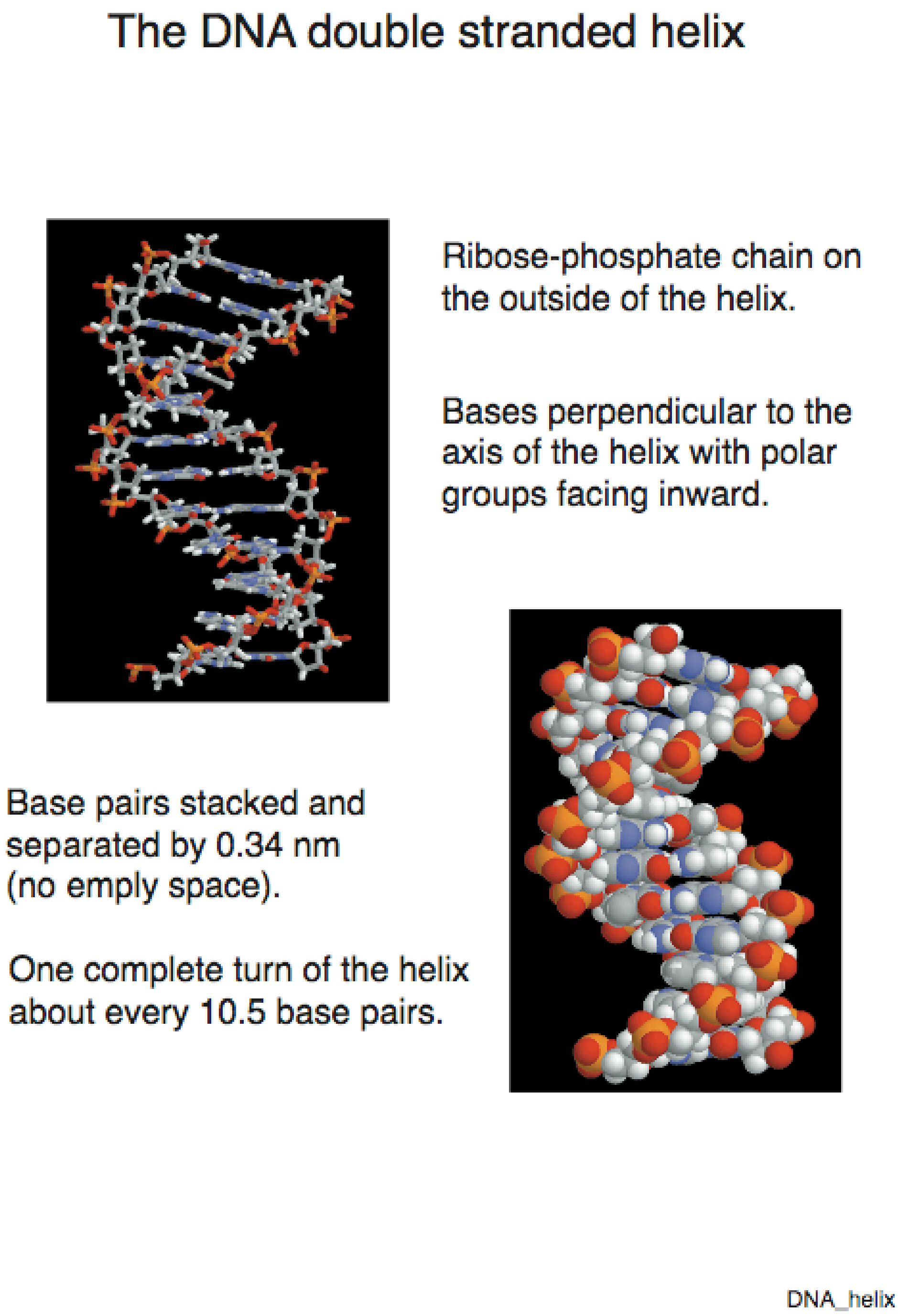
Chapter 3.2 Copying and reading DNA

The double stranded DNA helix

In the previous chapter we described the structure of a strand of DNA. However, DNA in the cell is not present as a single strand, but rather as two nucleotide strands running in opposite directions, twisting around each other to form a helix. In one of the most dramatic cases of form following function, the structure of this double stranded helix immediately suggests the mechanism for DNA replication.

A DNA helix is shown as a stick and a space filling model in the upper and lower panels of Figure dna_helix respectively. The ribose-phosphate backbones of the two strands (traced by the orange phosphates) form spirals on the outside of the helix while the flat bases, seen on edge, point in toward the middle. At each level of the helix the bases attached to the strands forms a pair. One member of the pair is a pyrimidine (a six atom ring) and the other is a purine (a six and a five atom ring fused together). The fact that there is always one large and one small base in the pair means that there is a constant distance between the ribose groups on the two chains, i.e. the helix has a constant diameter.

Figure DNA_helix



The DNA double stranded helix

Ribose-phosphate chain on the outside of the helix.

Bases perpendicular to the axis of the helix with polar groups facing inward.

Base pairs stacked and separated by 0.34 nm (no emply space).

One complete turn of the helix about every 10.5 base pairs.

DNA_helix

Figure DNA_helix. Two strands of DNA are twisted around each other to form a helix. The flat bases are stacked on each other and point toward the center, while the phosphate-sugar chains are on the outside. Each base on one strand is paired with a base on the opposite strand.

In the stick model it looks as if there might be space between the stacked base pairs, but the space filling model shows that not to be the case. The base pairs are in contact, and attractive forces between them, due to overlap of the electron orbital clouds and electrostatic attraction of partial charges on the rings, e.g. Figure guanine, are major forces giving the helix stability.

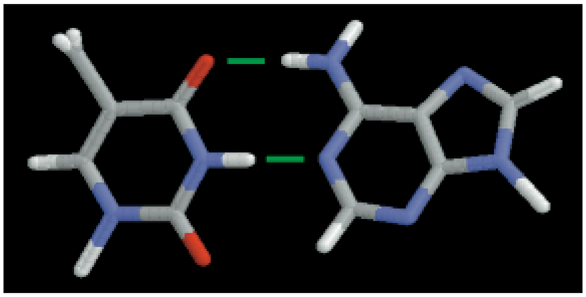
The two base pairs

Each base pair in DNA is not just any pyrimidine plus any purine, but rather is a specific pyrimidine-purine pair. As seen in Figure base_pair, thymine always pairs with adenine and cytosine always pairs with guanine.

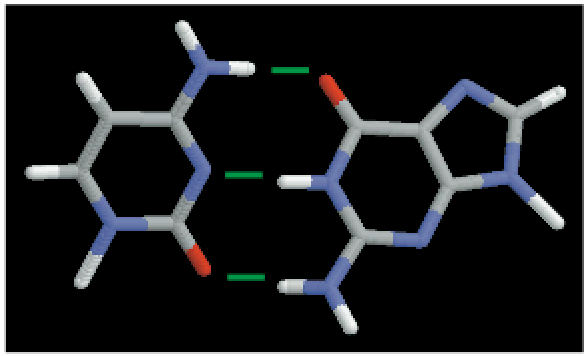
Figure base_pair

The two base pairs

(— are H bonds)



thymine : adenine



cytosine : guanine

basePair

Figure base_pair. In the DNA double helix an adenine base on one strand pairs with a thymine on the other while a cytosine pairs with a guanine.

