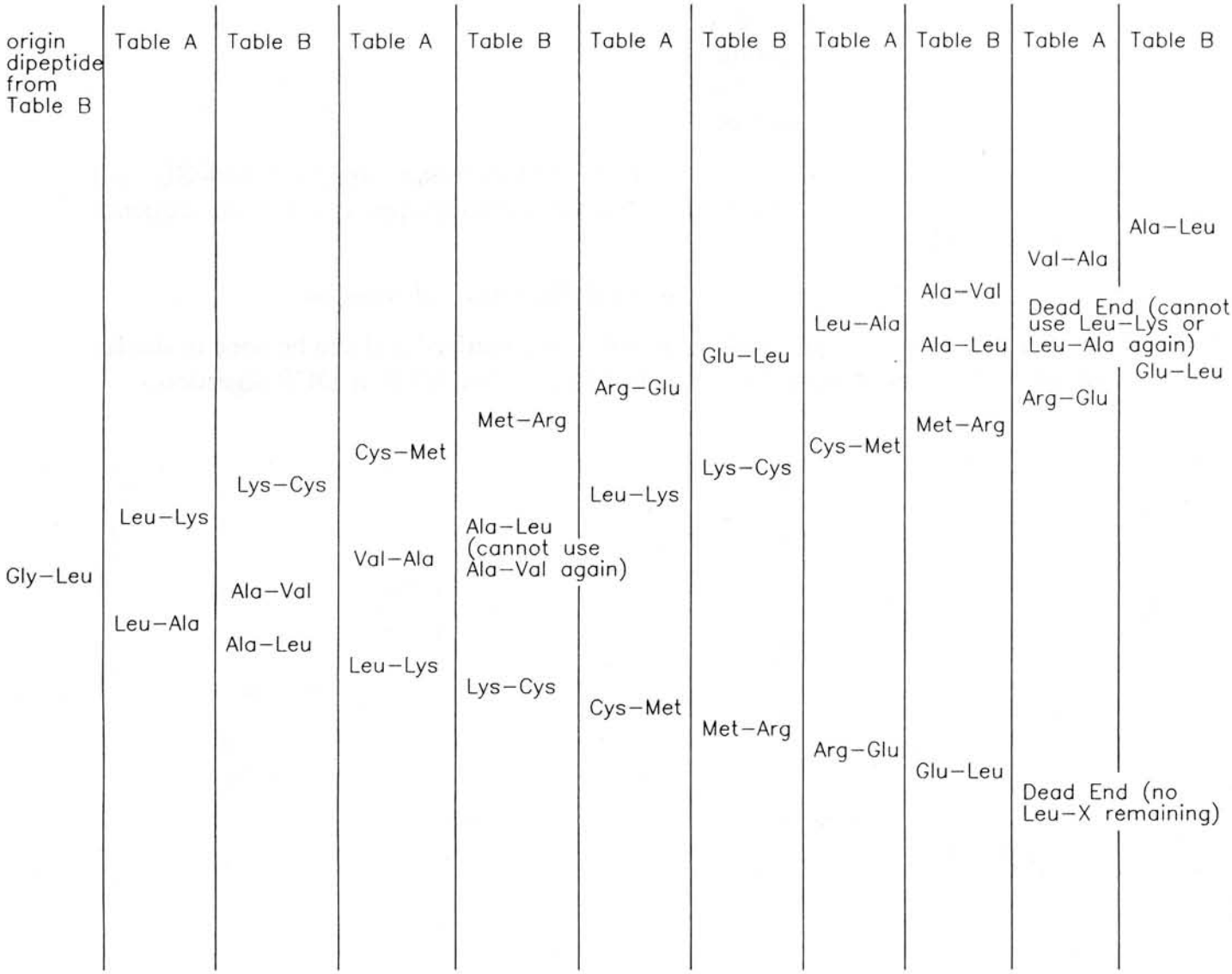



**Figure 5-16** Alignment of dipeptides obtained from diaminopeptidase treatment.

dipeptidase before and after addition of radioactive glycine to the C-terminal end, we get the following dipeptides:

Group A:

