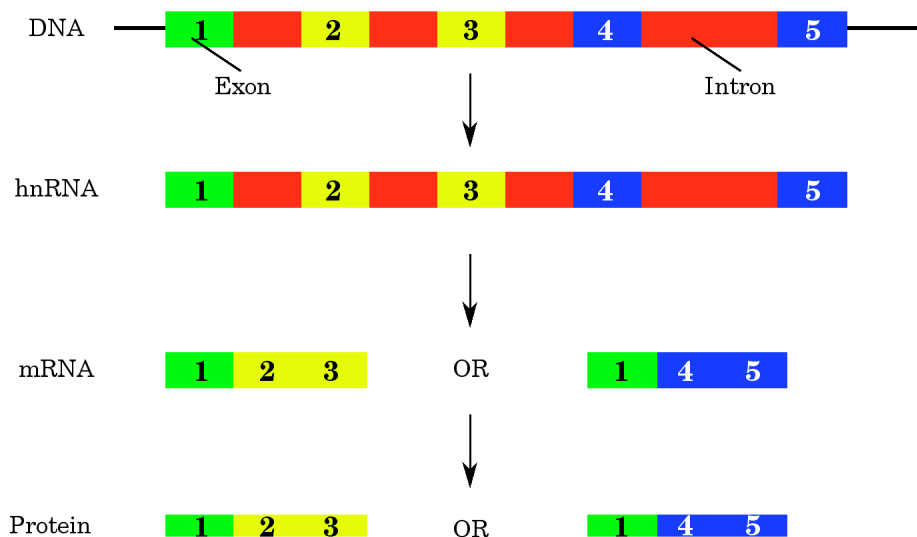


## RNA and Translation

### hnRNA

Following transcription, eukaryotes must modify their transcripts prior to making protein. The initial transcript is called **heterogeneous nuclear RNA**. The hnRNA must undergo a maturation process prior to becoming mRNA. The processing usually involves the removal of introns, and the addition of a poly (A) tail.

Note that removal of introns can mean alternative splicing; although eukaryotic transcripts do not contain the multiple coding sequences present in polycistronic prokaryotic messages produced from operons, a eukaryotic cell can vary the final mRNA by altering the splicing pattern, and therefore the final exon distribution of the mRNA. (The order of the exons is fixed by their sequence in the DNA, but the identity of the exons that become a part of the final mRNA is not entirely controlled by the DNA sequence.) In the example below, exon 1 is present in both alternately spliced mRNA molecules, but the remaining exons translated to produce the final protein are either 2 and 3 or 4 and 5, depending on how the RNA is spliced during the maturation process.



### mRNA structure

#### 5'-Cap

Eukaryotic mRNA molecules usually have a GTP attached to the 5' end of the 5'-most base of the RNA by *guanylyl transferase*. The guanine base is then usually methylated. The presence of the methyl guanosine residue increases the efficiency of translation initiation, and is thought to decrease enzymatic degradation of the RNA.

#### Poly(A) tail

Eukaryotic mRNA usually have a series of 100 to 200 adenosine residues added the 3' end. These are not included in the DNA sequence; instead these are added by a template-independent enzyme: *poly(A) polymerase*. As with the cap, the poly(A) tail is thought to stabilize the mRNA.

Of all common mRNAs, only those that code for histones lack poly(A) tails. The poly(A) tail is an extremely useful property for molecular biology, because it creates a common feature that allows eukaryotic mRNA to be readily purified. In normal cells, mRNA molecules comprise about 10% of the total cellular RNA; purification of mRNA avoids problems due to the presence of much greater amounts of rRNA.

Prokaryotes do not add poly(A) tails to their mRNA; it is therefore difficult to purify mRNA from prokaryotic organisms.

### 5' and 3' untranslated regions

Messenger RNA molecules, and especially eukaryotic mRNA molecules, contain significant sequences upstream and downstream of the actual coding regions. These untranslated regions (UTR) seem to have two major functions: 1) UTRs contain sequences that control translation, and 2) UTRs confer increased or reduced stability to the mRNA.

### Protein coding region

All mRNAs contain a region that actually codes for amino acids.

### Riboproteins

Most mRNAs have proteins associated with them; this is especially true for eukaryotic mRNAs. These proteins probably act to alter the stability of the message and to assist in regulating translation initiation for the mRNA.