Only these pairs result in alignment of partial positive and negative charged atoms to make hydrogen bonds, indicated by the green lines in the Figure. This pairing specificity, or complementarity, means that the base sequence along one DNA strand completely determines the base sequence along the other strand. Thus, each strand of a DNA helix contains the same information (but not the same sequence).

Base pairing enables replication

The structure of the DNA helix, deduced by Watson and Crick in 1953, immediately suggested to them a mechanism for DNA replication.

1. The two strands of the parental DNA helix unwind and separate.

2. Nucleotides diffuse from the cellular environment and form new hydrogen bonded complementary base pairs with the exposed parental DNA bases.

3. The new nucleotides are linked together to form new DNA strands and thus regenerate two double stranded DNA helixes.

The real chemistry of DNA replication

Base pairing is indeed the basis for DNA replication, but the actual process is at least one order of magnitude more complicated than the three-step "suggestion" above. Many of the complications in implementing the suggestion can serve as general paradigms for the other chemical processes that enable life. With this motivation and hope we plunge ahead into the details of DNA replication.

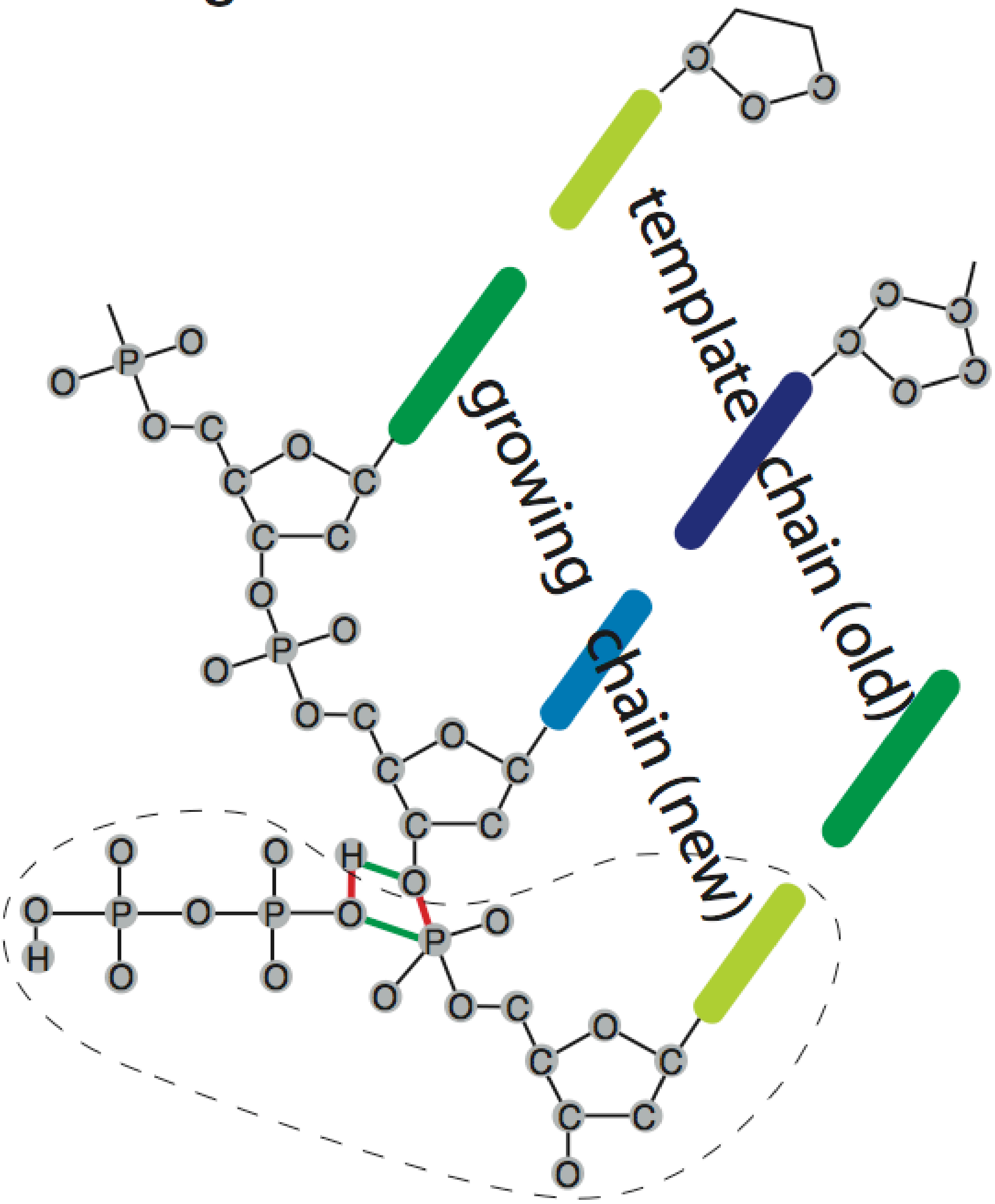
Replication must move down the energy hill

The problem we consider first is actually implementation of item 3 in the “suggestion”. If one takes a number of nucleoside monophosphates (base + sugar + phosphate), which are indeed the units (monomers) of the DNA chain, and arrange them next to each other just as they would be in a DNA helix, the nucleotides will not spontaneously link together to form a DNA strand. Even if you wait a thousand years, it won’t happen. It won’t happen because there is more energy in the chemical bonds of DNA chain than there is in the chemical bonds of the collection of nucleoside monophosphates. You are expecting a low energy situation to change to a high-energy configuration. It’s like placing a bowling bar at the bottom of a hill (a low energy situation) and waiting for it to roll up the hill to the top (a higher energy situation).

The solution to this problem is to change the chemical reaction so the starting molecules (the substrates) have more energy than the final molecule (the products). Then we will be moving down the energy hill as the reaction occurs. We are not at liberty to change the structure of the product since it is by definition the solution to the problem. However, we can change the substrates, as long as they can still serve to make the product. Living cells have evolved to use nucleoside triphosphates as the substrates for DNA replication. The structure of a nucleoside triphosphate is seen in Figure adding1nucleotide.