Ala-Leu
Ala-Val
Glu-Leu
Met-Arg
Lys-Cys
Leu (therefore, must be N terminal)

Group B:

Leu-Gly (therefore, C terminal)
Val-Ala
Leu-Ala
Arg-Glu
Cys-Met
Leu-Lys

This time (Fig. 5-17) we can start from the C-terminal dipeptide, Leu-Gly, and add dipeptides as before. From this scheme it is readily apparent that the sequence of the peptide is

Leu-Lys-Cys-Met-Arg-Glu-Leu-Ala-Val-Ala-Leu

The aligning of the dipeptides is readily computerized and can be done in similar fashion with tables of dipeptides obtained from either DAP or DCP digestion.

| B | A | B | A | B | A | B | A | B | A | B |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Dead End | | | |
| | | | | | | | | Leu–Ala | | |
| | | | | | | | | | Ala–Leu | |
| | | | | | | | Ala–Val | Val–Ala | | Leu–Gly |
| | | | | | | Leu–Ala | | | Glu–Leu | |
| | | | | | Glu–Leu | | | Arg–Glu | | |
| | | | | Arg–Glu | | | Met–Arg | | | |
| | | | Met–Arg | | | Cys–Met | | | | |
| | | Cys–Met | | | Lys–Cys | | | | | |
| | Lys–Cys | | | Leu–Lys | | | | | | |
| Leu–Lys | | | Ala–Leu | | | | | | | |
| | | Leu–Ala | | | | | | | | |
| | Dead End | | | | | | | | | |

**Figure 5-17** Alignment of dipeptides obtained from dicarboxypeptidase treatment.

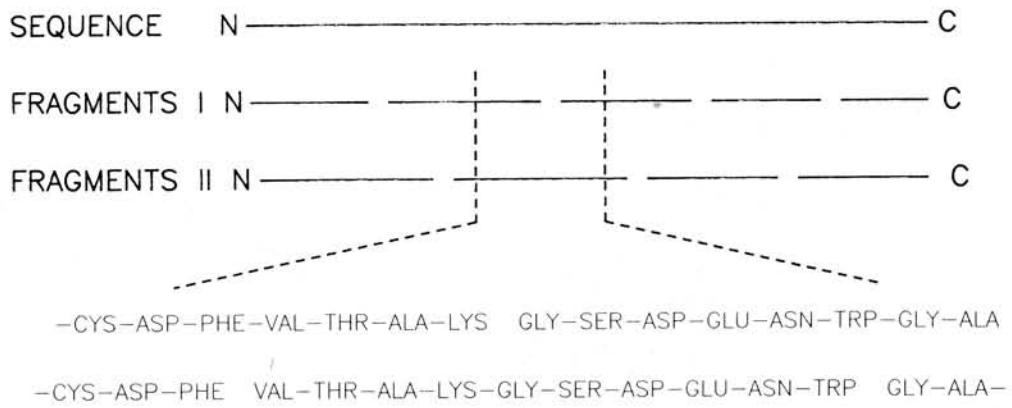



**Figure 5-18** Alignment of sequenced fragments by use of overlapping fragment. Fragments I are obtained by tryptic cleavage of the original peptide, while fragments II are obtained by chymotryptic cleavage.

## ALIGNMENT OF THE FRAGMENTS

Once all possible peptides from a particular fragmentation procedure have been isolated and sequenced, the question remains as to the order in which the fragments appear in the primary sequence of the protein. Usually, the original N-terminal and C-terminal peptides can be determined by labeling the particular terminal residues *prior* to fragmentation so that these two fragments are readily assigned. It is not possible, however, to align the remaining fragments (assuming that there is more than one) without further work. The process (shown in Fig. 5-18) requires a second fragmentation method to be employed and the resultant fragments purified and sequenced. The sequences of this second set are then used to align the fragments obtained in the first fragmentation.

