8. Peptide bonds, polypeptides and proteins

In which we consider the nature of proteins, how they are synthesized, how they are folded and assembled, how they get to where they need to go, how they function, how their activities are regulated, and how mutations can influence their behavior.

We have mentioned proteins many times, since there are few biological processes that do not rely on them. Proteins act as structural elements, signals, regulators, and catalysts in a wide arrange of molecular machines. Up to this point, however, we have not said much about what they are, how they are made, and how they do what they do. The first scientific characterization of what are now known as proteins was published in 1838 by the Dutch chemist, Gerardus Johannes Mulder (1802–1880).[193] After an analysis of a number of different substances, he proposed that they all represented versions of a common chemical core, with the molecular formula $C_{400}H_{620}N_{100}O_{120}P_1S_1$, and that the differences between them were primarily in the numbers of phosphate (P) and sulfur (S) atoms they contained. The name "protein", from the Greek word πρώτα ("prota"), meaning "primary", was suggested by the Swede, Jons Jakob Berzelius (1779–1848) based on the presumed importance of these compounds in biological systems.[194] As you can see, Mulder's molecular formula is not very informative, it tells us little or nothing about protein structure, but suggested that all proteins are fundamentally similar, which is confusing since they carry out so many different roles. Subsequent studies revealed that protein could be dissolved in either water or dilute salt solutions but aggregated and became insoluble when the solution was heated; as we will see this aggregation reflects a change in the structure of the protein. Mulder was able to break down proteins through an acid hydrolysis reaction into amino acids, named because they contained amino ($-NH_2$) and carboxylic acid (-COOH) groups. Twenty different amino acids could be identified in hydrolyzed samples of proteins. Since their original characterization as a general class of compounds, we now understand that while they share a common basic structure, proteins are remarkably diverse. They are involved in roles from the mechanical strengthening of skin to the regulation of genes, to the transport of oxygen, to the capture of energy, to the catalysis and regulation of essentially all of the chemical reactions that occur within cells and organisms.

Polypeptide and protein structure basics

While all proteins have a similar bulk composition, this obscures rather than illuminates their dramatic structural and functional differences. With the introduction of various chemical methods, it was discovered that different proteins were composed of distinct and specific sets

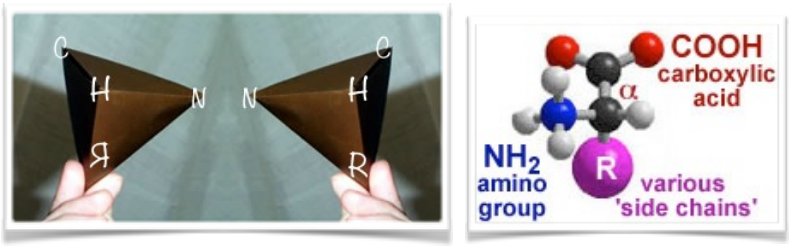
[193] From 'protein' to the beginnings of clinical proteomics: http://www.ncbi.nlm.nih.gov/pubmed/21136729

[194] While historically true, the original claim that proteins get their name from "the ancient Greek sea-god Proteus who, like your typical sea-god, could change shape. The name acknowledges the many different properties and functions of proteins." seems more poetically satisfying to us.

of subunits, and that each subunit is an unbranched polymer of amino acids with a specific sequence. Because the amino acids in these polymers are linked by what are known as peptide bonds, the polymers are known generically as polypeptides. At this point, it is important to reiterate that proteins are functional objects, In addition to polypeptides many proteins also contain other molecular components, known as co-factors or prosthetic groups (we will call them co-factors for simplicity's sake.) These co-factors can range from metal ions to various small molecules.

Amino acid polymers

As you might remember from chemistry, carbon atoms (C) form four bonds, and where these are all single bonds, the basic structure of the atoms bound to a C is tetrahedral. We can think of an amino acid as a (highly) modified form of methane (CH_4), with the C referred to as the alpha carbon (C_α). Instead of four hydrogens attached to the central C, there is one H, an amino group ($-NH_2$), a carboxylic acid group (-COOH), and a final, variable (R) group attached to the central C_α atom. The four groups attached to the α-carbon are arranged at the vertices of a tetrahedron. If all four groups attached to the α-carbon are different from one another, as they are in all amino acids except glycine, the resulting amino acid can exist in two possible stereoisomers, which are known as enantiomers. Enantiomers are mirror images of one another and are termed the L- and D- forms. Only L-type amino acids are found in proteins, even though there is no obvious reason that proteins could not have also been made using both types of amino acids or using only D-amino acids.[195] It appears that the universal use of L-type amino acids in the polypeptides found in biological systems is yet another example of the evolutionary relatedness of organisms, it appears to be a homologous trait. Even though there are hundreds of different amino acids known, only 22 amino acids (these include the 20 common amino acids and two others, selenocysteine and pyrrolysine) are found in proteins.

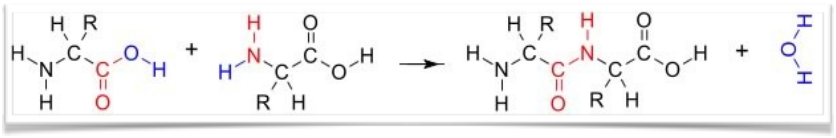


Amino acids differ from one another by their R-groups, which are often referred to as "side-chains". Some of these R-groups are large, some are small, some are hydrophobic, some are hydrophilic, some of the hydrophilic R-groups contain weak acidic or basic groups. The extent to which these weak acidic or basic groups are positively or negatively charged will change in response to environmental pH. Changes in charge will (as we will see) influence the structure of the polypeptide/protein in which they find themselves. The different R-groups provide proteins with a broad range of chemical properties, which are further extended by the presence of co-factors.

As we noted for nucleic acids, a polymer is a chain of subunits, amino acid monomers linked together by peptide bonds. Under the conditions that exist inside the cell, this is a thermodynamically unfavorable dehydration reaction, and so must be coupled to a thermodynamically favorable reaction. A

[195] It is not that D-amino acids do not occur in nature, or in organisms, they do. They are found in biomolecules, such as the antibiotic gramicidin, which is composed of alternating L-and D-type amino acids - however gramicidin is synthesized by a different process than that used to synthesize proteins.

molecule formed from two amino acids, joined together by a peptide bond, is known as a dipeptide. As in the case of each amino acid, the dipeptide has an N-terminal (amino) end and a C-terminal (carboxylic acid) end. To generate a polypeptide, new amino acids are added (exclusively) to the C-terminal end of the polymer. A peptide bond forms between the amino group of the added amino acid and the carboxylic acid group of the polymer. This reaction generates a new C-terminal carboxylic acid group. It is important to note that while some amino acids have a carboxylic acid group as part of their R-groups, new amino acids are not added there. Because of this fact, polypeptides are unbranched, linear polymers. This process of amino acid addition can continue, theoretically without limit. Biological polypeptides range from very short (5-10) to many hundreds (thousands) of amino acids in length. For example, the protein Titin (found in muscle cells) can be more than 30,000 amino acids in length. Because there is no theoretical constraint on which amino acids occur at a particular position within a polypeptide, there is a enormous universe of possible polypeptides that could exist. In the case of a 100 amino acid long polypeptide, there are 20^{100} possible different polypeptides that could be formed.



Specifying a polypeptide's sequence

Perhaps at this point you are asking yourself, if there are so many different possible polypeptides, and there is no inherent bias favoring the addition of one amino acid over another, what determines the sequence of a polypeptide, clearly it is not random. Here we connect to the information stored in DNA. We begin with a description of the process in bacteria and then extend it to archaea and eukaryotes. We introduce them in this order because, while basically similar, the system is simpler in bacteria (although you might find it complex enough for your taste.) Even so, we will leave most of the complexities for subsequent courses. One thing that we will do that is not common is that we will consider the network dynamics of these systems. We will even ask you to do a little analytics, with the goal of enabling you to make plausible predictions about the behavior of these systems, particularly in response to various perturbations. Another important point to keep in mind, one we have made previously, is that the system is continuous. The machinery required for protein synthesis is inherited by the cell, so each new polypeptide is synthesized in an environment full of pre-existing proteins and ongoing metabolic processes.

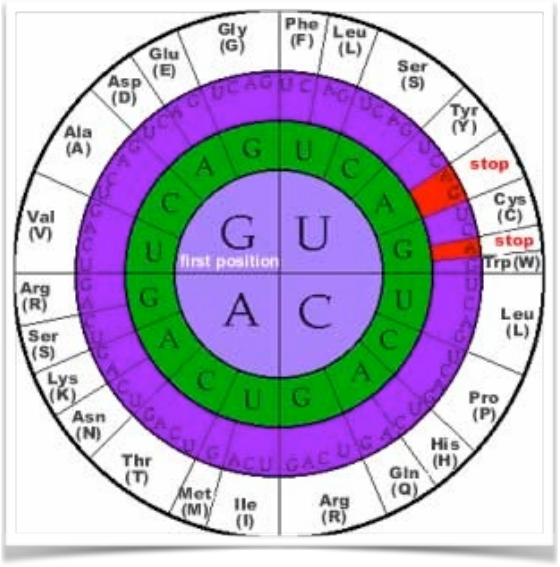
Making a polypeptide in a bacterial cell

A bacterial cell synthesizes thousands of different polypeptides. The sequences of these polypeptides are encoded within the DNA of the organism. The genome of most bacteria is a double-stranded circular DNA molecule that is millions of base pairs in length. Each polypeptide is encoded by a specific region of this DNA molecule. So, our questions are how are specific regions in the DNA recognized and how is nucleic acid-encoded information translated into polypeptide sequence.