Figure adding1nucleotide

Adding one nucleotide



adding1nucleotide

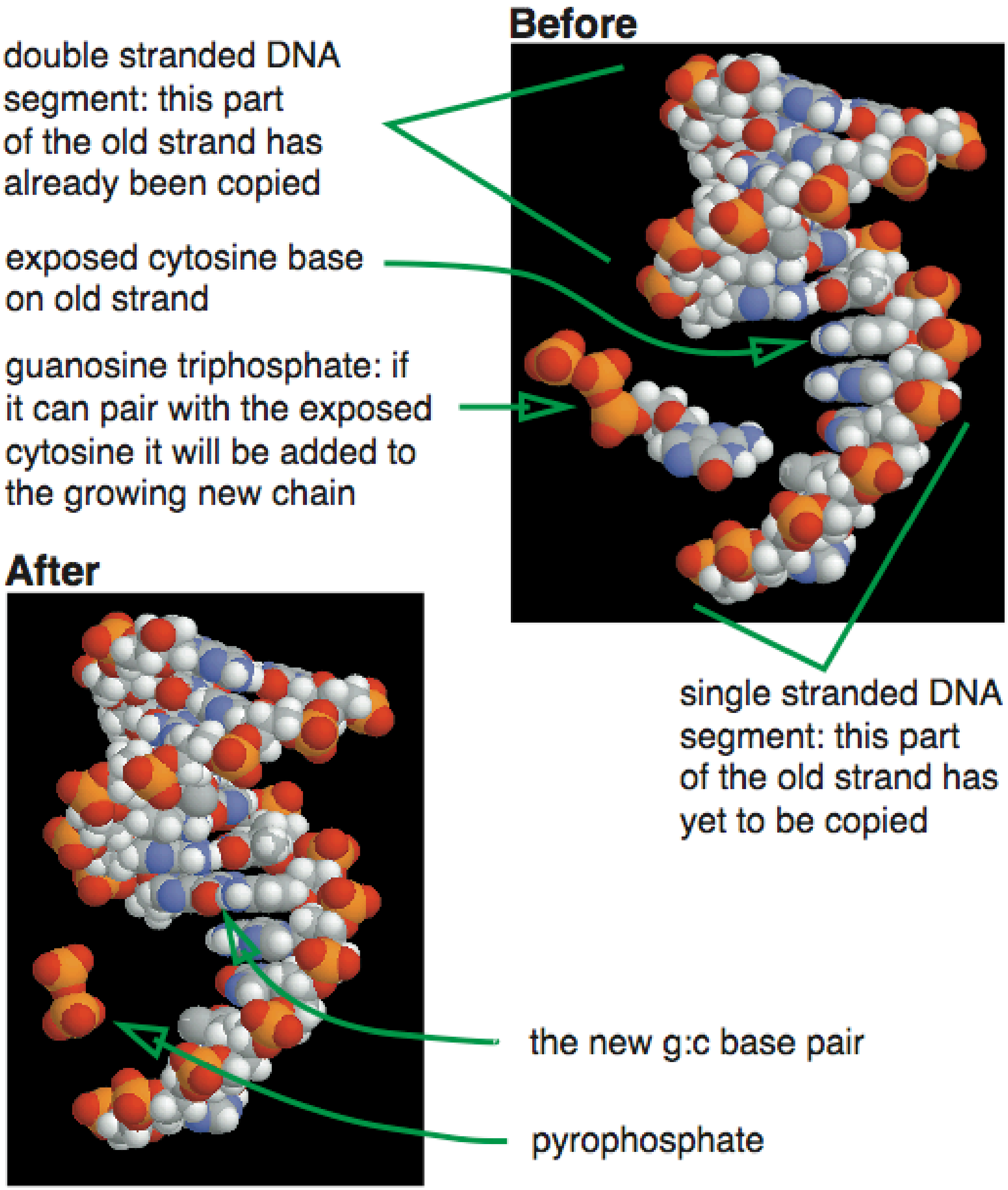
Figure adding1nucleotide. The template DNA chain is copied by the pairing of a complementary nucleotide triphosphate (enclosed by the dashed line) at the end of the new, growing chain (bases are seen edge on as solid bars). At the point of attachment of the new nucleotide, bond electrons move to convert old (green) bonds to new (red) bonds. The pyrophosphate group (two linked phosphates) diffuses away to leave a DNA chain one nucleotide longer.

As the name suggests, the nucleotide triphosphate has a chain of three phosphate groups bonded to the 5′ carbon of the deoxyribose. This produces a very different molecule from an energy point of view because the P-O-P links have a much higher energy than the P-O-C link. If the nucleoside triphosphate is added to the end of a growing DNA chain and the extra phosphates are released (the two green bonds replaced by the two red bonds in Figure adding1nucleotide) the total energy decreases. Now we are moving down the energy hill.

The addition of a new nucleotide is shown in the more realistic, but more complex space filling molecular model seen in Figure growing_dna. The DNA is a double helix in the top portion of the Figure, while the single strand template that is to be copied extends down on the right of the panel. The nucleoside triphosphate approaches from the left, and if it can pair with the template strand a new ribose-phosphate bond can be made. The pyrophosphate group will then be released to diffuse out from the DNA as shown in the lower panel.

Figure growing_dna

Adding one base to a new copy of an old DNA strand


Before

double stranded DNA segment: this part of the old strand has already been copied

exposed cytosine base on old strand

guanosine triphosphate: if it can pair with the exposed cytosine it will be added to the growing new chain

single stranded DNA segment: this part of the old strand has yet to be copied

After

the new g:c base pair

pyrophosphate

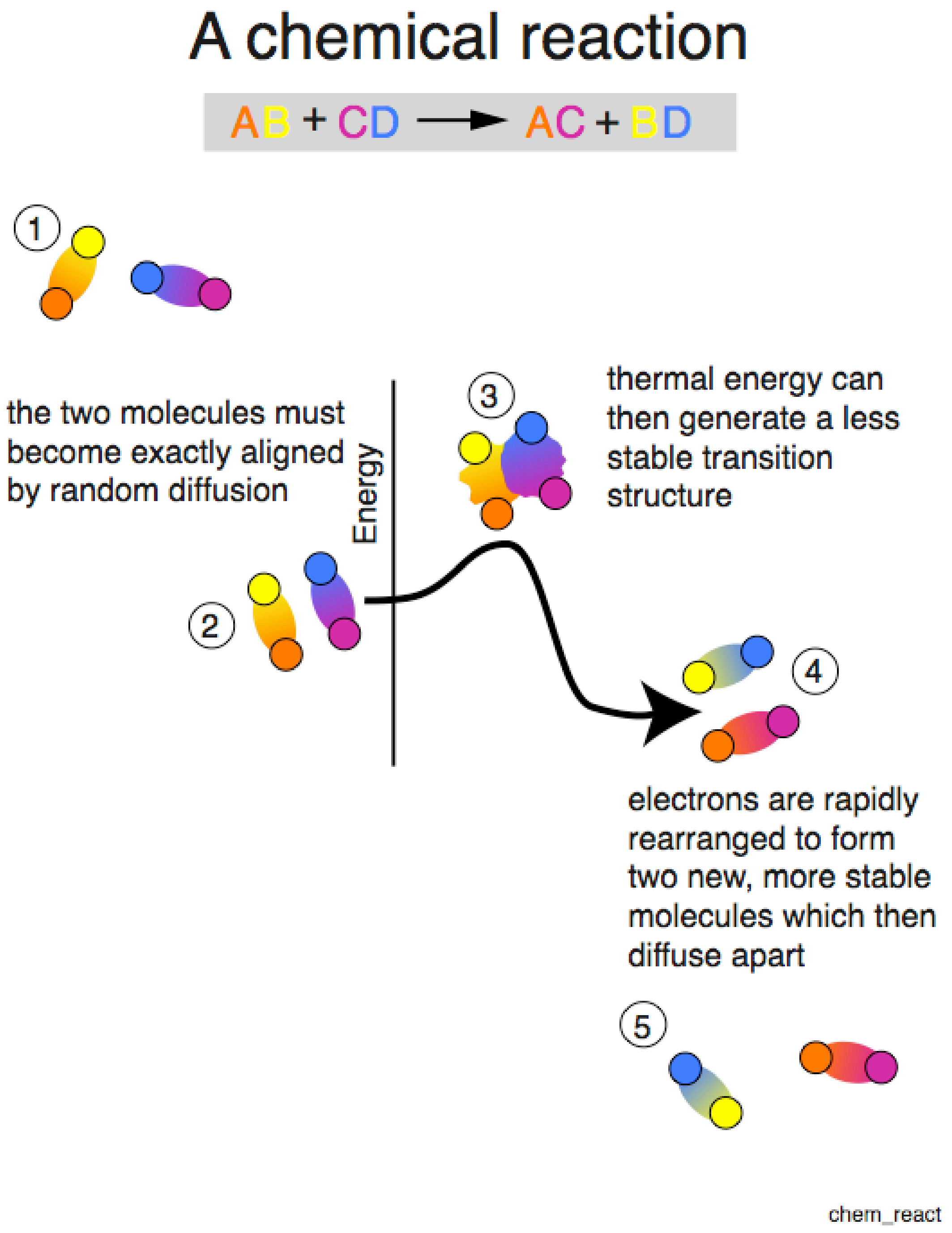
growing_dna

Figure growing_dna. A nucleotide triphosphate is added to a new, growing DNA chain to produce a complementary copy of the old template strand. In this example the new base is a guanine which pairs with the cytosine base in the old template strand. This is a more realistic but complex view of the process shown in Figure adding1nucleotide.