## ASSIGNMENT OF DISULFIDE BONDS

Consider a protein with the structure shown schematically in Fig. 5-19. Isolation of tryptic fragments from this protein results in two rather interesting fragments in addition to a number of "normal" fragments. A fragment with a single amino and carboxyl terminal is obtained (fragment I in the diagram), and a second fragment (II) with two amino and two carboxyl terminals per mole of fragment is also found. Consider first the characteristics of fragment I. This fragment, prior to reduction, has a single —SH group; however, amino acid analysis of the reduced and carboxymethylated fragment indicates the presence of three cysteine residues and thus the existence of a disulfide bond. Sequencing shows the positions of the three cysteines but not which two are involved in the disulfide. This ambiguity is overcome by chemically labeling free sulfhydryl groups, for example, by carboxymethylating the peptide *prior* to reduction with C-14 iodoacetamide. After reduction, C-12 iodoacetamide is used prior to sequencing to prevent reformation of disulfide bonds.

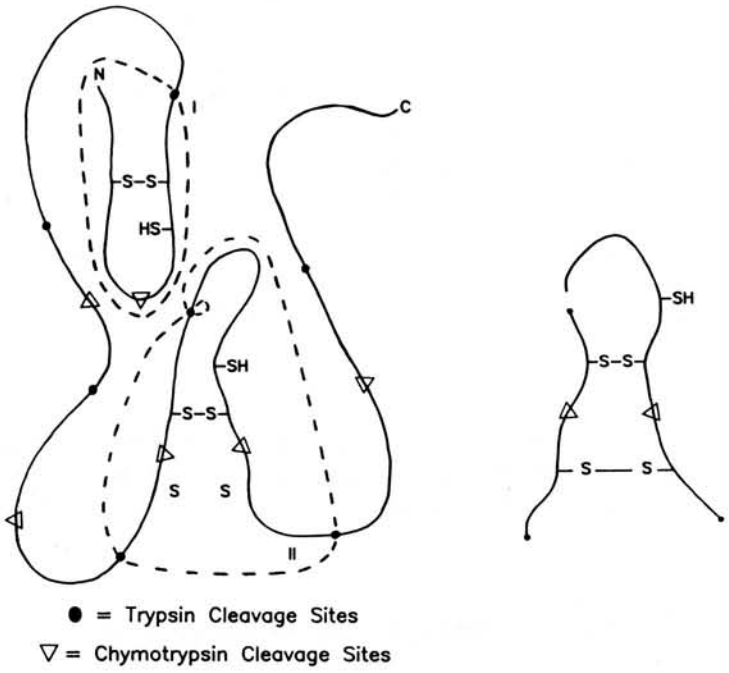



**Figure 5-19** Cleavage and isolation required in the identification of disulfide-bonded residues. Inset at right shows detail of tryptic fragment II.

With fragment II other problems are encountered. The fragment as isolated has two amino and two carboxyl terminal residues, indicating the presence of two polypeptides covalently linked by disulfides. Determination of the free and total cysteine content shows that two disulfides are present. As before, the free cysteine can be radioactively labeled prior to the two chains being separated. Each chain on sequencing has two unlabeled cysteines, each of which is involved in a disulfide bond; the question that remains is, which is bonded to which? In the absence of additional information, this cannot be known. However, if tryptic fragment II is subsequently digested with chymotrypsin *prior* to reduction of the disulfides, two further fragments are obtained (see Fig. 5-19), each of which contains a disulfide bond that can be analyzed as already outlined and allow the unique assignment of the disulfides.

It is entirely possible, of course, that a second cleavage site that allows separation of the disulfides in this manner may not exist. In such a case ambiguity remains concerning the exact partners in each disulfide bond.

## OTHER ASSIGNMENTS

Four other types of assignment with regard to the covalent structure of a polypeptide chain may be necessary. These involve the position of amides and the location of covalently attached groups such as carbohydrate chains, lipids, or phosphate groups.