To address the first question, thinking back to the structure of DNA, it was immediately obvious that the one-dimensional sequence of a polypeptide could be encoded in the one-dimensional

sequence of the polynucleotide chains in a DNA molecule.[196] The real question was how to translate the language of nucleic acids, which consists of sequences of four different nucleotide bases, into the language of polypeptides, which consists of sequences of the 20 different amino acids. As pointed out by the physicist George Gamow (1904-1968), when he was a professor at UC Boulder, the minimum set of nucleotides needed to encode all 20 amino acids is three; a sequence of one nucleotide (4^1) could encode at most four different animo acids, a two nucleotide length sequence could encode (4^2) or 16 different amino acids (not enough), while a three nucleotide sequence (4^3) could encode 64 different amino acids (more than enough).[197] Although the actual coding scheme that Gamow proposed was wrong, his thinking about coding capacity influenced those who experimentally determined the actual rules of the "genetic code".

The genetic code is not the information itself, but the algorithm by which nucleotide sequences are "read" to determine polypeptide sequences. A polypeptide is encoded by the sequence of nucleotides. This nucleotide sequence is read in groups of three nucleotides, known as a codon. Codons are read in a non-overlapping manner, with no spaces (that is, non-coding nucleotides) between them. Since there are 64 possible codons but only 20 (or 22 - see above) different amino acids used in organisms, the code is redundant, that is, certain amino acids are encoded by more than one codon. In addition, there are three codons, UAA, UAG and UGA that encode "stops"; they do not encode any amino acid but are used to mark the end of a polypeptide. The region of the nucleic acid that encodes a polypeptide begins with what is known as the "start" codon and continues until a stop codon is reached. This sequence is known as an open reading frame or an ORF.



There are a number of hypotheses on the origin of the genetic code. One is the frozen accident model in which the code used in modern cells is the result of an accident, a bottleneck event. Early in the evolution of life on Earth, there may have been multiple types of organisms, using different codes, but the code used reflects the fact that only one of these organisms gave rise to all modern organisms. Alternatively, the code could reflect specific interactions between RNAs and amino acids that played a role in the initial establishment of the code. What is clear is that the code is not necessarily fixed, there are examples in which certain codons are "repurposed" in various organisms. What these variations in the genetic code illustrate is that evolutionary mechanisms can change the genetic code.[198] Since the genetic code does not appear to be predetermined, the

---

[196] Nature of the genetic code finally revealed!: http://www.nature.com/nrmicro/journal/v9/n12/full/nrmicro2707.html

[197] The Big Bang and the genetic code: Gamow, a prankster and physicist, thought of them first: http://www.nature.com/nature/journal/v404/n6777/full/404437a0.html:

[198] The genetic code is nearly optimal for allowing additional information within protein-coding sequences: http://genome.cshlp.org/content/17/4/405 and Stops making sense: translational trade-offs and stop codon reassignment: http://www.ncbi.nlm.nih.gov/pubmed/21801361

general conservation of the genetic code among organisms is seen as strong evidence that all organisms (even the ones with minor variations in their genetic codes) are derived from a single common ancestor. It appears that the genetic code is a homologous trait between organisms.

An important feature of the genetic system is that the information stored in DNA is not used directly to direct polypeptide synthesis. Rather it has to be copied through the formation of an RNA molecule, known as a messenger RNA or mRNA. In contrast to the process involved in the transformation of the information stored in a nucleic acid sequence into a polypeptide sequence, both DNA and RNA use the same nucleotide language. Because of this fact, the process of DNA-directed RNA synthesis is known as ***transcription.*** The process of RNA-directed polypeptide synthesis is known as ***translation***, because the language of nucleic acids is different from the language of polypeptides.

Protein synthesis: transcription (DNA to RNA)

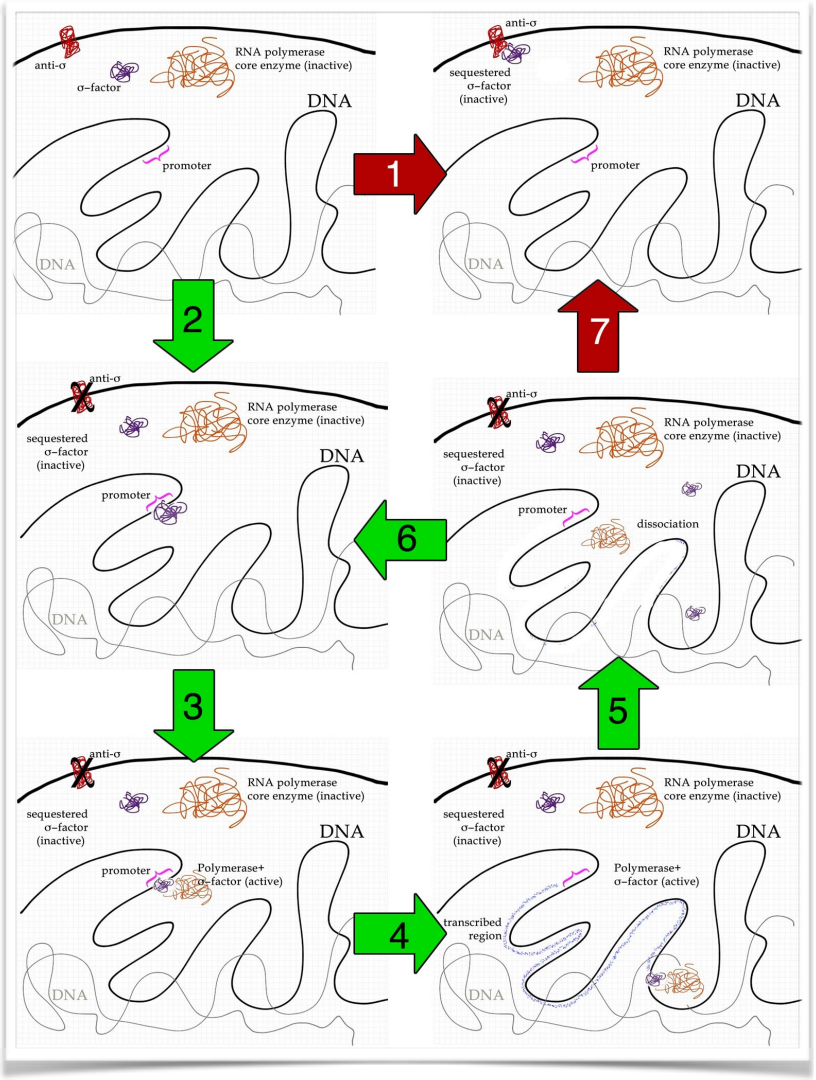
Having introduced the genetic code and RNA, however, briefly, we now return to the process by which a polypeptide is specified by a DNA sequence. Our first task is to understand how it is that we can find the specific region of the DNA molecule that encodes a specific polypeptide, since we are looking for a short region of DNA within millions or in eukaryotes, typically billions of base pairs of sequence). So while the double stranded nature of DNA makes the information stored in it redundant (a fact that makes DNA replication straightforward), the specific nucleotide sequence that will be decoded using the genetic code is present in only one of the two strands. From the point of view of polypeptide sequence the other strand is nonsense.

As we have noted, a gene is the region(s) of a larger DNA molecule. Part of the gene's sequence, its regulatory region, is used (as part of a larger system involving the products of other genes) to specify when, where, and how much the gene is "expressed". Another part of the gene's sequence is used to direct the synthesis of an RNA molecule (the transcribed or coding region). Once a gene's regulatory region is engaged, the synthesis of an RNA molecule is the next step in the expression of the gene. As a general simplification, we will say that a gene is expressed when the RNA that it encodes is synthesized. We can postpone further complexities to later (and subsequent classes). It is important to recognize that an organism as "simple" as a bacterium can contain thousands of genes, and that different sets of genes are used in different environments to produce specific behaviors. In some cases, these behaviors may be mutually antagonistic. For example, a bacterium facing a rapidly drying out environment might turn on genes that allow it to stop growing and dividing, and prepare it to survive in such a hostile environment. That means some genes (involved in active growth and replication) need to be turned off, while others, involved in survival, need to be turned on. Our goal is not to have you accurately predict the particular behavior of an organism, but rather to be able to make plausible predictions about how gene expression will change in response to various perturbations. This requires us to go into some detail about mechanisms, but rather superficially, in order to illustrate a few of the regulatory processes that are active in cells.