Replication must occur at a reasonable speed

If we take a number of nucleoside triphosphates and arrange them next to each other just as they would be in a DNA helix, the nucleotides will link together to form a DNA strand, but will take a very, very long time. We have to speed up the process a great deal, and the solution is actually just a variation of the previous solution. First, to see the problem from a more general, abstract viewpoint, follow the generic chemical reaction in Figure chem_react.

Figure chem_react



the two molecules must become exactly aligned by random diffusion

thermal energy can then generate a less stable transition structure

electrons are rapidly rearranged to form two new, more stable molecules which then diffuse apart

Figure chem._react. Chemical reactions require the rearrangement of bonds, which means shifts in bonding electrons. Even when the final product is more stable than the original molecules, there are intermediate states when the electron distributions have a higher energy. The overall chemical reaction is thus dependent on the small numbers of molecules that have the required energy to get over this transition barrier. The lower the transition barrier the greater the number of molecules that have the required energy, and the greater the rate of the chemical reaction.

In general a chemical reaction requires some chemical bonds to be broken and new chemical bonds to be made, i.e. chemical bonds are rearranged. When the bonds have covalent character, electron clouds around the atoms at the reaction site must be rearranged. In the Figure, two old bonds are broken and two new bonds are made, just as in the addition of a nucleotide to the growing DNA chain. However, in the middle of the transition the electrons are in neither stable configuration, and thus have a higher energy than the starting configuration. Using the hill analogy, the electrons must get up over a little bump before they can slide down the hill. Even if the initial and final bonds were completely ionic in character, the ion centers would have to move through positions having a higher energy, which would also be represented as an energy bump.

But, getting over little bumps in the energy landscape does not take a long time, and the smaller the bump the faster it is. Remember, all these molecules have thermal energy. They, and all the atoms that are part of the molecule, are moving, bending, twisting, rotating, and vibrating with an average kinetic energy proportional to the temperature, kT/2. Energy bumps that are low compared to this energy have no real ability to confine the behavior of the molecule. In addition, objects in a given environment have a wide distribution of energies, depending on what molecules they have bumped into recently. Thus, a molecule may not have sufficient energy to get over a medium sized bump at one instant, but wait a while and it will be able to. If the bump is large it will take a very long time, because the average time required to get over an energy bump increases exponentially with the height of the bump.

Don't heat

Since average energy increases with temperature, why not just increase the temperature to get over the energy bump quickly and complete the reaction? When chemists make new compounds in the laboratory, they often heat the reaction flask to do just that. If the energy released by the reaction is converted into heat, you may need to only heat a small portion of the reaction mix to start the reaction. Then the heat generated in that portion will enable more material to react and you are off and running with an autocatalytic reaction. This is what you are doing when you light a fire.

However, heating up DNA to speed the replication reaction is not a good idea. Many of the bonds that hold the DNA helix together are quite weak, and even a modest temperature increase would destroy its structure. In addition, there are other molecules near the DNA that are at least as sensitive to an elevated temperature.

Use an enzyme

If we can decrease the size of the energy bump the starting molecules must pass over as they rearrange to form products, more molecules will have the energy to get over the bump. Again this can be done by changing the reaction, but instead of changing the substrates we create a new intermediate product by having the nucleotide triphosphate and DNA combine with an enzyme. This will accomplish our goal because the energy required to form this enzyme complex is less than the original energy bump. The enzyme complex then decomposes to form the desired products and the original enzyme. Thus, the enzyme is only a transient player in the reaction, and can play its role over and over again.

Enzymes are very important because they are involved in the great majority of chemical reactions that occur in the cell. In fact, the main function of DNA is to provide the information needed to make the many different enzymes required by the cell. There are many different enzymes because most enzymes are very specific; they speed up only one chemical reaction.

You could focus on the fact that enzymes makes it possible for chemical reactions to occur at a high rate; they increase the output of a chemical factory. However, cells might have evolved with fewer enzymes. Enzymes could be less specific, so they could accelerate a variety of reactions. Cells could have evolved to generate substrates that react spontaneously without the need for enzymes.

An alternate view for the role of enzymes is associated with seeing the cell as an information network, with each chemical species being a link in the network. The enzymes are then the nodes which connect the links and control the flow of information by defining the rate of the corresponding chemical transformation. The need for a specific enzyme to catalyze each reaction is thus a feature, not a bug.

In fact, it is known that the relative amounts of specific enzymes are controlled by the cell to regulate the rates of the corresponding chemical reactions. In addition, the activity of many enzymes is known to be modulated by the concentration of chemical species that are not directly related to the reaction catalyzed. However, in most cases one can see an evolutionary advantage for the activity of the enzyme to be modulated by this chemical species. We can think of the enzyme as a transistor with the chemical reaction it controls being the emitter-collector current and the modulating chemical species being the base current.

What is an enzyme?

Most enzymes are proteins, which are polymers of amino acids. While DNA is a polymer containing four different kinds of nucleotides, proteins contain twenty different kinds of amino acids. While DNA is a long double helix with a structure almost independent of nucleotide sequence, the proteins are typically compact, globular molecules with a complex and varied structure that is highly dependent on the sequence of amino acids. The sequence of amino acids determines the three dimensional shape of the protein, but even now the rules for this specification are only partially understood by us humans.