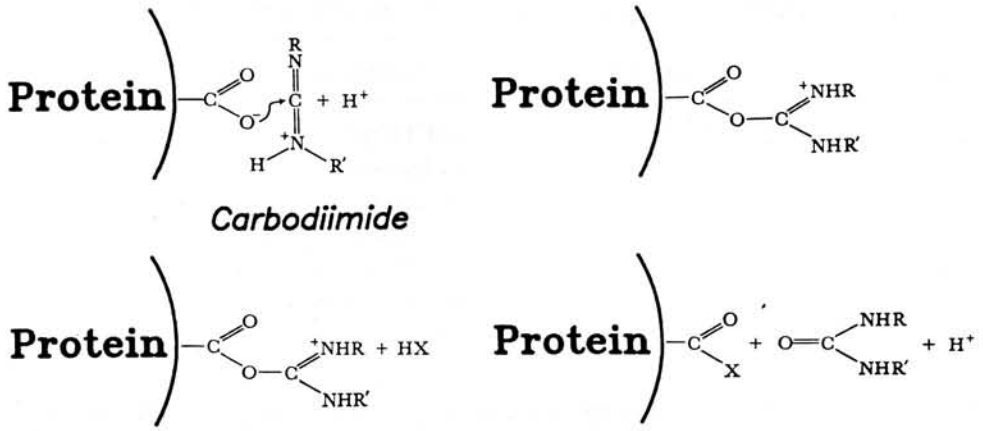



**Figure 5-20** Carbodiimide activation of free carboxyl groups to allow labeling with radioactive nucleophile.

### *Amides*

There are three means used to establish whether an amide (Gln or Asn) is at a particular position or whether the free acid (Glu or Asp) occurs. Where ambiguity remains, the residue is often listed as Asx or Glx. Where the peptide has not been exposed to conditions that might convert the amide to the free acid, it is possible to separate the phenylthiohydantoin derivatives of the amide from the free acid by TLC or by gas chromatography during Edman sequencing. The second approach involves the electrophoretic mobility of isolated peptides that contain a single Asx or Glx residue, which usually indicates whether the amide or the free acid is involved. If the Asx or Glx is close to either the amino or carboxyl terminus, the use of the appropriate exopeptidase removes the ambiguous residue and a comparison of the electrophoretic mobility before and after exopeptidase treatment distinguishes between Asn or Gln and Asp or Glu. In cases where a peptide contains multiple Asx or Glx, it is necessary to use further fragmentation methods to produce peptides, each of which contain a single ambiguous residue, before this method can be used.

The final method depends on the covalent derivatization of free carboxyl groups via the carbodiimide–glycine methyl ester coupling method, which is outlined in Fig. 5-20. This approach gives radioactive derivatives of the free carboxyls, which are then readily distinguished from the amides even if the amides are subsequently converted to the free carboxyls.

### *Carbohydrates, Lipids, and Phosphate*

These covalent derivatives are usually linked to quite unique amino acid side chains: carbohydrates through asparagine or serine (N-linked and O-linked can be distinguished on the basis of their acid lability: O-linked is acid labile), fatty acids via Ser, Thr, or Cys residues via ester linkages, and phosphate groups via serine (though tyrosine and histidine derivatives can occur). Identification of the exact residue involved requires extensive degradation and isolation of the derivatized

**TABLE 5-3** Specific sequences for covalent substitution

| Substitution | Specific sequences |
|---|---|
| N-linked CHO | Asn(CHO)-X-Ser |
| O-linked CHO | Not known |
| Fatty acids and Lipids | Not known |
| Phosphate | Lys(Arg)-Ser-Asn-Ser($PO_4$), Lys(Arg)-Arg-Ala-Ser($PO_4$), Arg-Thr-Leu-Ser($PO_4$) |

peptide such that only a single copy of the appropriate linkage amino acid remains in the peptide. Assignment is then done by inference. In certain cases these covalent substituents are associated with a specific sequence that may be identified in the primary structure, and thus represent potential sites for the appropriate derivatization. Some of these specific sequences are summarized in Table 5-3.