So you need to think, what are the molecular components that can recognize a gene's regulatory sequences? The answer is proteins. The class of proteins that do this are known generically

as transcription factors. Their shared property is that they bind with high affinity to specific sequences of nucleotides within DNA molecules. For historical reasons, in bacteria these transcription factor proteins are known as sigma (σ) factors. The next question is how is an RNA made based on a DNA sequence? The answer is DNA-dependent RNA polymerase, which we will refer to as RNA polymerase. In bacteria, groups of genes share regulatory sequences recognized by specific σ factors. As we will see this makes it possible to regulate groups of specific genes in a coordinated manner. Now let us turn to how, exactly (although at low resolution), this is done, first in bacteria and then in eukaryotic cells.

At this point, we need to explicitly recognize common aspects of biological systems. They are highly regulated, adaptive and homeostatic - that is, they can adjust their behavior to changes in their environment (both internal and external) to maintain the living state. These types of behaviors are based on various forms of feedback regulation. In the case of the bacterial gene expression system, there are genes that encode specific σ factors. Which of these genes are expressed determines which σ factor proteins are present and which genes are actively expressed. Of course, the gene encoding a specific σ factor is itself regulated. At the same time, there are other genes that encode what are known as anti-σ factors. One class of anti-σ factors are membrane-associated proteins. For a σ factor to activate a gene, it must be able to bind to the DNA, which it cannot do if it is bound to the anti-σ factor. So a gene may not be expressed (we say that it is "off") because the appropriate σ factor is not expressed or because even though that σ factor is expressed, the relevant anti-σ factor is also expressed, and its presence acts to block the action of the σ factor (arrow 1). We can, however, turn on our target gene if we inactivate the anti-σ factor. Inactivation can involve a number of mechanisms, including the destruction or modification of the anti-σ factor so that it no longer interacts with the σ factor. Once the σ factor is released, it can diffuse through out the cell and bind to its target DNA sequences (arrow 2). Now an inactive RNA polymerase can bind to the DNA-σ factor complex (arrow 3). This activates the RNA polymerase, which initiates DNA-dependent RNA synthesis (arrow 4). Once RNA polymerase has been activated, it will



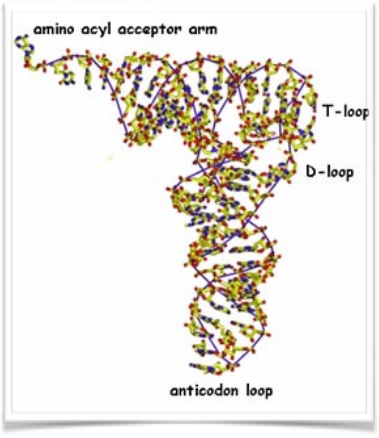
move away from the σ factor. The DNA bound σ factor could bind another polymerase (arrow 6) or the σ factor could release from the DNA and then diffuse around and rebind to other sites in the DNA or to the anti-σ factor if that protein is present (arrow 7).

As a reminder, RNA synthesis is a thermodynamically unfavorable reaction, so for it to occur it must be coupled to a thermodynamically favorable reaction, in particular nucleotide triphosphate hydrolysis (see previous chapter). The RNA polymerase moves along the DNA (or the DNA moves through the RNA polymerase, your choice), to generate an RNA molecule (the transcript). Other signals lead to the termination of transcription and the release of the RNA polymerase (arrow 5). Once released, the RNA polymerase returns to its inactive state. Another gene can be transcribed if the RNA polymerase interacts with a σ factor bound to its promoter (arrow 6). Since multiple types σ factor proteins are present within the cell and RNA polymerase can interact with all of them, which genes are expressed within a cell will depend upon the relative concentrations of σ factors and anti-σ factor proteins present and active, and the binding affinities of particular σ factors for specific DNA sequences (compared to their general low-affinity binding to DNA in general).

Protein synthesis: translation (RNA to polypeptide)

Translation involves a complex cellular organelle, the ribosome, which together with a number of accessory factors reads the code in a mRNA molecule and produces the appropriate polypeptide.[199] The ribosome is the site of polypeptide synthesis. It holds the various components (the mRNA, tRNAs, and accessory factors) in appropriate juxtaposition to one another to catalyze polypeptide synthesis. But perhaps we are getting ahead of ourselves. For one, what exactly is a tRNA?

While we have focussed on mRNA up to now, the process of transcription is also used to generate other types of RNAs; these play structural, catalytic, and regulatory roles within the cell. Of these non-mRNAs, two are particularly important in the context of polypeptide synthesis. The first are molecules known as transfer RNAs (tRNAs). These small single stranded RNA molecules fold back on themselves to generate a compact L-shaped structure (→). In the bacterium *E. coli*, there are 87 tRNA encoding genes (there are over 400 such tRNA encoding genes in human). For each amino acid and each codon there are one or more tRNAs. The only exception being the stop codons. A tRNA specific for the amino acid phenylalanine would be written $tRNA^{Phe}$. Two parts of the tRNA molecule are particularly important and functionally linked: the part that recognizes the codon on the mRNA and the amino acid acceptor stem, which is where an amino acid is attached to the tRNA. Each specific type of tRNA can recognize a particular codon in an mRNA through base pairing interactions with what is known as the anti-codon. The rest of the tRNA molecule mediates interactions with protein catalysts (enzymes) known as amino acyl tRNA synthetases. There is a distinct amino acyl tRNA synthetase for each amino acid, so that there is a phenylalanine-tRNA synthetase and a proline-tRNA synthetase, etc. An amino acyl tRNA synthetase binds the appropriate tRNA and amino acid and, through a reaction coupled to a thermodynamically favorable nucleotide triphosphate hydrolysis

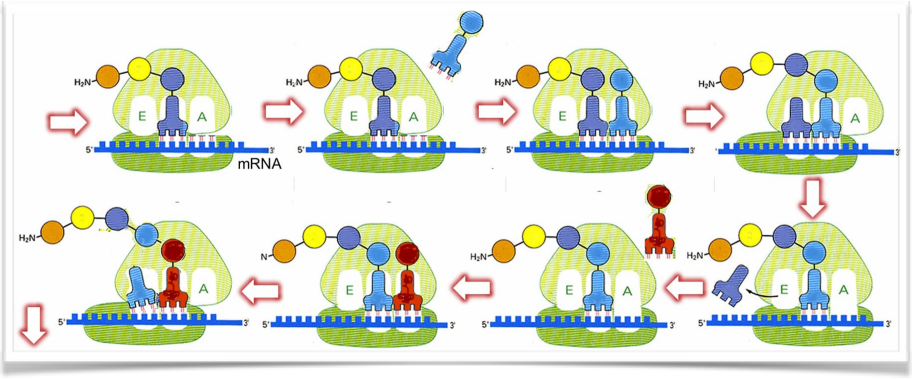


[199] Can't stop yourself? go here for a more detailed description of translation.http://www.nature.com/nsmb/journal/v19/n6/full/nsmb.2313.html?WT.ec_id=NSMB-201206

reaction, catalyzes the formation of a covalent bond between the amino acid acceptor stem of the tRNA and the amino acid, to form what is known as a charged or amino acyl-tRNA. The loop containing the anti-codon is located at the other end of the tRNA molecule. As we will see, in the course of polypeptide synthesis, the amino acid group attached to the tRNA's acceptor stem will be transferred from the tRNA to the growing polypeptide.

Ribosomes: Ribosomes are composed of roughly equal amounts (by mass) of ribosomal (rRNAs) and ribosomal polypeptides. An active ribosome is composed of a small and a large ribosomal subunit. In the bacterium *E. coli*, the small subunit is composed of 21 different polypeptides and a 1542 nucleotide long rRNA molecule, while the large subunit is composed of 33 different polypeptides and two rRNAs, one 121 nucleotides long and the other 2904 nucleotides long.[200] It goes without saying (so why are we saying it?) that each ribosomal polypeptide and RNA is itself a gene product. The complete ribosome has a molecular weight of ~3 x 10^6 daltons. One of the rRNAs is an evolutionarily conserved catalyst, known as a ribozyme (in contrast to protein based catalysts, which are known as enzymes). This catalytic rRNA lies at the heart of the ribosome - it catalyzes the transfer of an amino acid bound to a tRNA to the carboxylic acid end of the growing polypeptide chain.

The growing polypeptide chain is bound to a tRNA, known as the peptidyl tRNA. When a new aa-tRNA enters the ribosome's active site (site A), the growing polypeptide is added to it, so that it becomes the peptidyl tRNA (with a newly added amino acid, the amino acid originally associated with incoming aa-tRNA). This attached polypeptide group is now one amino acid longer. Again, the use of an RNA based catalysts is a conserved feature of polypeptide synthesis in all known organisms, and appears to represent an evolutionarily homologous trait.