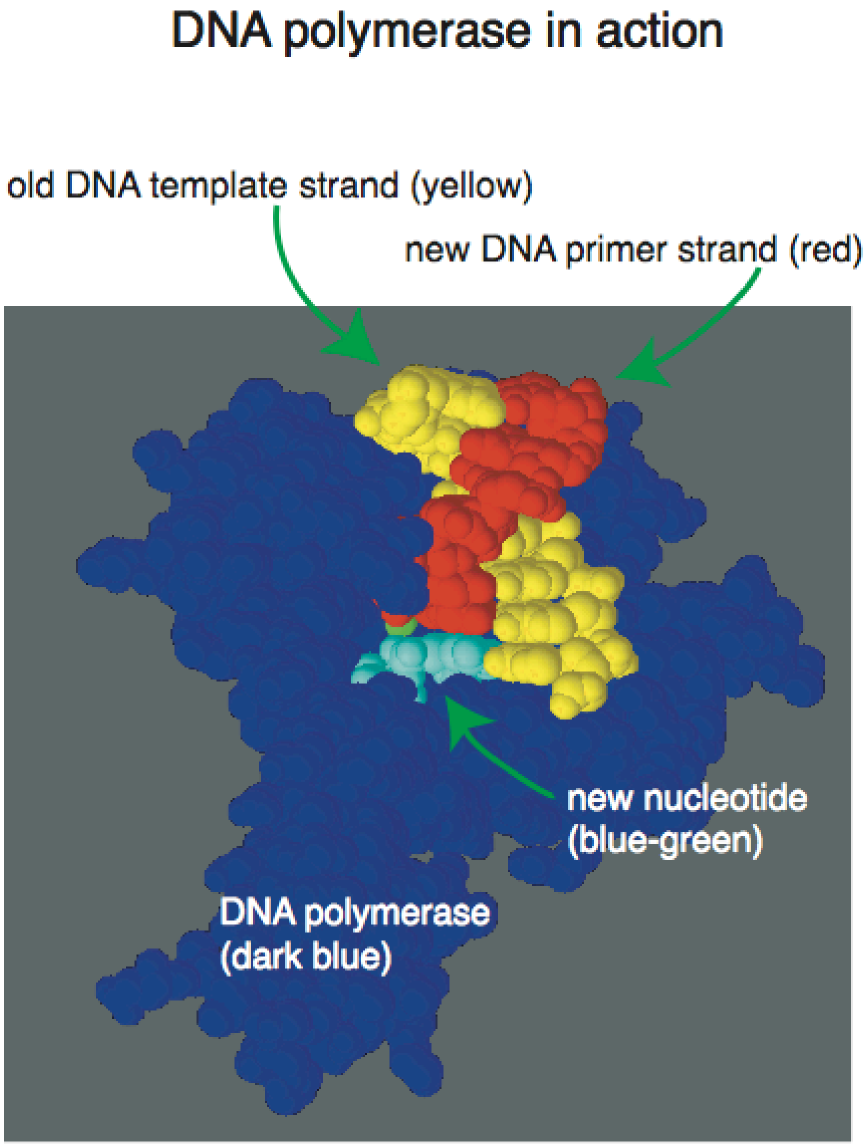
Some enzymes, in fact some that are most central to life, are made of RNA (ribonucleic acid), or RNA-protein hybrid molecules. Chemically RNA is very similar to DNA; it is a polymer of four different nucleotides. However, its functional role and three dimensional structure are quite different from DNA.

DNA polymerase

The enzyme that binds and reacts with the nucleotide triphosphate and DNA to first form a transient intermediate and then adds the nucleotide to the DNA is the protein called DNA polymerase[1].

[1] There are rules for constructing enzyme names. Typically the first part of the name is one of the characteristic substrates, the middle part describes the process (here the enzyme is linking units together to form a polymer), and the suffix "ase" indicates the object is an enzyme.

Figure dna_polymerase



dna_polymerase

Figure dna_polymerase. The large polymerase enzyme binds to the growing end of the new DNA strand and forms a pocket for the new nucleotide. For the new nucleotide to form a stable complex it must be complementary to the nucleotide on the template strand.

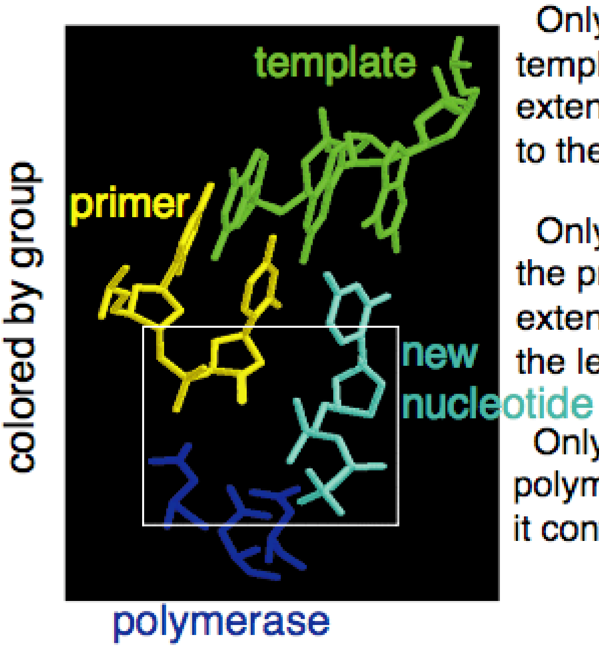
The polymerase can be thought of as a scaffold, a fairly rigid object with a pocket shaped to fit around the growing end of the DNA. It binds to the DNA by a large number of weak, non-covalent bonds, which are however, sufficiently strong to modify and define the geometry of the DNA helix in the binding region. The incoming nucleoside triphosphate starts to bind to the polymerase even before a complementary base pair has formed. The polymerase thus stabilizes a potentially successful complex, increasing the probability it will progress along the reaction path to final addition. It can increase or decrease the probability of an incorrect nucleoside triphosphate being incorporated into the growing DNA as well as increasing the speed of the reaction. The polymerase can define almost any aspect of the nucleoside triphosphate addition, except of course the overall energy balance; the energy lost or gained during the reaction is determined solely by the molecules at the beginning and end of the reaction.

Chemical bonds are formed

Formation of the transition molecule occurs at the active site of the polymerase, the region where old bonds are replaced by new. Obtaining a realistic picture of the active site is not so easy because it is intrinsically three-dimensional, located in a cleft of the large polymerase molecule. The polymerase is a polymer of 335 amino acids and the average amino acid is about as complex as the deoxyribose sugar. The only way to obtain a detailed understanding of the structure would be to spend an hour in front of a computer monitor twisting and turning a virtual model of the complex to see all the important spatial relationships (to do this see Appendix www_structure). We do our best here by looking at Figure DNAp_transition.

Figure polXitionComplex

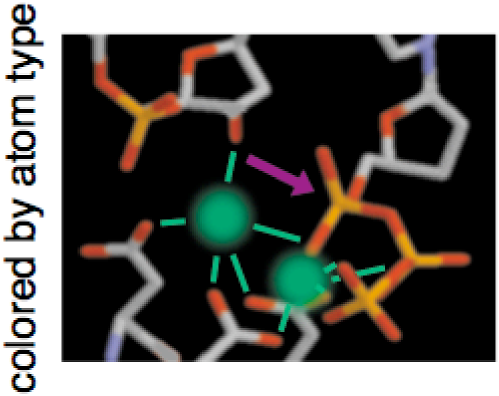
Adding a nucleotide: the transition complex



Only three nucleotides of the template are shown; it typically extends millions of nucleotides to the left and right in this figure

Only the last two nucleotides of the primer are shown; it typically extends millions of nucleotides to the left.

Only three amino acids of the polymerase enzyme are shown; it contains a total of 335.



Negative oxygen atoms (red) of the terminal primer ribose, the new nucleotide, and polymerase are linked by two positive magnesium ions (green).

Electrons associated with the primer ribose oxygen move toward the positive phosphate on the new nucleotide (purple arrow) to form a covalent bond.

polXitionComplex

Figure polXitionComplex. In the top panel we have zoomed in on the polymerase-DNA complex of the previous Figure to show the area where the new bond between the incoming nucleotide and the growing end of the primer DNA chain will be made. The bottom panel enlarges the area in the white rectangle of the top panel, reveals two magnesium atoms, and colors the stick model according to the atom type. The purple arrow indicates the migration of electrons that creates the new bond.