## *ALTERNATIVE STRATEGY TO SEQUENCING A PROTEIN*

As indicated at the outset of this chapter, the protein chemist can resort to the tools of the molecular biologist in order to obtain a complete linear amino acid sequence. Sequencing of a specific DNA fragment is extremely rapid and relatively inexpensive. Consequently, once cloned in a suitable vector and isolated, the sequence of a gene is readily determined, and from this the primary sequence can be derived. However, this approach is complicated by the fact that within a given nucleotide sequence, there can exist more then one potential reading frame. A partial protein sequence of only five to six amino acid residues, perhaps from a chymotrypic peptide, can quickly serve to orient the direction and reading frame for transcription and translation of the protein of interest. Uncertainty regarding the site of translational intitation is readily overcome with an *N*-terminal analysis of the protein by standard techniques. Problems exists if a pre-protein is proteolytically processed to yield a mature protein, and in this case N-terminal analysis of the precursor protein is necessary to determine the actual translational start site.

The first stage in obtaining a nucleotide sequence requires that the gene of interest be cloned, identified, and isolated using recombinant DNA techniques. There are three basic approaches to obtaining the desired gene and many variations on them have been successfully utilized. They are: (1) selection by complementation, (2) detection with antibody, and (3) the oligonucleotide approach.

Selection by complementation involves a direct screening for the activity of the desired gene product via expression in a host lacking the activity (complementation). This approach has been very useful for obtaining genes from prokaryotic sources and some lower eukaryotes, but in general is not applicable to higher eukaryotes. It usually requires efficient expression of the gene in order to work, and low-level expression is often difficult to detect, thereby limiting the sensitivity of this approach.

The second method involves the use of antibody either to detect clones producing the protein of interest or to obtain (or enrich for) nucleic acid that is actively being translated. The detection of clones producing the desired protein requires that part or all of the gene be accurately transcribed and translated. Clones producing the protein of interest are detected after lysis and immobilization of macromolecules on a filter membrane, usually nitrocellulose. Antibody is added under conditions that allow specific binding, and positive reactions are disclosed by addition of either radiolabeled staphylococcal A protein or a secondary labeled or enzyme-conjugated antibody directed against the primary antibody. Positive clones are then isolated and propagated from a master plate. This technique is extremely sensitive, and depending on the affinity of the antisera can easily detect nanogram quantities of a gene protein.

Specific antibody is also useful in significantly enriching for the mRNA encoding the desired protein. In vitro, one can obtain translation of mRNA from either eukaryotic or prokaryotic sources. Selective precipitation of mRNA–ribosome–protein complexes (polysomes) can be done with antibody and the mRNA deproteinized. In this way, nucleic acid encoding the gene of interest can be significantly purified or enriched.

The oligonucleotide approach for detecting or identifying specific cloned sequences is perhaps the most sensitive and most powerful. It does not require transcription or translation of the cloned gene and is readily adaptable to screen either eukaryotic cDNA or genomic libraries as well as gene libraries from prokaryotic sources. In this method a partial protein sequence is needed. From this, one can derive a number of possible nucleotide sequences. Since the triplet code is degenerate, it is advantageous if the protein sequence used contains amino acids that are coded for by the least number of triplets. Particularly suitable are methionine and tryptophan, since each has a single triplet codon. Phenylalanine, tyrosine, histidine, glutamine, asparagine, lysine, aspartate, glutamate, and cysteine are also desirable since each has just two triplet codons. Serine should be avoided, if possible, since it shows degeneracy at both the 2 and 3 positions in the triplet. The goal of this approach is to synthesize a probe oligonucleotide which will hybridize strongly to the appropriate segment of DNA, and maximum complementarity will be obtained with the least degenerate probe possible. Mixed oligonucleotide probes are frequently used in instances where ambiguity exists so that a complementary sequence will be represented in the probe mixture. These probes are generally end-labeled with $^{32}P$ and used to screen clones after lysis and immobilization on a filter membrane.

### *Cloning Hosts and Vectors*

The gram-negative bacterium *Escherichia coli* is the most readily employed host for cloned genes from all sources for a variety of reasons: It is easily grown and maintained, and foreign DNA is easily introduced into this organism by transformation, transfection, or infection with bacteriophage. DNA is stably maintained and readily recovered, and a plethora of "genetic tricks" exist that allow the investigator to manipulate cloned DNA efficiently and inexpensively. Other prokaryotes, yeast, and

mammalian cells are also used as hosts for cloned DNA, but it is desirable to do most DNA manipulation using *E. coli*, for the reasons already given.