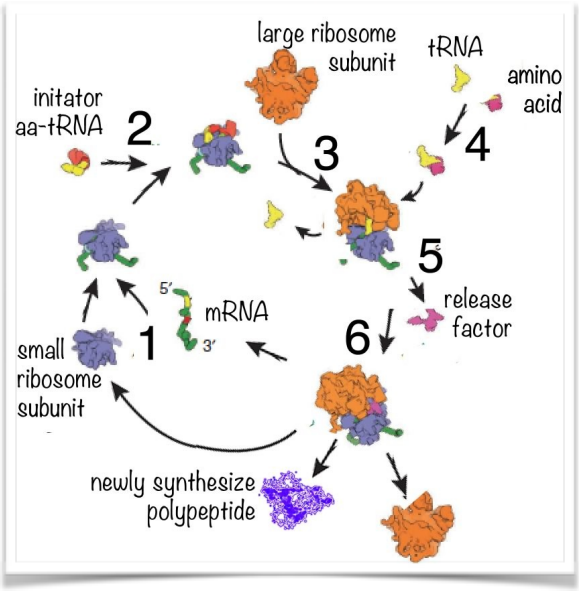
The cytoplasm of cells is packed with ribosomes. In a rapidly growing bacterial cell, approximately 25% of the total cell mass is ribosomes. Although structurally similar, there are characteristic differences between the ribosomes of bacteria, archaea, and eukaryotes. This is important from a practical perspective. For example, a number of antibiotics selectively inhibit polypeptide synthesis by bacterial, but not eukaryotic ribosomes. Both chloroplasts and mitochondria have ribosomes of the bacterial type. This is yet another piece of evidence that chloroplasts and mitochondria are descended from bacterial endosymbionts and a reason that translational blocking anti-bacterial antibiotics are mostly benign, since most of the ribosomes inside a eukaryotic cell are not effected by them.

[200] In the human, the small ribosomal subunit is composed of 33 polypeptides and a 1870 nucleotide rRNA, while the large ribosomal subunit contains 47 polypeptides, and three rRNAs of 121, 156, and 5034 nucleotides in length.

The translation (polypeptide synthesis) cycle

In bacteria, there is no barrier between the cell's DNA and the cytoplasm, which contains the ribosomal subunits and all of the other components involved in polypeptide synthesis. Newly synthesized RNAs are released directly into the cytoplasm, where they can begin to interact with ribosomes. In fact, because the DNA is located in the cytoplasm in bacteria, the process of protein synthesis (translation) can begin before mRNA synthesis (transcription) is complete.

We will walk through the process of protein synthesis, but at each step we will leave out the various accessory factors involved in regulating the process and coupling it to the thermodynamically favorable reactions that make it possible. These can be important if you want to re-engineer or manipulate the translation system, but are unnecessary conceptual obstacles that obscure a basic understanding. Here we will remind you of two recurring themes. The first is to recognize that all of the components needed to synthesize a new polypeptide (except the mRNA) are already present in the cell; another example of biological continuity. The second is that all of the interactions we will be describing are based on stochastic, thermally driven movements. For example, when considering the addition of an amino acid to a tRNA, random motions have to bring the correct amino acid and the correct tRNA to their binding sites on the appropriate amino acyl tRNA synthetase, and then bring the correct amino acid charged tRNA to the ribosome. Generally, many unproductive collisions will occur before a productive (correct) one, since there are more than 20 different amino acid/tRNA molecules bouncing around in the cytoplasm.

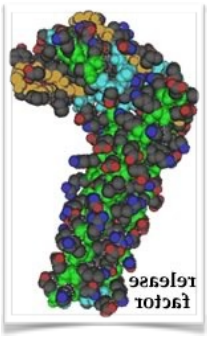


The first step in polypeptide synthesis is the synthesis of the specific mRNA that encodes the polypeptide. (1) The mRNA contains a sequence[201] that mediates its binding to the small ribosomal subunit. This sequence is located near the 5' end of the mRNA. (2) the mRNA-small ribosome subunit complex now interacts with and binds to a complex containing an initiator (start) amino acid:tRNA. In both bacteria and eukaryotes the start codon is generally an AUG codon and inserts the amino acid methionine (although other, non-AUG start codons are possible).[202] This interaction defines the beginning of the polypeptide and the reading frame within the mRNA. (3) The met-tRNA:mRNA:small ribosome subunit complex can now form a functional complex with a large ribosomal subunit to form the functional mRNA:ribosome complex. (4) Catalyzed by amino acid tRNA synthetases, charged amino acyl tRNAs will be present and can interact with the mRNA:ribosome complex to generate a polypeptide. Based on the mRNA sequence and the reading frame defined by the start codon, amino acids will be added

[201] Known as the Shine-Delgarno sequence for its discovers

[202] Hidden coding potential of eukaryotic genomes: nonAUG started ORFs: http://www.ncbi.nlm.nih.gov/pubmed/22804099

sequentially. With each new amino acid added, the ribosome moves along the mRNA. An important point, that we will return to when we consider the folding of polypeptides into their final structures, is that the newly synthesized polypeptide threads through a molecular tunnel in the ribosome. Only after the N-terminal end of the polypeptide begins to emerge from this tunnel can it begin to fold. (5) The process of polypeptide polymerization continues until the ribosome reaches a stop codon, that is a UGA, UAA or UAG.[203] Since there is no tRNA for this codon, the ribosome pauses, waiting for a charged tRNA which will never arrive. Instead, a polypeptide known as release factor, which has a shape something like a tRNA, binds to the polypeptide:mRNA:ribosome complex instead.
(6) This leads to the release of the polypeptide, the disassembly of the ribosome into small and large subunits, and the release of the mRNA.



When associated with the ribosome, the mRNA is protected against interaction with proteins (ribonucleases) that could degrade it, that is, break it down into nucleotides. Upon its release the mRNA may interact with a new small ribosome subunit, and begin the process of polypeptide synthesis again or it may interact with a ribonuclease and be degraded. Where it is important to limit the synthesis of particular polypeptides, the relative probabilities of these two events (new translation or RNA degradation) will be skewed in favor of degradation. Typically this is mediated by specific nucleotide sequences in the mRNA. The relationship between mRNA synthesis and degradation will determine the half-life of a population of mRNA molecules, the steady state concentration of the mRNA in the cell, and indirectly, the level of polypeptide present.

Bursting synthesis and alarm generation

At this point, let us consider a number of interesting behaviors associated with translation. First, the onset of translation begins with the small ribosomal subunit interacting with the 5' end of the mRNA. Multiple ribosomes can interact with a single mRNA, each moving down the mRNA molecule, synthesizing a polypeptide. Turns out, the initial interaction between an mRNA and the first ribosomal subunit makes it more likely that other ribosomal subunits can add, once the first ribosome begins moving away from the ribosomal binding site on the mRNA. This has the result that the synthesis of polypeptides from an RNA often involves a burst of multiple events. Since the number of mRNA molecules encoding a particularly polypeptide can be quite small (less than 10 per cell in some cases), this can lead to noisy protein synthesis. Bursts of new polypeptide synthesis can then be followed by periods when no new polypeptides are made.

The translation system is dynamic and a major consumer of energy within the cell.[204] When a cell, particularly a bacterial cell, is starving, it does not have the energy to generated amino acid charged tRNAs. The result is that uncharged tRNAs accumulate. Since uncharged tRNAs fit into the

[203] In addition to the common 19 amino and 1 imino (proline) acids, the code can be used to insert two other amino acids selenocysteine and pyrrolysine. In the case of selenocysteine, the amino acid is encoded by a stop codon, UGA, that is in a particular context within the mRNA. Pyrrolysine is also encoded by a stop codon. In this case, a gene that encodes a special tRNA that recognizes the normal stop codon UAG is expressed. see Selenocysteine: http://www.ncbi.nlm.nih.gov/pubmed/8811175

[204] Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources: http://www.ncbi.nlm.nih.gov/pubmed/24766808

amino-acyl-tRNA binding sites on the ribosome, their presence increases the probability of unproductive tRNA interactions with the mRNA-ribosome complex. When this occurs, the stalled ribosome generates a signal (see [205]) that can lead to adaptive changes in the cell that enable it to survive for long periods in a "dormant" state.[206]

Another response that can occur is a more social one. Some cells in the population can "sacrifice" themselves for their (generally closely related) neighbors (remember kin selection and inclusive fitness.) This mechanism is based on the fact that proteins, like nucleic acids, differ in the rates that they are degraded within the cell. Just as ribonucleases can degrade mRNAs, proteases degrade proteins and polypeptides. How stable a protein/polypeptide is depends upon its structure, which we will be turning to soon.

A common system within bacterial cells is known as an addiction module. It consists of two genes, encoding two distinct polypeptides. One forms a toxin molecule which when active can kill the cell. The second is an anti-toxin (a common regulatory scheme, think back to σ factors and anti-σ factors.) The key feature of the toxin-anti-toxin system is that the toxin molecule is stable, it is has a long half life. The half-life of a molecule is the time it takes for 50% of the molecules in a population to be degraded (or otherwise disappear from the system.) In contrast, the anti-toxin molecule's half-life is short. The result is that if protein synthesis slows or stops, the level of the toxin will remain high, while the level of the anti-toxin will drop rapidly, which leads to loss of inhibition of the toxin, and cell death. Death leads to the release of the cell's nutrients which can be used by its neighbors. A similar process can occur if a virus infects a cell, if the cell kills itself before the virus replicates, it destroys the virus and protects its neighbors (who are likely to be its relatives.