The top panel shows four components that make up the transition complex. The last two nucleotides of the growing DNA chain are in yellow. Each base of this primer forms a hydrogen-bonded pair with the template DNA chain, shown in green. The new, incoming nucleotide triphosphate, in light blue, is paired with the next nucleotide in the green template. Portions of three amino acids (all aspartic acid) of the polymerase are represented in deep blue at the bottom of the panel.

The bottom panel is an enlargement of the lower middle portion of the top panel. The two diffuse green spheres are positively charged magnesium ions that bond together red electro-negative oxygen atoms in the primer, in the polymerase, and in the nucleotide triphosphate. Not all enzymes have metal atoms at the active site, but many do. Metals that are essential components of other enzymes and proteins include iron, copper, cobalt, and zinc.

This is important chemistry

We have only started the description of DNA replication, but already have covered fundamental chemical paradigms that make life possible. Indeed, the motivation for examining DNA nucleotide addition in detail was not just that it was important, after all replication is the central property of life, but also that we could use this process as an example. Several generalizations can now be made about biosynthetic reactions, when typically a small molecule is added to a larger one.

The overall process must run down an energy hill; the chemical energy of the starting substrates must be greater than the end products. To satisfy this condition the starting substrates are usually not just fragments of the desired product. At least one of the reactants linked to chemical groups by bonds that release energy when replaced by the desired bond. Pyrophosphate is often the group that is released when the new bond is made.

The initial stages of the reaction consists of random diffusion of the small reacting substrates. As the reaction progresses diffusion is increasingly modified by the micro-environment, a large enzyme molecule. If the small molecule has the correct geometry and electronic distribution, it forms a complex with the enzyme. The small molecule is then added to the larger one in a second step. After this step the reaction is essentially irreversible. Without the enzyme the reaction proceeds with such a low probability that it would normally be undetectable.

The entire process can be considered as evolution at a molecular level. In the initial stage many molecular configurations occur sequentially by random thermal motion. These are analogous to a large, diverse collection of individual organisms. If and when a configuration occurs that satisfies detailed geometric and electronic specifications, the electronic distribution around a few atoms shift to form new bonds. This new molecule is stable compared to all the previous configurations. This is analogous to an individual being born that has a survival and reproductive advantage. The individual, its offspring, and its genes have a higher stability than the rest of the population.

The replication fork

The previous section described the synthesis, or extension, of a single strand of DNA by an enzyme, using the complementary DNA strand as the template. If one strand of a double stranded DNA helix is completely copied by this process the result will be one DNA double stranded helix containing one old and one new strand, and one single strand of old DNA.

However, DNA in the cell is a double stranded helix, and the replication process must generate two new double stranded helixes. If DNA replication were simple the two parental strands would separate and each strand would be copied from opposite directions to form two progeny helixes. However, in the real world both strands of the DNA helix are copied in a small regions, the replication forks. The forks moves along the parental DNA helix and two new progeny DNA helixes are formed in its wake. However, there are two obvious problems in making the fork machine work.

The parental DNA helix must be unwound

In order for the helix to be replicated, for the fork to move, the parental DNA helix must be unwound. You could imagine that this might be accomplished by either rotating the entire parental helix, or by rotating the two daughter helixes around each other. Either one of these solutions might look feasible from Figure fork1, but in the real world the helixes extend in each direction from the replication fork for millions of bases, and thus can’t be twisted at the speed necessary for effective replication.

Figure fork1

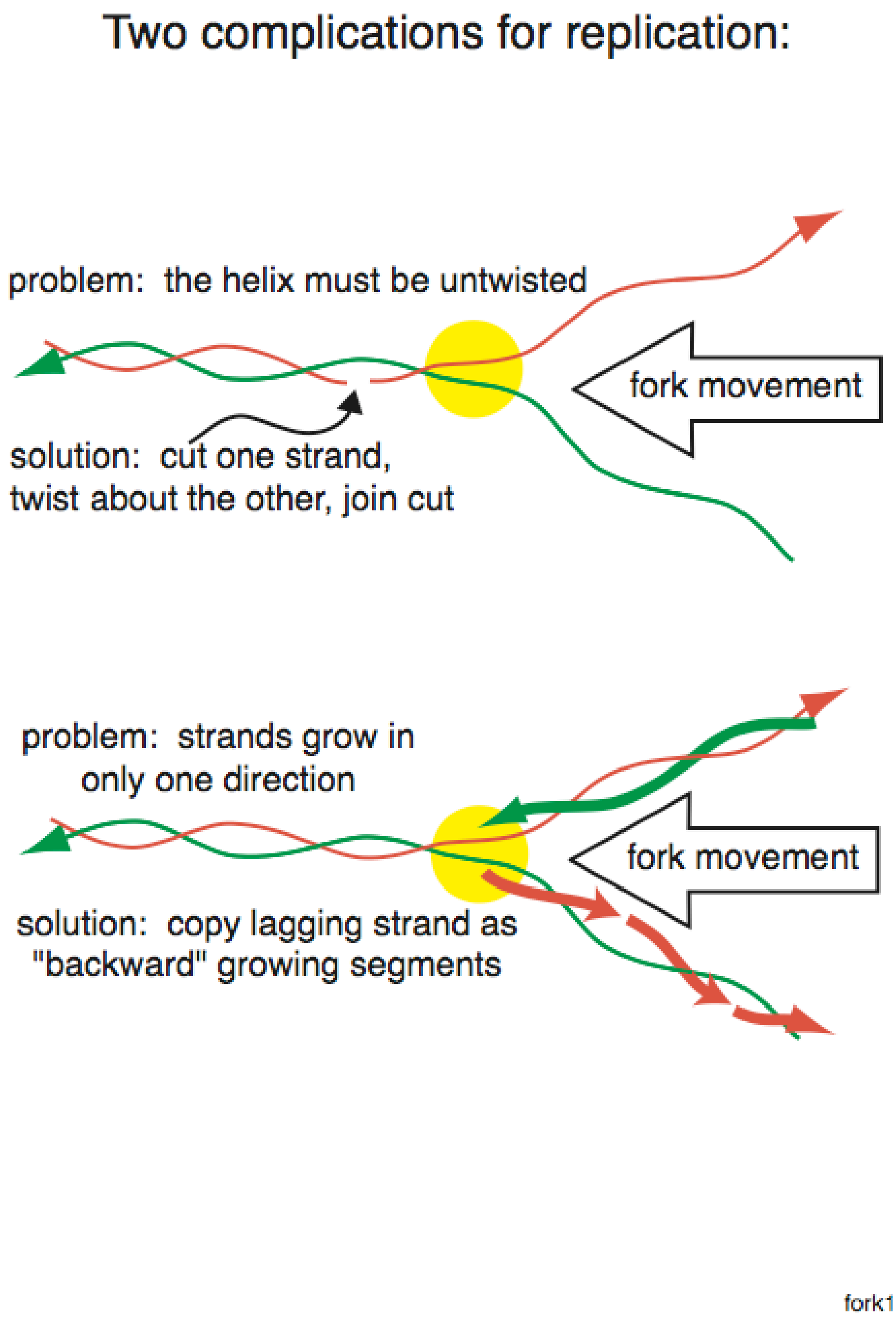

Two complications for replication:
problem: the helix must be untwisted
fork movement
solution: cut one strand,
twist about the other, join cut
problem: strands grow in
only one direction
fork movement
solution: copy lagging strand as
"backward" growing segments
fork1

Figure fork1. The top panel shows how the old DNA helix is unwound for replication. One strand is cut, the replication fork complex rotates, and the break is repaired. In the bottom panel, we see that only the red strand can be copied by polymerase in the same direction that the replication fork moves. The green strand must be replicated in the opposite direction. Thus, the green strand must be replicated in short segments by a polymerase moving in the direction opposite to the replication fork.