## Genetic code

### Codons

The genetic code requires groups of three nucleotides to specify each amino acid. RNA has 4 possible nucleotides (A,C,G,U); groups of three bases can therefore result in  $4^3 = 64$  possible codons. The vast majority of organisms need only 21 different “letters” (20 amino acids and a stop signal) in their genetic code. The result is that the genetic code contains redundancy – several amino acids have more than one codon. The genetic code is shown below.

First Position	Second Position								Third Position
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	<b>Stop</b>	UGA	<b>Stop</b>	A
	UUG	Leu	UCG	Ser	UAG	<b>Stop</b>	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	<b>Met</b>	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Examination of the genetic code reveals that most of the redundancy in the code is found in the last base of the codon (*e.g.*, glycine is GGN, and proline is CCN, where the N indicates any of the four possible bases).

### **Universality**

The genetic code is essentially universal. With minor exceptions, all organisms use exactly the same genetic code. The major exceptions are mitochondria, in which a few of the codons have different meanings (*e.g.*, four differences from the standard code exist in the code used by mammalian mitochondria).

### **Reading frame**

The fact that it takes three bases to specify an amino acid means that the coding region can be read in more than one way, with only one of these ways being correct. This means that the protein synthesis machinery must have some mechanism for determining the correct reading frame. Because the choice of reading frame is made only at the initiation of translation, DNA mutations that insert or remove a nucleotide (**frame-shift mutations**) change the meaning of the entire sequence after that point.

### **Start codon**

Prokaryotic, and to a lesser extent, eukaryotic mRNA molecules, contain more than one signal for the beginning of protein synthesis. The most important signal is the start codon. Nearly all genes use the AUG methionine codon. (A few genes use GUG (one of the valine codons); note that the only difference between AUG and GUG is the substitution of one purine for another). This means that, as synthesized, most proteins begin with methionine.

### **Stop codons**

In contrast to the start codon, the stop codons do not code for amino acids; instead they are signals that cause a halt in protein synthesis. Most organisms use three stop codons: UAA, UAG, and UGA.

### **tRNA Functions**

All tRNA molecules are fairly small (~75 bases) single-stranded nucleic acids used as adapter molecules. The protein synthesis machinery uses tRNAs as carriers of amino acids.

### **tRNA Synthesis**

tRNAs are synthesized as larger precursor RNAs. These are then post-transcriptionally modified to form the mature tRNA. One modification is the cleavage of the precursor to release the actual tRNA (cleavage at the 5' end is performed by RNase P, a riboprotein complex, in which the RNA portion contains the catalytic activity). Other modifications include alterations of some of the bases (deamination of some adenines to form inosine, methylation, formation of pseudouridine). The modifications also include the addition of a CCA sequence to the 3' end of the tRNA.

### CCA acceptor

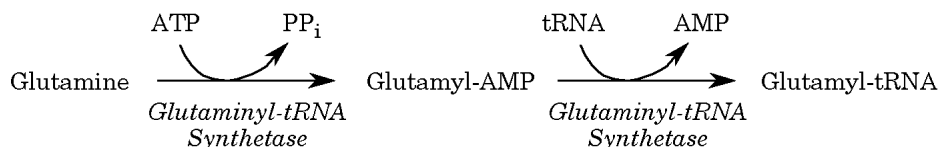
Two parts of the tRNA are specific to that particular tRNA. One of these is the structure formed at the 5' and 3' ends (which are in close proximity in three dimensional space). This structure must have a CCA sequence at the 3' end in order to accept amino acids.

### Anticodon

The other major specific part of the tRNA is the anticodon. The anticodon is located in the center of the tRNA sequence. The anticodon is somewhat flexible, especially in the first base of the anticodon (which corresponds to the last base of the codon). This means that the last base is the least important for establishing specificity of codon recognition. This also means that fewer than 64 different tRNA molecules are required to allow full expression of the genetic code.

### Aminoacyl-tRNA synthetase

Each amino acid has at least one enzyme that specifically charges the corresponding tRNA molecule. The charging of the tRNA is a two-step process. Because the aminoacyl-tRNA synthetase has to bind to the amino acid twice, it has two chances to select the correct amino acid, allowing discrimination between closely related amino acids (such as glutamine and asparagine). Each step has an error rate of 1 in 100; this results in a 1:10,000 overall error.