Questions to answer & to ponder:

- What are the "natural" limits to the structure of an R-group in a polypeptide?
- How would a condensation reaction be effected by the removal of water from a system?
- Why do we think that the use of a common set of amino acids is a homologous trait?
- What factors can you imagine influenced the set of amino acids used in organisms?
- Why so many tRNA genes?
- Why does the ribosome tunnel inhibit the folding of the newly synthesized polypeptide?
- What types of molecules does DNA directly encode? How about indirectly?
- How might a DNA molecule encode the structure of a lipid?
- How, in the most basic terms, do different tRNAs differ from one another?
- What is the minimal number of different tRNA-amino acid synthetases in a cell?
- What could happen if a ribosome started translating an mRNA at the "wrong" place?
- Why don't release factors cause the premature termination of translation at non-stop codons?
- What does it mean to say the genetic code is an algorithm?
- What is meant when people call the genetic code a "frozen accident"?
- What is (seriously) unrealistic about this tutorial [http://youtu.be/TfYf_rPWUdY]?
- Design a process (and explain the steps) by which you might reengineer an organism to use a new (non-biological) type of amino acid in its proteins.

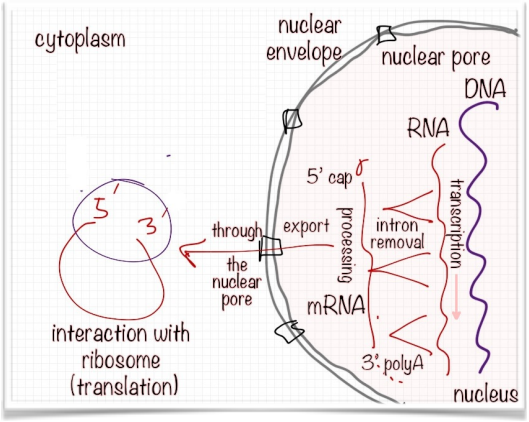
---

[205] http://virtuallaboratory.colorado.edu/BioFun-Support/labs/Adaptation/section_03.html

[206] Characterization of the Starvation-Survival Response of Staphylococcus aureus*:* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC107086/

Getting more complex: gene regulation in eukaryotes

At this point, we will not take vary much time to go into how gene expression in particular, and polypeptide synthesis in general differ between prokaryotes and eukaryotes except to point out a few of the major differences, some of which we will return to, but most will be relevant only in more specialized courses. The first and most obvious difference is the presence of a nucleus, a distinct domain within the eukaryotic cell that separates the cell's genetic material, its DNA, from the cytoplasm. The nucleus is a distinct compartment, with a distinct environment. This distinction is maintained by active processes that serve to both restrict the movement of molecules into and out of the nucleus (from the cytoplasm), and to re-establish the nuclear environment in situations in which it breaks down. As we will see later on this can occurs during cell division (mitosis). The barrier between nuclear interior and cytoplasm is known as the nuclear envelope (no such barrier exists in prokaryotic cells, the DNA is in direct contact with the cytoplasm.) The nuclear envelope consists of two lipid bilayer membranes that are punctuated by macromolecular complexes (protein machines) known as nuclear pores. While molecules of molecular weight less than ~40,000 atomic units (known as daltons) can generally pass through the nuclear pore, larger molecules must be transported actively, that is, in a process that is coupled to a thermodynamically favorable reaction, in this case the hydrolysis of guanosine triphosphate (GTP) instead of adenine triphosphate (ATP). The movement of larger molecules into and out of the nucleus through nuclear pores is regulated by what are known as nuclear localization and nuclear export sequences, present in polypeptides. These are recognized by proteins associated with the pore complex, and lead to movement of the polypeptide into or out of the nucleus.

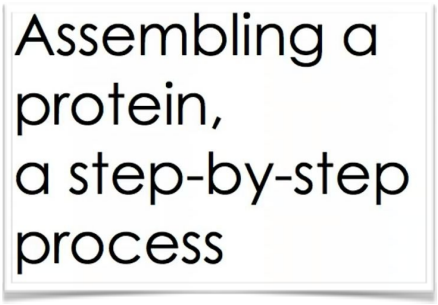


Aside from those within mitochondria and chloroplasts, the DNA molecules of eukaryotic cells are located within the nucleus. One difference between eukaryotic and bacterial genes is that the transcribed region of eukaryotic genes often contains what are known as intervening sequences or introns. After the RNA is synthesized, these non-coding introns are removed enzymatically, resulting in a shorter mRNA. As a point of interest, which sequences are removed can be regulated, this can result in mRNAs that encode somewhat (and often dramatically) different polypeptides. In addition to removing introns, the mRNA is further modified at both its 5' and 3' ends. Only after RNA processing as occurred is the mature mRNA exported out of the nucleus, through a nuclear pore into the cytoplasm, where it can interact with ribosomes. One further difference from bacteria is that the mRNA recognition of the small ribosomal subunit involves the formation of a complex in which the 5' and 3' ends of the mRNA are brought together into a circle. The important point here is that unlike the situation in bacteria, where mRNA is synthesized into the cytoplasm and so can immediately interact with ribosomes and begin translation (even before the synthesis of the RNA is finished), the coupling of transcription and translation does not occur in eukaryotes because of the nuclear envelope. Transcription occurs within the nucleus and the mRNA must be transported to the cytoplasm (where the ribosomes are located) before it can be translated. This makes processes like RNA splicing, and the generation of multiple,

functionally distinct RNAs from a single gene possible. This leads to significantly greater complexity from only a relatively small increase in the number of genes.

Turning polypeptides into proteins

Protein structure is commonly presented in a hierarchical manner. While this is an over-simplification, it is a good place to start. When we think about how a polypeptide folds, we have to think about the environment it will inhabit, how it interacts with itself, and were it is part of a *multi*-polypeptide protein, how its interactions with other subunits are established. As we think about polypeptide structure, it is typical to see it referred to in terms of primary, secondary, tertiary, and quaternary structure. The primary structure of a polypeptide is the sequence of amino acids in a polypeptide, chain, written from its N- or amino terminus to its C- or carboxyl terminus. As we will see below, the secondary structure of a polypeptide consists of local folding motifs: the α-helix, the β-sheet, and connecting domains. The tertiary structure of a polypeptide is the overall three dimensional shape a polypeptide takes in space (as well as how its R-chains are oriented). Quaternary structure refers to how the various polypeptides and co-factors that combine to make up a functional protein are arranged with respect to one another. In a protein that consists of a single polypeptide and no co-factors, its tertiary and quaternary structures are the same. As a final complexity, a particular polypeptide can be part of a number of different proteins. This is one reason that a gene can play a role in a number of different processes and be involved in a number of different phenotypes.



see http://youtu.be/Rq7DwrX0Uoc

Polypeptide synthesis (translation), like most all processes that occur within the cell, is a stochastic process, meaning that it is based on random collisions between molecules. In the specific case of translation, the association of the mRNA with ribosomal components occurs stochastically; similarly, the addition of a new amino acid depends on the collision of the appropriate amino acid-charged tRNA with the RNA-ribosome complex. Since there are many different amino-acid charged tRNAs in the cytoplasm, the ribosomal complex must be able to productively bind only the tRNA that the mRNA specifies, that is the tRNA with the right anticodon. This enables its attached amino acid to interact productively to add the amino acid to the growing polypeptide chain. In most illustrations of polypeptide synthesis, you rarely see this fact illustrated. From 12 to 21 amino acids are added per second in bacterial cells (and about half that rate in mammalian cells).[207]

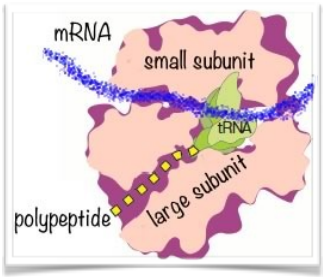
Now you might wonder if there are errors in polypeptide synthesis, as there are in nucleic acid synthesis. In fact there are. For example, if a base is skipped, the reading frame will be thrown off. Typically, this leads to a completely wrong sequence of amino acids added to the end of the polypeptide and generally quickly leads to a stop codon, which terminates translation, releasing a polypeptide that cannot fold correctly and is (generally) rapidly degraded.[208] Similarly, if the wrong amino acid is inserted at a particular position and it disrupts normal folding, the polypeptide could be

[207] see http://bionumbers.hms.harvard.edu/default.aspx

[208] Quality control by the ribosome following peptide bond formation: http://www.ncbi.nlm.nih.gov/pubmed/19092806

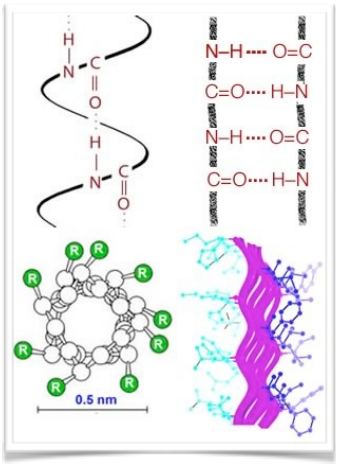
degraded. What limits the effects of mistakes during translation is that most proteins (unlike DNA molecules) have finite and relatively short half-lives; that is, the time an average polypeptide exists before it is degraded by various enzymes. Normally (but not always) this limits the damage that even an antimorphic polypeptide can do to the cell and organism.