There are a huge variety of cloning vectors, which vary in their host range, in the amount of DNA that can be accommodated, and in the regulatory functions contained on the vector. Vectors exist which are designed for the generation of deletions in DNA, for DNA sequencing, or for the expression of foreign DNA. In general, there are four classes of vectors: plasmids, viral (prokaryotic and eukaryotic), cosmids, and integration vectors.

Plasmid cloning vectors are small, autonomously replicating elements that usually contain a marker for selection, such as an antibiotic-resistance determinant, as well as a mechanism to detect if foreign DNA is inserted. The pUC plasmids developed in Joachim Messing's laboratories are perfect examples. The ampicillin-resistance gene allows positive selection of transformants, and cloning into a number of unique restriction sites destroys $\beta$-galactosidase activity, so that clones containing recombinant plasmids can be differentiated on appropriate medium. Plasmids can accommodate a wide range of sizes of DNA fragments, from a few base pairs to several kilo-base pairs (kbp).

*E. coli* bacteriophage vectors are extremely useful cloning vectors and have played a central role in molecular biology. Bacteriophage lambda has been widely used for cloning cDNA or genomic DNA from all sources. Varying amounts of a nonessential region of the phage DNA can be exchanged for foreign DNA, and the recombinant DNA molecule can be packaged into viable phage particles in vitro. The recombinant phage can be propagated in an appropriate *E. coli* host. Up to 23 kbp of DNA can be cloned in some vectors.

The single-stranded filamentous *E. coli* phage, $m_{13}$, has been developed by Messing for use in the Sanger dideoxy chain-termination DNA-sequencing system. Foreign DNA is inserted into the double-stranded replicative form of the phage at a polylinker region containing a number of unique restriction endonuclease sites. High yields of single-stranded DNA necessary for the Sanger sequencing method are easily obtained after transfection with the recombinant replicative form and propagation.

Eukaryotic viral vectors have been very useful in studies on gene regulation and development. They allow the study of gene function and DNA sequences and have already been useful in examining sequence–function relationships among some of the oncogenes.

Cosmids are hybrids of plasmids and bacteriophages capable of autonomous replication and can be recognized and packaged into phage particles in vitro. They are useful for cloning very large DNA fragments (approximately 38 kbp).

Insertion vectors have been developed for both prokaryotes and eukaryotes. They have no replication origin for a particular host or a replication origin with a conditional mutation that allows inactivation of replication functions, generally by a temperature shift. These vectors facilitate the stable integration of foreign DNA into the host's chromosome. These types of vectors are very powerful tools in regulation studies.

Of all the cloning vectors, plasmids and single-stranded bacteriophage can be used directly in experiments where one wants to create specific sequence alterations in a nucleic acid coding region.

### *Cloning Strategies*

Genes from most prokaryotes and some eukaryotes do not contain introns and can therefore be cloned directly from partial or complete digestion of chromosomal DNA into any of the vectors described previously. If the desired gene is expressed, clones synthesizing the desired gene product can be detected by screening for activity, or by using antibody against the gene product, as already described. In the absence of sufficiently strong expression, the oligonucleotide probe approach is a feasible alternative.

Genes containing introns can be cloned directly, but more often than not they are not transcribed; and even with transcription, prokaryotes cannot properly process mRNA for translation. It is therefore desirable to isolate a copy of the gene representative of the mature mRNA. This can be done by isolating mature, spliced, polyadenylated mRNA from cells that are actively producing the gene product of interest. Because most mature eukaryotic mRNAs are polyadenylated at their 3′ end, oligo-dT can be employed to prime the message for reverse transcription. In the presence of dNTPs, reverse transcriptase will synthesize a copy of DNA from the mRNA. DNA polymerase can then be used to complete the population of double-stranded copy DNA. Synthetic linkers can be inserted into a compatible restriction site in a plasmid or phage vector to construct a cDNA "bank" or "library".