The solution is direct, if apparently inelegant. An enzyme cuts one of the parental strands, allowing the helix to twist for a few turns around the remaining strand, after which the same enzyme reforms the bond to repair the helix. The need to rotate a DNA helix is common to many steps in DNA processing, and enzymes that allow rotation are called helicases. The energy in the strained helix may be sufficient to accomplish the transfer of bond connectivity or ATP hydrolysis may be coupled to the reaction to supply the energy.

DNA polymerase works in only one direction

Addition of a nucleotide triphosphate to the end of a DNA chain is a very specific process, as seen in Figures dna_polymerase and polXitionComplex. Specific atoms in the primer, template, and incoming nucleotide triphosphate must be in precise positions relative to specific atoms in the large, asymmetric polymerase. Thus, the nucleotide can only be added at the 3′ hydroxyl group of the ribose at the end of the growing chain. However, to copy a double stranded helix you need to copy both strands, and only one, represented by the red line in Figure fork1, can be extended in the same direction the replication fork is moving. This strand is called the leading strand. The strand represented by the green line, called the lagging strand, must be extended in the direction opposite from the direction the replication fork, in short segments (Okazaki fragments, named after their discoverer) by the polymerase moving in the direction opposite to the replication fork. Then these segments must be joined to form a continuous strand. Joining DNA segments is also a common process, and the enzymes that perform this function are called ligases.

New strands must be started

In our descriptions of DNA synthesis strands are always extended. The verb extended was carefully chosen because in fact DNA polymerases can only extend a preexisting primer strand. Replication of the leading strand, requires only one initiation event at the start of the process, and thus we might not question how that is done. However, new DNA strands need to be initiated every thousand or so bases in the process of copying the lagging strand in short segments.

Initiation is accomplished by a primase. Even this polymerase can't initiate a DNA strand, but rather starts out making a complementary RNA chain (RNA is very similar to DNA, as we will see in the next chapter). After ten or so bases the primase switches to using deoxyribonucleotide triphosphates for a few dozen more bases at which time it stops and falls of the DNA template strand. This primer strand can then be extended by "ordinary" DNA polymerases. Later we will need to get rid of the RNA head segments before the DNA fragments can be joined by a ligase.

The enzyme that removes most of the RNA head is called RNase H1 (H because it works on an RNA chain that is part of a helix). However, RNase H1 can not remove the last ribonucleotide. Another enzyme involved in trimming the RNA head is the flap endonuclease (FEN1 protein). The flap in this enzyme's name comes from the fact that the RNA segment is typically displaced from the DNA template by the polymerase that is extending the upstream region of the strand. But the flap nuclease can't remove the last RNA nucleotide if it is tightly base paired to the DNA template, and several other enzymes (which we mercifully don't mention here) are required to do this job. This is just another case of an apparently simple problem requiring a surprisingly large collection of enzymes to do the job.

Single stranded DNA must be transiently stabilized

As the replication fork moves along, there will be segments of both stands that need to remain single strands for a short time. If the two exposed parental strands reformed helix replication would be slowed or stopped. In addition, stretches of DNA several thousand bases long may contain short sequences that are complementary by chance, and such segments could hybridize to from "hairpin loops" which could also slow or stop replication. Proteins that bind to and thus stabilize single stranded DNA solve this problem.

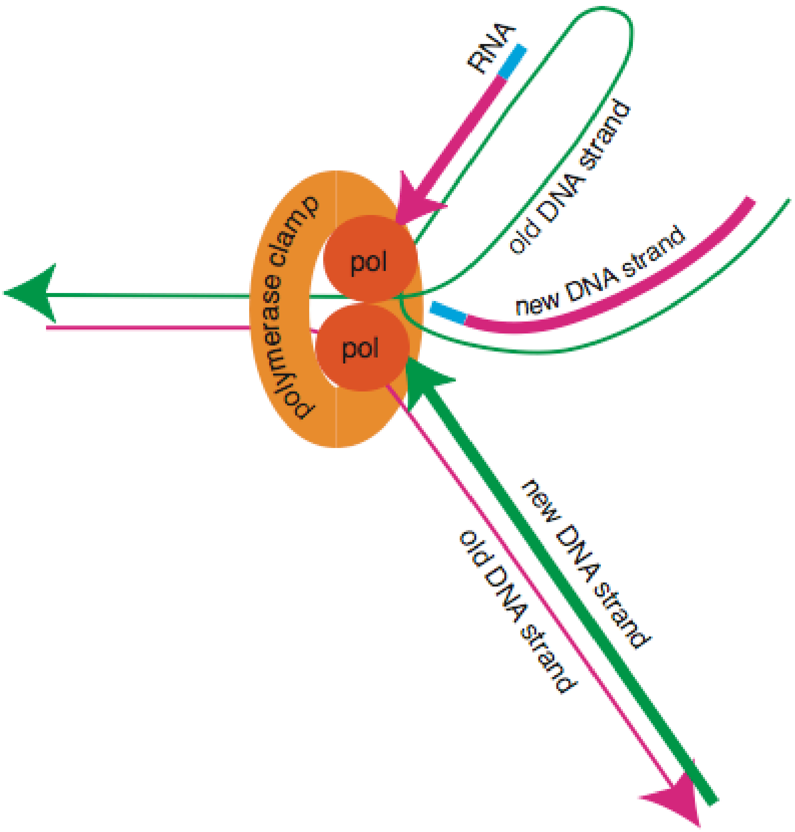
DNA replication must be highly progressive

In the context of DNA synthesis progressive has nothing to do with politics, but means that once a replication fork starts it must progress on its way down the DNA helix without falling off. The typical replication fork finishes its job by running into another replication fork moving in the opposite direction, although some will end up at the physical end of the helix. Finishing at the end of a helix is a big and important problem, but we put it off for the time being.

Progression is important because control of DNA synthesis is mainly at the initiation stage. Control of DNA synthesis is important in many situations, but consider the most obvious; the DNA in one cell must be copied only once. When the cell divides there must be exactly two copies, one for each daughter cell. After a replication fork has been created at a specific initiation site, no new replication fork is started at that site or anywhere on the segment the fork is replicating. This strategy for balanced replication works only if all replication forks proceed to the end without falling off the helix.

Figure fork2

Enzymes at the replication fork



fork2

Figure fork2. This is an enlarged view of the replication fork seen in the previous Figure. Now we see that there must be at least two polymerase molecules at work. The lower polymerase can just copy the old DNA strand, since it moves in the same direction as the replication fork. Polymerases copying the other strand must make short segments of DNA, which will be later joined to make one strand. Each short segment is initiated as an RNA copy by another enzyme, and the RNA fragment is extended by the DNA polymerase. The RNA fragment is then removed.