Factors influencing polypeptide folding and structure: Polypeptides are synthesized, and they fold, in a vectorial, that is, directional manner. The polypeptide is synthesized in an N- to C- terminal direction and exits the ribosome through a tunnel approximately 10 nm long and 1.5 nm in diameter. This tunnel is narrow enough to block the folding of the newly synthesized polypeptide chain. As the polypeptide emerges from the tunnel, it encounters the crowded cytoplasmic environment; at the same time it begins to fold. As it folds, the polypeptide needs to avoid low affinity, non-specific, and non-physiologically significant interactions with other cellular components. These arise due to the fact that all molecules interact with each other via van der Waals interactions. If it is part of a multi-subunit protein, it must "find" its partner polypeptides, which again is a stochastic process. If the polypeptide does not fold correctly, it will not function correctly and may damage the cell. A number of degenerative neurological disorders are due, at least in part, to the accumulation of misfolded polypeptides (see below).



http://youtu.be/i8rGTYQ6oZ8

We can think of the folding process as a “drunken” walk across an energy landscape, with movements driven by thermal fluctuations and thermodynamic factors. The goal is to find the lowest point in the landscape, the energy minimum of the system. This is generally assumed to be the native or functional state of the polypeptide. That said, this state is not necessarily static, since the folded polypeptide (and the final protein) will be subject to thermal fluctuations; it is possible that it will move between various states with similar, but not identical stabilities. The problem of calculating the final folded state of a polypeptide is an extremely complex one. Generally two approaches are taken, in the first the structure of the protein is determined directly by X-ray crystallography or Nuclear Magnetic Resonance spectroscopy. In the second, if the structure of a homologous protein is known (and we will consider homologous proteins later on), it can be used as a framework to model the structure of a previously unsolved protein.

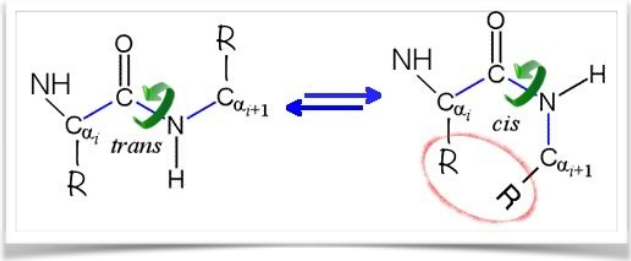


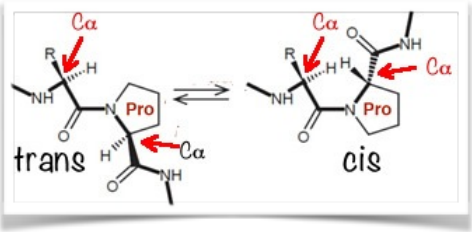
There are a number of constraints that influence the folding of a polypeptide. The first is the peptide bond itself. All polypeptides contain a string of peptide bonds. It is therefore not surprising that there are common patterns in polypeptide folding. The first of these common patterns to be recognized, the α-helix, was discovered by Linus Pauling and Robert Corey in 1951. This was followed shortly thereafter by their description of the β-sheet. The forces that drive the formation of the α-helix and the β-sheet will be familiar. They are the same forces that underlie water structure.

In an α-helix and a β-sheet, all of the possible H-bonds involving the peptide bond's donor and acceptor groups (–N–H : O=C– with “:” indicating a H-bond) are formed within the polypeptide. In the α-helix these H-bond interactions run parallel to the polypeptide chain. In the β-sheet they occur between

polypeptide strands. These strands can be within the same polypeptide and can run parallel or anti-parallel to one another. requiring one or more bends in the polypeptide. It is also possible to have β-sheet interactions between polypeptides located in different polypeptides. In an α-helix, the R-groups point outward from the helix axis. In β-sheets the R-groups point in an alternating manner either above or below the sheet. While all amino acids can take part in either α-helix or β-sheet structures, the imino acid proline cannot - the N-group coming off the α-carbon has no H, so its presence in a polypeptide chain leads to a break in the pattern of intrachain H-bonds.

Peptide bond rotation and proline: Although drawn as a single bond, the peptide bond behaves more like a double bond, or rather like a bond and a half. In the case of a single bond, there is free rotation around the bond axis in response to thermal motion. In contrast, rotation around a peptide bond requires more energy to move from the trans to the cis configuration and back again, that is, it is more difficult to rotate around the peptide bond. In addition, in the cis configuration the R groups of adjacent amino acids are on the same side of the polypeptide chain. If these R groups are large, they can bump into each other. If they get too close the repulsions between the outer electrons of each group make this arrangement less stable. This will usually lead to the polypeptide chain to prefer (at least locally) to be in the *trans* arrangement. In both α-helix and β-sheet configurations, the peptide bonds are in the *trans* configuration because the *cis* configuration disrupts their regular organization. However peptide bonds containing a proline residue have a different problem. The amino group is “locked” into a particular shape by the ring and therefore inherently destabilizes both α-helix and β-sheet structures (see above). Prolines are found in the cis configuration ~100 times as often as those between other amino acids. This *cis* configuration leads to a bend or kink in the polypeptide chain. The energy involved in the rotation around a proline bond is much higher than that of a standard peptide bond; so high, that there exist protein catalysts (peptidyl proline isomerases) that facilitate *cis-trans* rotations in such bonds. That said, the polypeptide chain folds as a unit, so increased stability elsewhere in the folded molecule can lead to an otherwise unfavorable local configuration elsewhere.





Hydrophobic R-groups: Many polypeptides and proteins exist primarily in an aqueous (water-based) environment. Yet, a number of their amino acid R-groups are hydrophobic. That means that their interactions with water will decrease the entropy of the system. Very much like the process that drives the assembly of lipids into micelles and bilayers, a typical polypeptide, with hydrophobic R groups along its length will, in aqueous solution, collapse onto itself so as to minimize the interactions of its hydrophobic residues with water. All else being equal minimizing their interaction with water will be thermodynamically favorable (since entropy will increase.) In practice this means that the first step in the folding of a polypeptide as it is synthesized is generally to move hydrophobic R-groups out of contact with water. This drives the collapse of the polypeptide into a compact and dynamic "molten

globule." In contrast where there are no (or few) hydrophobic R groups in the polypeptide, it will tend to adopt an elongated configuration. In contrast, if a protein comes to be embedded within a membrane (and we will briefly consider how this occurs later on), then the hydrophobic R-groups will be located on the surface of the folded polypeptide, so that they interact with the hydrophobic interior of the lipid bilayer. Hopefully this makes sense to you, thermodynamically.

The path to the native (that is, most stable) state is not necessarily a smooth or predetermined one. The folding polypeptide can get "stuck" in a local energy minimum; there may not be enough energy (derived from thermal collisions) for it to get out again. If a polypeptide gets stuck, there are active mechanisms to unfold it and let it try again to reach its native state. This process of partial unfolding is carried out by proteins known as chaperones. There are many types of protein chaperones; some interact with specific polypeptides as they are synthesized and attempt to keep them from getting into trouble, that is, folding in an unproductive way. Others can recognize inappropriately folded polypeptides and couple ATP hydrolysis with polypeptide unfolding, allowing the polypeptide a second (or third or …) chance to fold correctly. In the "simple" eukaryote, the yeast *Saccharomyces cerevisiae*, there are at least 63 distinct molecular chaperones [209]

chaperone video
http://youtu.be/b39698t750c

One class of chaperones are known as "heat shock proteins." The genes that encode these proteins are activated in response to increased temperature (as long as the increase is not so severe that it kills the cell immediately.) Given what you know about polypeptide/protein structure, you should be able to develop a plausible model by which to regulate the expression of heat shock genes. Heat shock proteins recognize unfolded polypeptides which are more likely to be present at higher temperatures. Heat-shock chaperones couple ATP hydrolysis reactions to unfold misfolded polypeptides. They then release the unfolded polypeptides giving them another chance to refold correctly. The chaperone does not directly control the behavior of the polypeptide. You might be asking now, how do chaperones recognize unfolded or abnormally folded proteins? Well unfolded proteins will tend to have hydrophobic regions exposed on the surface. the chaperones can recognize and interact with these regions and then help the polypeptide refold.

Heat shock proteins can be used to help an organism adapt. In classic experiments, when bacteria were grown at temperatures sufficient to turn on the expression of the genes that encode heat shock proteins, the bacteria had a higher survival rate when exposed to elevated temperatures compared to bacteria that had been grown continuously at lower temperature. Heat shock response-mediated survival at higher temperatures is an example of the ability of an organism to adapt to its environment - it is a physiological response. The presence of the heat shock system itself, however, is likely to be a selectable trait, encouraged by temperature variation in the environment. It is the result of evolutionary factors.