Most aminoacyl-tRNA synthetases recognize a stretch of about 10 bases in the tRNA. The Alanine tRNA synthetase is an exception; it recognizes a single G-U pair in the Ala-tRNA amino acid arm.

### Codon usage

Different codons are used at different frequencies in different organisms. This can be important in expressing foreign proteins in a given organism.

### Ribosome function

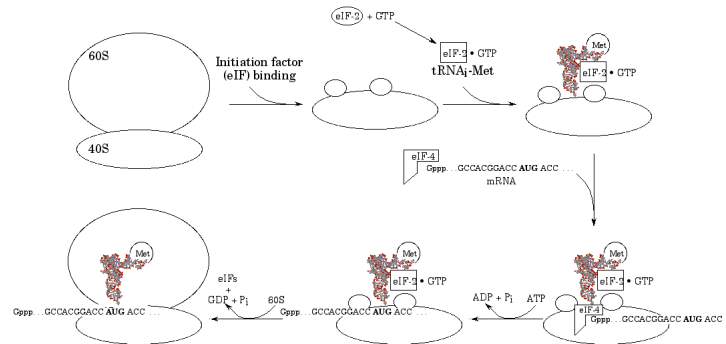
Protein synthesis occurs using a 5' to 3' readout of the mRNA, resulting in peptide synthesis beginning at the N-terminus.

The general features of prokaryotic and eukaryotic translation are similar. The following describes eukaryotic translation; most of this information also applies to prokaryotic translation.

### Initiation

The first required step is the initiation of translation. The initiation process begins when initiation factors (eIF = eukaryotic initiation factors) bind to the ribosome,

and induce separation of the 40S and 60S subunits. eIF-2 binds GTP and associates with a specific initiator methionine tRNA charged with methionine; this complex then binds to the 40S subunit. eIF-2 acts as a regulatory point; phosphorylation of eIF-2 prevents initiation of translation.



eIF-4 binds to the 5'-cap of the mRNA to be translated. Following binding of the charged methionine tRNA, the eIF-4 complex then to the 40S subunit. The eIF-4 complex is the second regulatory point; phosphorylation of eIF-4 (by different kinases from those that phosphorylate eIF-2) stimulates translation initiation. Alternatively, inhibitory proteins can bind to parts of the eIF-4 complex and prevent initiation.

The mRNA must then be translocated (in an ATP-dependent process) to place the AUG start codon in the correct position. The AUG used as the start codon is determined by sequences near the AUG; the sequences are called Shine-Dalgarno sequences in *E. coli*. The Shine-Dalgarno site is complementary to the 3'-end of the 16S rRNA. (The sequence is 3'AUUCCUCCAC). Initiation of translation in prokaryotes requires a stretch of 4 to 9 bases of the mRNA that will base pair with the 16S rRNA.

In eukaryotes, the sequences near the AUG appear to be less important; usually the first AUG in the mRNA is used to initiate translation.

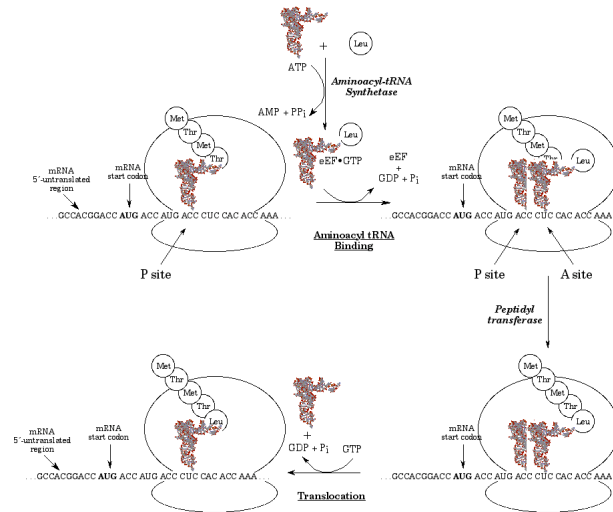
After the translocation of the mRNA, the 60S subunit binds, and allows translation to occur.

## Elongation

Elongation is a spiral process. The growing peptide chain is bound to a tRNA in the **P site** (= peptide site) of the ribosome. (Note that the Met-tRNA that binds the start codon is the only tRNA that binds the P site first.)