In instances in which the mRNA of interest is present as a significant portion of total mRNA, it is sufficient to prepare a cDNA from whole mRNA. In this case, the desired clone should represent a significant amount of the cDNA library and detection should be possible. For very low abundance message, the mRNA encoding the desired clone can be significantly enriched by polysome precipitation using specific antisera, as described previously. Then a cDNA library can be prepared from the precipitated mRNA, thereby increasing the representation of the clone of interest in the total population. Another ingenious technique has been used to clone low-abundance mRNAs which are cell-type specific. In this "subtraction cloning" protocol, mRNA from a population of mRNAs containing the desired message is allowed to hybridize on filters with a cDNA library from cells that do not make the desired mRNA. Message that does not hybridize is washed off the filter and collected. A DNA library can then be constructed from this RNA. This method "subtracts" the mRNAs common to both populations and facilitates recovery of cell-type-specific message. It has been successfully employed in obtaining genes expressed only in T cells by subtracting T-cell mRNA with a cDNA bank from B cells.

The detection of a successful cloning event from among the many clones generated when a cDNA or genomic library is constructed is often a difficult and labor-intensive task. Two methods are available and widely employed. A clone *can* be detected in a

cDNA library by the use of antibody as described previously, but because eukaryotic promotor and translational initiation sites are not usually recognized by *E. coli*, a prokaryotic version of these regulatory sequences must be provided. For example, the popular bacteriophage vector λGT11 allows the expression of in-frame fusions of the *E. coli* β-galactosidase gene and foreign proteins. Plasmid vectors containing λ or *E. coli* regulatory sequences are also available.

The most sensitive method for identifying recombinants from cDNA libraries cloned in either phage or plasmid vector is the use of mixed oligonucleotide probes, as described previously. Expression of the gene is not required and therefore this method is also useful for identifying genomic clones. Specific oligonucleotides which are complementary to the desired gene can also be employed to prime synthesis of a cDNA library, so that the majority, if not all, of the cDNA thus generated originated from the sequences related to the desired genes. Also, oligonucleotides can be used to prime synthesis from a specific mRNA in the presence of radiolabeled precursors, and this transcript can be used directly to probe a cDNA library.

Another very useful strategy for detecting cloned genes is to use clones of genes from other species as radiolabeled probes. For example, porcine factor VIII was used to detect the human gene, and mammalian *ras* oncogene has been used to identify and clone a related protein from yeast.

### *Sequencing the Gene*

The success of sequencing a gene depends on the experimental separation of single-stranded DNAs having different lengths by polyacrylamide gel electrophoresis in the presence of urea. Two basic techniques have been developed for the generation of the required sets of fragments, one dependent on random termination of the labeled oligonucleotide during synthesis in the presence of [$^{32}$P]dNTP, and the other dependent on random chemical cleavage after labeling the 5′ end with $^{32}$P. The Sanger method is an enzymatic approach relying on the ability of dideoxynucleoside triphosphates to function as efficient nascent DNA chain terminators. The Maxam–Gilbert method involves the random chemical cleavage of end-labeled DNA fragments. Polyacrylamide gels with between 8 and 12% acrylamide give good separation of oligonucleotide fragments between 10 and 300 nucleotides in length generated in the aforementioned sequencing methods. Since oligonucleotides differing in length by a single nucleotide can readily be separated in this manner, it is possible to sequence an oligonucleotide if some way of differentially identifying the terminal residue of each fragment is available. If a separate series of fragments, each terminating in A, C, G, or T, can be generated and separately electrophoresed, it is a simple matter (Fig. 5-21) to read off the nucleotide sequence. The urea prevents the formation of base pairing, which leads to double-stranded DNA or secondary structure that could result in anomalous migration.