Acidic and basic R-groups: Some amino acid R-groups contain carboxylic acid or amino groups and so act as weak acids and bases. Depending on the pH of their environment these groups may be uncharged, positively charged, or negatively charged. Whether a group is charged or uncharged can have a dramatic effect on the structure, and therefore the activity, of a protein. By regulating pH, an

[209] An atlas of chaperone–protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2710862/

organism can modulate the activity of specific proteins. There are, in fact, compartments within eukaryotic cells that are maintained at low pH in part to regulate protein structure and activity. In particular, it is common for internal spaces of vesicles associated with endocytosis to become acidic (through the ATP-dependent pumping of H^+ across their membrane), which activates a number of enzymes involved in the hydrolysis of proteins and nucleic acids.

Subunits and prosthetic groups: Now you might find yourself asking yourself (a philosophically complex task), if most proteins are composed of multiple polypeptides, but polypeptides are synthesized individually, how are proteins assembled in a cytoplasm crowded with other proteins and molecules? This is a process that often involves specific chaperone proteins that bind to the newly synthesized polypeptide and either stabilize its folding, or hold it until it interacts with the other polypeptides it must interact with to form the final, functional protein. The absence of appropriate chaperones can make it difficult to assemble multisubunit proteins into functional proteins *in vitro*.

Many functional proteins also contain non-amino acid-based components, known generically as co-factors. A protein minus its cofactors is known as an apoprotein. Together with its cofactors, it is known as a holoprotein. Generally, without its cofactors, a protein is inactive and often unstable. Cofactors can range in complexity from a single metal ion to quite complex molecules, such as vitamin B12. The retinal group of bacteriorhodopsin and the heme group (with its central iron ion) are co-factors. In general, co-factors are synthesized by various anabolic pathways, and so they represent the activities of a number of genes. So a functional protein can be the direct product of a single gene, many genes, or (indirectly) entire metabolic pathways.

Questions to answer & to ponder

- How does entropy drive protein folding and assembly?
- Why does it matter that rotation around a peptide bond is constrained?
- How might changing the pH of a solution alter a protein's structure and activity?
- What happens to a typical protein if you place it in a hydrophobic solvent?
- What would be your prediction for the structure of a polypeptide if all of its R-groups were hydrophilic?
- How might a chaperone recognize a misfolded polypeptide?
- How would a chaperone facilitate the assembly of a protein composed of multiple polypeptides?
- Summarize the differences in structure between a protein that is soluble in the cytoplasm and one that is buried in the membrane.
- Why might proteins that require co-factors misfold in the absence of the co-factor?
- How might surface hydrophobic R-groups facilitate protein-protein interactions.
- Suggest a reason why cofactors would be necessary in biological systems (proteins)?
- Map the ways that a mutation in a gene encoding a chaperone influence a cell or organism?

Regulating protein localization

Translation of proteins occurs in the cytoplasm, where mature ribosomes are located. Generally, if no information is added, a newly synthesized polypeptide will remain in the cytoplasm. Yet even in the structurally simplest of cells, those of the bacteria and archaea, there is more than one place that a protein may need to be to function correctly: it can remain in the cytoplasm, it can be inserted into the plasma membrane or it may be secreted from the cell. Both membrane and secreted polypeptides must

be inserted into, or pass through, the plasma membrane.

Polypeptides destined for the membrane or for secretion are generally marked by a specific tag, known as a signal sequence. The signal sequence consists of a stretch of hydrophobic amino acids, often at the N-terminus of the polypeptide. As the signal sequence emerges from the ribosomal tunnel it interacts with a signal recognition particle (SRP) - a complex of polypeptides and a structural RNA. The binding of SRP to the signal sequence causes translation to pause. SRP acts as a chaperone for a subset of membrane proteins. The nascent mRNA/ribosome/nascent polypeptide/SRP will find (by diffusion), and attach to, a ribosome/SRP receptor complex on the cytoplasmic surface of the plasma membrane (in bacteria and archaea.) This ribosome/SRP receptor is associated with a polypeptide pore. When the ribosome/SRP complex docks with the receptor, translation resumes and the nascent polypeptide passes through a protein pore and so through the membrane. As the polypeptide emerges on the external, non-cytoplasmic face of the membrane, the signal sequence is generally removed by an enzyme, signal sequence peptidase. If the polypeptide is a membrane protein, it will remain within the membrane. If it is a secreted polypeptide, it will be released into the periplasmic space, that is the region outside of the cell's plasma membrane and inside its cell wall. Other mechanisms can lead to the release of the protein from the cell.

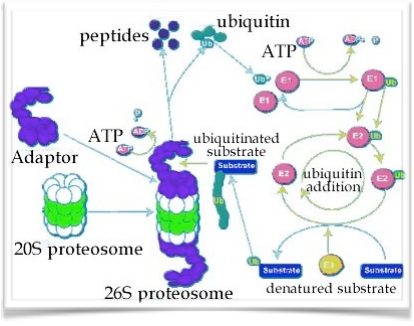
Eukaryotic cells are structurally and topologically more complex than bacterial and archaeal cells; there are more places for a newly synthesized protein to end up. While we will not discuss the details of those processes, one rule of thumb is worth keeping in mind. Generally, in the absence of added information, a newly synthesized polypeptide will end up in the cytoplasm. As in bacteria and archaea, a eukaryotic polypeptides destined for secretion or insertion into the cell's plasma membrane or internal membrane systems (that is the endoplasmic reticulum) are directed to their final location by a signal sequence/SRP system. Proteins that must function in the nucleus generally get there because they have a nuclear localization sequence, other proteins are actively excluded from the nucleus using a nuclear exclusion sequence (see above). Likewise, other localization signals and sequences are used to direct proteins to other intracellular compartments, including mitochondria and chloroplasts. While details of these targeting systems are beyond the scope of this course, you can assume that each specific targeting event requires signals, receptors, and various mechanisms that drive what are often thermodynamically unfavorable reactions.

Regulating protein activity

Proteins act through their interactions with other molecules. Catalytic proteins (enzymes) interact with substrate molecules; these interactions lower the activation energy of the reaction's rate limiting step, leading to an increase in the overall reaction rate. At the same time, cells and organisms are not static. They must regulate which proteins they produce, the final concentrations of those proteins within the cell (or organism), how active those proteins are, and where those proteins are located. It is primarily by altering proteins (and so indirectly gene expression) that cells (and organisms) adapt to changes in their environment.

A protein's activity can be regulated in a number of ways. The first and most obvious is to control the total number of protein molecules present within the system. Let us assume that once synthesized, a protein is fully active. With this simplifying assumption, the total concentration of a protein in a system [P_{sys}] is proportional to the rate of that protein's synthesis (dSynthesis/dt) minus the rate of that protein's degradation (dDegradation/dt). dt indicates per unit time. The combination of these two processes, synthesis and degradation, determines the protein's half-life. The degradation of proteins is mediated by a special class of enzymes (proteins) known as proteases. Proteases cleave peptide bonds via hydrolysis reactions. Proteases that cleave a polypeptide chain internally are known as endoproteases - they generate two polypeptides. Those that hydrolyze polypeptides from one end or the other, to release one or two amino acids at a time, are known as exoproteases. Proteases also can act more specifically, recognizing and removing a specific part of a protein in order to activate it or to inactivate it, or to control where it is found in a cell. For example, nuclear proteins become localized to the nucleus (typically) because they contain a nuclear localization sequence or they can be excluded because they contain a nuclear exclusion sequence. For these sequences to work they have to be able to interact with the transport machinery associated with the nuclear pores; but the protein may be folded so that they are hidden. Changes in a protein's structure can reveal or hide such targeting sequences, thereby altering the protein's distribution within the cell and its activity. For example, many proteins are originally synthesized in a longer, and inactive "pro-form". To become active the pro-peptide must be removed - it is cut by an endoprotease. This proteolytic processing activates the protein. Proteolytic processing is itself often regulated (see below).

Controlling protein levels: Clearly the amount of a protein within a cell (or organism) is a function of the number of mRNAs encoding the protein, the rate that these mRNAs are recognized and translated, and the rate at which functional protein is formed, which in turn depends upon folding rates and their efficiency. It is generally the case that once translation begins, it continues at a more or less constant rate. In the bacterium *E. coli*, the rate of translation at 37ºC is about 15 amino acids per second. The translation of a polypeptide of 1500 amino acids therefore takes about 100 seconds. After translation, folding and, in multisubunit proteins, assembly, the protein will function (assuming that it is active) until it is degraded.



Many proteins within the cell are necessary all of the time. Such proteins are considered "constitutive." Protein degradation is particularly important for controlling the levels of "regulated" proteins, whose presence or concentration within the cell may lead to unwanted effects in certain situations. The regulated degradation of a protein typically begins when the protein is specifically marked for degradation. This is an active and highly regulated process, involving ATP hydrolysis and a multi-subunit complex known as the proteosome. The proteosome degrades the polypeptide into small peptides and amino acids that can be recycled. As a mechanism for regulating protein activity, however, degradation has a serious drawback, it is irreversible. Since both a protein's synthesis and degradation can be regulated, its half-life can be regulated.