The appropriate aminoacyl-tRNA (determined by the codon) binds to the **A site** (=acceptor site) of the ribosome (this binding requires GTP hydrolysis). The *peptidyltransferase* activity of the ribosome then forms a peptide bond (the energy required is contained in the aminoacyl-tRNA, which required 2 ATP equivalents for synthesis). The *peptidyltransferase* activity is thought to be mediated by the rRNA. The mRNA/tRNA/peptide is translocated in a GTP-dependent process, moving the tRNA to the P site, and exposing an empty A site. The old tRNA dissociates, and the ribosome is ready for the next tRNA.

Elongation requires 4 ATP equivalents for each amino acid (two for the charging of the tRNA, one for tRNA binding, and one for translocation). This energy is required both to form the peptide bonds and to ensure that the information from the mRNA is accurately translated into protein sequence.



## Termination

Stop codons usually do not have tRNA; instead, a releasing factor binds to the ribosome, and expends one GTP in hydrolyzing the peptide-tRNA bond.

## Polysomes

Ribosomes are large particles. On the other hand, mRNA is a long molecule; it is therefore possible for multiple ribosomes to bind (and translate) the same mRNA at the same time. Steric constraints prevent ribosomes binding closer than 80-100 bases apart on the same mRNA.

In prokaryotes, translation can begin before transcription is complete.

## Protein synthesis rates

*E. coli* in log phase growth contain about 15,000 ribosomes: ~25% of the total cell weight.

Prokaryotic protein synthesis occurs at ~20 amino acids per second. Each ribosome can only synthesize one protein at a time; however, multiple ribosomes can synthesize protein from the same message. Eukaryotic protein synthesis is considerably slower, and occurs at a rate of between two and four amino acids per second.

## Protein synthesis and infections

### Antibiotics

Although all ribosomes are related, some species differences do exist. These differences are more significant between eukaryotic ribosomes and the smaller and simpler bacterial ribosomes.

The critical role of protein synthesis has led to the evolution of a wide variety of defensive molecules that bind to and inhibit ribosomes of other species while not binding to the ribosomes of the host organism. These compounds are useful both as probes of ribosome function and as therapeutic antibiotics.

**Puromycin** binds the A site of both eukaryotic and prokaryotic ribosomes, and usually results in early chain termination, because it structurally resembles the 3'-end of a tRNA but cannot engage in translocation. **Cycloheximide** is a eukaryotic peptidyltransferase inhibitor. **Tetracycline** binds to the A site of bacterial ribosomes. **Streptomycin** inhibits initiation and causes misreading of the mRNA. **Chloramphenicol** inhibits bacterial and mitochondrial ribosomes by inhibiting peptidyltransferase activity. Tetracycline, streptomycin, and chloramphenicol have been used as antibiotic drugs due to their ability to inhibit bacterial protein synthesis and, therefore, bacterial growth.

### **Toxins**

**Diphtheria toxin** is a gene product of a bacteriophage that infects some bacteria. The protein contains an enzymatic activity that covalently modifies eEF-2 and therefore inhibits protein synthesis. **Ricin toxin** is a protein made by castor beans; it removes a single adenine from the 23S rRNA, a modification that is sufficient to inactivate the ribosome. Since both diphtheria toxin and ricin toxin are enzymes, single molecules can be enough to kill cells.

### **Viruses**

Viruses must use the cellular machinery to produce their proteins. Viruses use several methods.

Some viruses use strong promoters that produce large amounts of mRNA; these viral mRNA molecules then simply compete numerically with the cellular mRNA. Some viral RNA molecules bind to the ribosomes with higher affinity than do the cellular mRNA molecules, which allows preferential translation of the viral mRNA. Finally, some viruses produce proteins that inactivate (or stimulate cellular inactivation of) initiation factors. These viruses usually contain initiation factor substitutes that can then use the ribosomes without interference.

## **Post-translation modifications**

### **Protein folding**

Folding is a post-translational process (although it may begin during translation). Protein folding is a complicated, incompletely understood process. Some proteins fold without assistance; others require the assistance of other proteins to allow proper folding.

### **Covalent modifications**

#### **N-terminal methionine**

The vast majority of proteins begin with methionine (or N-formyl-methionine in prokaryotes). A specific *methionine aminopeptidase* removes the methionine from some newly synthesized proteins (usually if the second amino acid is small).