*Sanger Method.* This method requires a single-stranded DNA template of the gene of interest. In general, this is obtained by cloning into the replicative form of

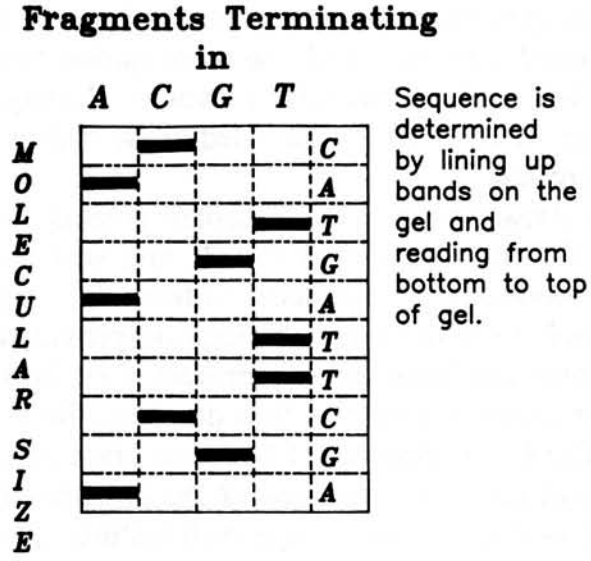



**Figure 5-21** Outline of PAGE method of sequencing an oligonucleotide.

the single-stranded $\phi m_{13}$ bacteriophage by transfecting an appropriate host, and allowing the phage machinery to produce particles carrying the single-stranded DNA. A short oligonucleotide (15 to 22 kbp) complementary to a region upstream of the DNA to be sequenced is used to prime DNA synthesis. The 79-kDa fragment of *E. coli* polymerase I, the Klenow fragment, containing the DNA-polymerizing activity is used to synthesize a complementary DNA strand. Reaction mixes containing the single-stranded DNA primer, all four dNTPs [$\alpha^{32}$P]dNTP labeled, and one of four dideoxynucleotides (ddNTP) at appropriate concentrations and the Klenow fragment are constructed. The ddNTPs are inserted randomly by the polymerase, resulting in termination of DNA synthesis. In this way a random pool of different-size labeled fragments which all terminate at a specific nucleotide are generated. The double-stranded DNA can be melted and applied to a urea gel and the sequences read off as previously. This method is rapid and efficient. About 300 to 400 bases can be read from a single gel. By cloning and sequencing different portions of a gene in different orientations, one can quickly obtain an entire gene sequence. Alternatively, a single bacteriophage carrying the entire gene can be sequenced by synthesizing primers for different regions of the DNA.

*Maxam–Gilbert Method.* In an alternative approach, the piece of DNA to be sequenced is first labeled at the 5′ end by using polynucleotide kinase and $^{32}$P. The radioactive DNA is then subjected to various chemical cleavages with differing base specificity to generate randomly cleaved fragments, which are then analyzed similar to the dideoxynucleoside triphosphate approach. Because of the chemical similarity of the purines it is not possible to obtain unique cleavage at either A or G; however, after methylation of the DNA with dimethylsulfate, heating at neutral pH preferentially cleaves at G, while incubation with acid leads to preferential cleavage at A. As a result, alternating light and dark bands are seen on the autoradiograph, the

dark bands corresponding to sites of preferential cleavage and the light bands corresponding to the dark bands obtained with the other purine cleavage pattern. Treatment with hydrazine (followed by piperidine) leads to cleavage (with equal facility) at cytosine and thymine. If 2 *M* NaCl is included in the hydrazine reaction, cleavage at the thymine is suppressed.

These four chemical reactions give four sets of fragments: with cleavage at G > A, with cleavage at A > G, with cleavage at C + T, and with cleavage at C > T. As before, the nucleotide sequence is easily deduced from the gel patterns.

Although these tools from molecular biology do provide an easy approach for obtaining the linear amino acid sequence of a protein, there is much essential description of the covalent structure of a protein that demands the classical approaches of the protein chemist. The DNA sequence of a protein gives no information regarding covalent modifications of particular residues, or the location of disulfide bonds, and a complete description of proteins containing such features depends on the isolation and characterization of peptide regions containing such modifications. The aid to these tasks comes from the fact that if the complete linear sequence of a protein is known, it is easier to identify particular fragments once they have been purified. It is even possible in some instances to uniquely identify, for example, tryptic fragments of a protein on the basis of their amino acid composition. In most cases, however, when the linear sequence is known, several cycles of Edman degradation are sufficient to completely assign a particular fragment to its location in the sequence. It is fair to say that the development of gene cloning and sequencing techniques have aided but not replaced, the classical aims of protein chemistry.