Allosteric regulation

A reversible form of regulation is known as allosteric regulation, where regulatory molecules bind reversibly to the protein altering its conformation, which in turn alters the protein's activity and can alter its location within the cell and its half-life. Such allosteric effectors are not covalently attached to the protein and can act either positively or negatively. The nature of such factors is broad, they can be a small molecule or another protein. What is important is that the allosteric binding site is distinct from the enzyme's catalytic site. In fact allosteric means other site. Because allosteric regulators do not bind to the same site on the protein as the substrate, changing substrate concentration generally does not alter their effects.

Of course there are other types of regulation as well. A molecule may bind to and block the active site of an enzyme. If this binding is reversible, then increasing the amount of substrate can overcome the inhibition. An inhibitor of this type is known as a competitive inhibitor. In some cases, the inhibitor chemically reacts with the enzyme, forming a covalent bond. This type of inhibitor is essentially irreversible, so that increasing substrate concentration does not overcome inhibition. These are therefore known as non-competitive inhibitors. Allosteric effectors are also non-competitive, since they do not compete with substrate for binding to the active site. That said, binding of substrate could, in theory, change the affinity of the protein for its allosteric effectors, just as binding of the allosteric effector changes the binding affinity of the protein for the substrate.

Post-translational regulation

Proteins may be modified after synthesis - this process is known as post-translational modification. A number of post-translational modifications have been found to occur within cells. In general where a protein can be modified it can also be unmodified. The exception, of course, is when the modification involves protein degradation. The first, and most common type of modification we will consider involves the covalent addition of specific groups to specific amino acid side chains on the protein - these groups can range from phosphate groups (phosphorylation), an acetate group (acetylation), the attachment of lipid/hydrophobic groups (lipid modification), or carbohydrates (glycosylation). Such post-translational modifications are generally reversible, one enzyme adds the modifying group and another can remove it. For example, proteins are phosphorylated by enzymes known as protein kinases, while protein phosphotases remove these phosphate groups. Post-translational modifications act in much the same way as do allosteric effectors, they modify the structure and, in turn, the activity of the polypeptide to which they are attached. They can also modify a protein's interactions with other proteins, the protein's localization within the cell, or its stability.

Questions to answer & to ponder

- A protein binds an allosteric regulator - what happens to the protein?
- How is the post-translational modification of a protein like allosteric regulation? how is it different?
- Why are enzymes required for post-translational modification?
- Why is a negative allosteric regulator not considered a "competitive" inhibitor?
- Why do post-translational modifications (and their reversals) require energy?
- How does a signal sequence influence translation?
- How would a cell recover from the effects of an irreversible, non-competitive inhibitor?
- Why might a cell want a specific protein to have a short half-life?

- What would happen if you somehow put a signaling sequence at the beginning of a normally cytoplasmic polypeptide?
- Draw out the factors and their interactions that control the half-life, activity, and location of a particular protein within a biological system.

Diseases of folding and misfolding

If a functional protein is in its native (or natural) state, a dysfunctional misfolded protein is said to be denatured. It does not take much of a perturbation to unfold or denature most proteins. In fact, under normal conditions, proteins often become partially denatured spontaneously, normally these are either refolded (often with the help of chaperone proteins) or degraded (through the action of proteosomes and proteases). A number of diseases, however, arise from protein misfolding.

Kuru was among the first of these protein misfolding diseases to be identified. Beginning in the 1950s, D. Carleton Gadjusek (1923 – 2008)[210] studied a neurological disorder common among the Fore people of New Guinea. The symptoms of kuru, which means "trembling with fear", are similar to those of scrapie, a disease of sheep, and variant Creutzfeld-Jakob disease (vCJD) in humans. Among the Fore people, kuru was linked to the ritual eating of the dead. Since this practice has ended, the disease has disappeared. The cause of kuru, scrapie and vCJD appears to be the presence of an abnormal form of a normal protein, known as a prion. We can think of prions as a type of anti-chaperone. The idea of proteins as infectious agents was championed by Stan Prusiner, who was awarded the Nobel Prize in Medicine in 1997.[211]

The protein responsible for kuru and scrapie is known as PrPc. It normally exists in a largely α-helical form. There is a second, abnormal form of the protein, PrPsc for scrapie; whose structure is primarily of β-sheet. The two polypeptides have the same primary sequence. PrPsc acts as an anti-chaperone, catalyzing the transformation of PrPc into PrPsc. Once initiated, this leads to a chain reaction and the accumulation of PrPsc. As it accumulates, PrPsc assembles into rod-shaped aggregates that appear to damage cells. When this process occurs within the cells of the central nervous system it leads to severe neurological defects. There is no natural defense, since the protein responsible is a normal protein.

Disease transmission: When the Fore ate the brains of their beloved ancestors, they inadvertently introduced the PrPsc protein into their bodies. Genetic studies indicate that early humans evolved resistance to prion diseases, suggesting that cannibalism might have been an important selective factor during human evolution. Since cannibalism is not nearly as common today, how does anyone get such diseases in the modern world? There are rare cases of iatrogenic transmission, that is, where the disease is caused by faulty medical practice, for example through the use of contaminated surgical instruments or when diseased tissue is used for transplantation.

But where did people get the disease originally? Since the disease is caused by the formation of PrPsc, any event that leads to PrPsc formation could cause the disease. Normally, the formation of

---

[210] Carleton Gajdusek: http://www.theguardian.com/science/2009/feb/25/carleton-gajdusek-obituary

[211] Stanley Prusiner: 'A Nobel prize doesn't wipe the skepticism away': http://www.theguardian.com/science/2014/may/25/stanley-prusiner-neurologist-nobel-doesnt-wipe-scepticism-away and http://youtu.be/yzDQ8WgFB_U

PrPsc from PrPc is very rare. We all have PrPc but very few of us spontaneously develop kuru-like symptoms. There are, however, mutations in the gene that encodes PrPc that greatly enhance the frequency of the PrPc → PrPsc conversion. Such mutations may be inherited (genetic) or may occur during the life of an organism (sporadic). Fatal familial insomnia (FFI) is due to the inheritance of a mutation in the *PRNP* gene, which encodes PrPc. This mutation changes the normal aspartic acid at position 178 of the PrPc protein to an asparagine. When combined with a second mutation in the *PRNP* gene at position 129, the FFI mutation leads to Creutzfeld-Jacob disease (CJD). If one were to eat the brain of a person with FFI or CJD one might well develop a prion disease.

So why do PrPsc aggregates accumulate? To cut a peptide bond, a protease must position the target peptide bond within its catalytic active site. If the target protein's peptide bonds do not fit into the active site, they cannot be cut. Because of their structure, PrPsc aggregates are highly resistant to proteolysis. They gradually accumulate over many years, a fact that may explain the late onset of PrP-based diseases.

Why do harmful alleles persist?

At this point, you might well ask yourself, given the effectiveness of natural selection, why do alleles that produce severe diseases exist at all? There are a number of possible scenarios. One is that a new mutation arose spontaneously, either in the germ line of the organism's parents or early in the development of the organism itself, and that it will disappear from the population with the death of the organism. The prevalence of the disease will then reflect the rate at which such pathogenic mutations occur. The second, more complex reason involves the fact that many organisms carry two copies of each gene (they are diploid), and that carrying a single copy of the allele might either have no discernible effect on the organism's reproductive success or, in some cases, might even lead to an increase in reproductive success. In this case, the allele will be subject to positive selection, that is, it will increase in frequency. This increase will continue until the number of individuals carrying the allele reaches a point where the number of offspring with two copies of the mutant (pathogenic) allele becomes significant. These individuals (and the alleles they carry) are subject to strong negative selection. We will therefore arrive at a steady state population where the effects of positive selection (on individuals carrying one copy of the allele) will be balanced by effects of negative selection on individuals that carry two copies of the allele. You could model this behavior in an attempt to predict the steady state allele frequency by considering the sizes of the positive and negative effects and the probability that a mating will produce an organism with one (a heterozygote) or two (a homozygote) copies of the allele.

Generally the process of selection occurs gradually, over many (hundreds to thousands) of generations, but (of course) the rate depends on the strength of the positive and negative effects of a particular allele on reproductive success. As selection acts, and the population changes, the degree to which a particular trait influences reproductive success can also change. The effects of selection are themselves not static, but evolve. For example, a trait that is beneficial when rare may be less beneficial when common. New mutations that appear in the same or different genes can further influence the trait, and so how the population will change over time. For example, alleles that were

"neutral" or without effect in the presence of certain alleles at other genes (known as the genetic background) can have effects when moved into another genetic background. A (now) classic example of this effect was described by studies on the laboratory evolution of the bacterium *Escherichia coli*. A mutation with little apparent effect occurred in one lineage and its presence made possible the emergence of a new trait (the ability to use citrate for food) about 20,000 generations later.[212] We will return to how this works exactly toward the end of the course, but what is important here is that it is the organism (and its traits and all its alleles) that is "selected". Only in rare cases of extremely strong positive or negative selection, does it make sense to say that specific alleles are selected.

Questions to answer & to ponder

- How does the presence of PrPsc lead to the change in the structure of PrPc?
- Why is it, do you think, that FFI and CJD are late onset diseases?
- Which do you think would be more susceptible to proteolytic degradation, a compact or an extended polypeptide?

[212] Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli: http://www.pnas.org/content/105/23/7899.long