## **N-acetylation**

The addition of an acetyl group (derived from acetyl-CoA) is a common modification of eukaryotic proteins (~70% of cytosolic proteins are N-acetylated). N-acetylation removes the positive charge at the N-terminus.

## **Signal peptides**

### Secreted/Membrane proteins

Proteins destined for secretion (and some plasma membrane-bound proteins) have leader sequences. Leader sequences are short (~30 amino acid) sequences, usually at the N-terminus, that direct the protein to their extracellular or cell surface destination.

As the leader is translated, it binds to the **Signal Recognition Particle** (SRP), which inhibits further translation. The SRP-ribosome complex then binds to the endoplasmic reticulum (the presence of many of these results in the formation of “rough” endoplasmic reticulum).

Binding to the endoplasmic reticulum allows resumption of translation; during synthesis, the new polypeptide is fed across the membrane. As part of this process, the signal peptide is removed, and protein folding occurs.

### Hormones

Many peptide hormones are synthesized as larger precursors (**preprohormones**). The “pre” sequence is the export leader. The prohormone (following cleavage of the leader) is further processed inside the secretory vesicle to generate the mature hormone.

### Mitochondrial proteins

Most mitochondrial proteins are synthesized from genes found in the nucleus. These genes code for mitochondrial import leader sequences. Unlike secreted proteins, mitochondrial proteins are completely translated in the cytoplasm. However, in most cases, mitochondrial proteins are not folded until after translocation into the mitochondrion. The translocation and folding process usually involves removal of the leader.

### Bacterial translocation

Most bacterial membrane, periplasmic space, and secreted proteins use methods similar to the mitochondrial system, in which the protein is synthesized first, is translocated, and then the N-terminal signal sequence is removed. A few proteins, however, use a method similar to the SRP system for eukaryotes.

## **Lipid conjugation**

Proteins that are intended to become membrane associated are often conjugated to lipids. These include N-myristoylation (conjugation of the protein N-terminus to the fatty acid myristate) and prenylation (conjugation of isoprene derivatives, such as farnesyl or geranylgeranyl, to residues near the C-terminus of the protein). These hydrophobic modifications then act as membrane anchors, allowing hydrophilic proteins to become tightly associated with the membrane.

## **Phosphorylation**

Phosphorylation is a common phenomenon. The most common sites of phosphorylation are on amino acid side chains that contain hydroxyl groups, although other residues can also be phosphorylated. Unlike some of the other post-translational modifications, phosphorylation is a reversible process. Phosphorylation of proteins frequently acts as a functional switch that turns a protein on or off.

## **Methylation**

A number of proteins are methylated (in an S-adenosyl methionine dependent process), usually on lysine, arginine, histidine, glutamate, or aspartate residues. This was originally thought to be permanent modification, but more recent studies suggest that it is reversible, and may be involved in regulation of the protein function in a manner similar to phosphorylation.

The role of methylation is not clearly understood, and probably differs significantly between proteins.

## **Zymogens**

Some proteins must remain inactive until needed. This is especially true for proteases, which, when uncontrolled, can severely damage cells. Such proteins are often synthesized as larger inactive precursors. Activation of the protein occurs as a result of proteolytic cleavage.

## **Glycosylation**

Most plasma membrane proteins, and many secreted proteins, are glycosylated. The carbohydrate is added to the protein in the endoplasmic reticulum and Golgi apparatus.

The carbohydrate is either N-linked (attached to the amide group of asparagine residues) or O-linked (attached to serine or threonine hydroxyls).

In general, N-linked carbohydrate groups consist of larger, more complex chains of carbohydrates. The carbohydrate can be a significant part of the protein, comprising 50% or more of the molecular weight. N-linked carbohydrates are synthesized on dolichol pyrophosphate, and then transferred co-translationally to protein. The carbohydrate groups are then modified by removal or addition of monosaccharide units.

Glycosylation sites are probably specified by the three-dimensional structure of protein. This is especially true for O-linked carbohydrates, which do not seem to have any sequence signals. While N-linked carbohydrates are attached to the asparagine of Asn-X-Ser/Thr sequence, this attachment only occurs at specific sites, which again suggests an additional signal inherent in the three-dimensional structure of the protein.

The carbohydrate groups are frequently heterogeneous, with different protein molecules containing at least slightly different modifications.

The carbohydrate portion of proteins is thought to assist in maintaining the proper conformation.

The carbohydrate can also act as a signal for processing. This is especially the case for asialoglycoproteins. The enzyme neuraminidase removes terminal sialic acid residues, and then the liver asialoglycoprotein receptor removes the protein from circulation.

### **Disulfide bonds**

The cysteine sulfhydryl side-chain can interact with other cysteine side chains. This can either link different parts of the same peptide, or can link different peptides together. Disulfide bonds are relatively rare in the reducing environment inside cells, but are frequently found in cell-surface and secreted proteins.

### **Collagen**

Collagen is an extremely abundant protein, comprising about 25% of the total protein in animals. The majority of collagen consists of a triple helix consisting of three protein chains wound around one another.

The sequence of the helical portion is  $(\text{Gly-X-Y})_n$  with X often being proline, and Y often being either hydroxyproline or hydroxylysine. Some hydroxylysine residues are O-glycosylated (a structure unique to collagen).

The enzymes proline hydroxylase and lysine hydroxylase both require ascorbic acid and  $\alpha$ -ketoglutarate as cofactors, and only perform their reaction on collagen. Inhibition of collagen synthesis is responsible for some of the symptoms of scurvy.

Collagen is synthesized as procollagen; the N- and C-termini are globular protein-like domains that allow proper formation of the triple helix; the termini are cleaved as part of the post-translational modification process.

## **Non-covalent modifications**

### **Multimeric protein assembly**

Many proteins contain more than one polypeptide. In some cases, the assembly of the complex occurs spontaneously as a result of high affinity non-covalent interactions between the polypeptides. In other cases, some assistance may be necessary. The process of multimeric complex assembly is poorly understood.

### **Cofactor binding**

Many proteins contain tightly bound or covalently bound co-factors (*e.g.*, heme, iron-sulfur centers, pyridoxal phosphate, biotin, FMN, or FAD). These are added post-translationally.

## **Protein turnover**

### **Protein half-lives**

Protein half-lives vary greatly; some proteins have life-times that are comparable to those of their host cells (such as hemoglobin, which has a half-life of more than the

120 day life-span of the erythrocyte). Other proteins are produced and rapidly degraded.

Although the structural features that control protein life-times are poorly understood, one factor appears to be the amino terminal residue. Met, Ser, Ala, Gly, and Val are associated with long lifetimes.

Arg is associated with very short lifetimes.

### **Lysosomes**

Lysosomes are intracellular organelles that contain a variety of enzymes capable of degrading biological macromolecules. Specificity for degradation is mediated by import into the lysosomes. The lysosomal interior has a low pH (~4 to 5). Most lysosomal enzymes are only active under low pH conditions, which helps protect the cell from damage if the lysosomes are disrupted.

### **Proteasomes**

Proteasomes are large (MW ~1.7 million) barrel-shaped multiprotein (~30 subunits) complexes that degrade proteins in an ATP-dependent reaction. Most proteins are targeted for degradation by conjugation to **ubiquitin**. Ubiquitin is a highly conserved 76 amino acid protein, found in all eukaryotic organisms. (Ornithine decarboxylase, a protein with an extremely short half-life, is an exception.)

A specific class of proteins attaches ubiquitin to lysine residues on the target protein. Additional ubiquitin molecules are then attached to the ubiquitin. The protein is then either degraded by the proteasome or the ubiquitin is removed (in either case, the ubiquitin is recycled).

The signals that induce ubiquitin-conjugation are poorly understood. One feature seems to be “PEST” sequences (regions containing proline, glutamate, serine, and threonine). These stimulate phosphorylation; for a number of proteins, phosphorylation is a signal for degradation. On the other hand, in some cases, phosphorylation prevents ubiquitination. Amino-terminal sequences also seem to be related to protein stability. Proteins with small side chains on the amino terminus, and proteins with methionine at the amino terminus are relatively stable. Proteins with polar or aromatic N-terminal amino acids have short half-lives.

When proteins are being degraded for energy, the ubiquitin-proteasome pathway largely mediates the process.

### **Other pathways**

In addition to the lysosomes, and the proteasome pathway, cells have some specific proteases that cleave one or a few proteins, or which are only active under very specific conditions. For example, proteolytic cleavage of some peptides is part of the maturation process.

**Apoptosis** involves (among other things), activation of some proteases that degrade cellular proteins during the cell suicide pathway.