A high probability of progression to the end is obtained by using a molecular clamp; a trimer of protein subunits that form a circle around the helix as the replication fork is created. The construction of this processivity factor around the replication fork is aided by a multi-subunit loading protein, for historical reasons called replication factor C. The clamp stays at the fork as it moves along the helix and helps hold everything together while replication factor C is only needed during fork initiation.

The replisome: the cluster of enzymes at the replication fork

We have only described some of the enzyme subunits that are found at the replication complex. The complete molecular machine needed to copy a DNA double helix at the replication fork is called a replisome. Some of the enzymes associated with the replisome are listed in Table replisome.

NAME	STRUCTURE	FUNCTION
DNA polymerase δ	125, 50 kd	adds nucleotide, proofreads
proliferating cell nuclear antigen (PCNA)	three 36 kd	circles fork to insure progression of the polymerases
replication factor C	140, 40, 38, 37, 36 kd	loads PCNA
DNA helicases (several)	various	allows parental helix to rotate
primase	180, 70, 58, 48 kd	starts chains
replication factor A	70, 34, 11 kd	stabilizes single stranded DNA segments
FEN-1	46 kd	removes RNA
RNase H1		removes RNA
DNA ligase	various	joins Okazaki fragments

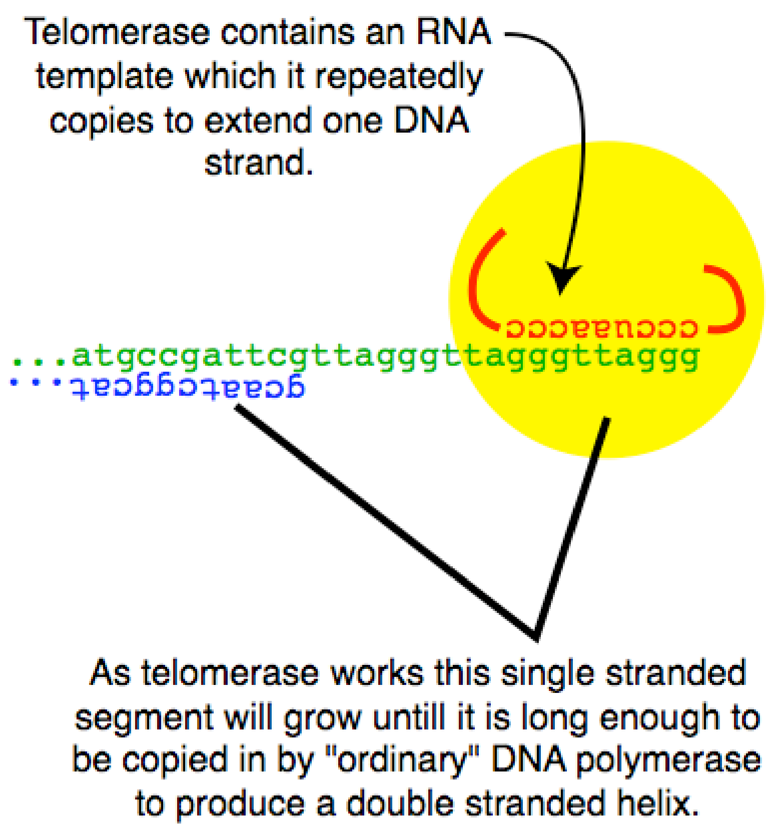
The telomere at the end of the helix

DNA polymerase can extend the leading chain to the end with no problem, it just keeps going. However, the lagging strand is a problem because the primase can only initiate a new strand every several thousand nucleotides or so. Thus there will be a terminal segment of the lagging strand that can not be replicated. Unless something is done to solve this problem each successive generation of the DNA double helix would have progressively shorter lagging strands at the ends.

The solution to losing ends is the enzyme telomerase which adds a series of short, identical nucleotide sequences to the 3' ends of DNA strands. This repeating sequence is entirely defined by the telomerase, it is not the complement of a parental DNA strand nor does it carry genetic information. As long as the group of telomere sequences at the end of the DNA strand is longer than the region that is not copied during generation of the lagging strand, all that is lost during the replication is telomere sequences, and they can be regenerated later by other telomerase enzyme molecules.

Figure telomere

DNA replication at the end



Telomerase contains an RNA template which it repeatedly copies to extend one DNA strand.

As telomerase works this single stranded segment will grow untill it is long enough to be copied in by "ordinary" DNA polymerase to produce a double stranded helix.

telomer

Figure telomere. Telomeres are repetitive sequences at the ends of DNA molecules which are made by telomerase, a very special DNA polymerase (yellow). This enzyme contains a RNA chain (red) which pairs with the DNA telomere sequence (green). The adjacent segment of the RNA is then used as a template to extend the DNA with another telomere fragment. The telomerase then shifts to the right and repeats the process. The complementary DNA strand (blue) is made later by a "regular" DNA polymerase.

Telomerase is able to extend a DNA strand without the normal complementary template because it has a template as an integral component. This template is an RNA chain. One segment of the RNA pairs with a sequence on the existing DNA, and the other segment provides a template that can be copied to extend the DNA chain. Note that one consequence of this mechanism is that once a DNA strand has lost all copies of the telomere sequence it can no longer be extended by telomerase.

Many cells in the adult body don't need to replicate many times. In fact a few, e.g. red blood cells, don't have DNA and thus can't replicate even once. However, other tissues must be constantly replenished by cell replication throughout the lifetime of the animal. As examples, white blood cells, skin cells, and the endothelial cells that line the gastrointestinal system must be continually replenished by daughter cells arising from division of specific stem cells. These stem cells and the germinal cells that become eggs and sperm must be to replicate many times.

Without telomerase a cell can not divide many times, since each round of replication would produce DNA molecules with more nucleotides missing at the ends. Eventually the original telomere sequences will be gone and the shorter DNA will loose essential genes at the ends. Stem cells and tumors have high telemerase activity, which is consistent with this model for cell proliferation[2].

In addition to telomerase a number of other proteins bind near the telomere sequence and appear to function as a protective cap to keep the ends from becoming attached to other ends of DNA. End to end joining would result in monster chromosomes that could not be properly segregated to daughter cells properly.

The speed of DNA replication

It is possible to estimate the speed of DNA replication without knowledge of the detailed biochemistry of DNA polymerase or the other parts of the machinery at the replication fork: DNA in a cell must be replicated each time the cell divides. Rapidly growing human cells can divide in about 8 hours, or about 30,000 seconds. Since the total DNA per human cell is 6 X 10^9 base pairs, this means an average replication rate of 2 X 10^5 base pairs per second.

[2] Of course this model is not complete. If either telomere length or the level of telomerase (essentially the same parameter) is to control the number of times a cell type divides, there needs to be some graceful way for the lack of a telomere to stop cell division. Just loosing genes at the ends of DNA molecules could generate "junk" cells that do no good, but consume nutrients and take up space, or worse yet loose control of cell growth and division and become cancer cells. In addition, a complete model must describe how telomere length or telomerase activity is controlled.

However, experimental observation of the movement of replication forks reveals that they copy about 50 base pairs per second at a maximum. The complexity of the replication machine and the fact that the movement and function of its parts must ultimately rely on random diffusion, makes it amazing, at least to me, that it is able to move that rapidly. However, even though I may think it does a creditable job, it still moves 40,000 fold too slow to duplicate the DNA of a rapidly growing cell in the allotted time.

Figure fork3

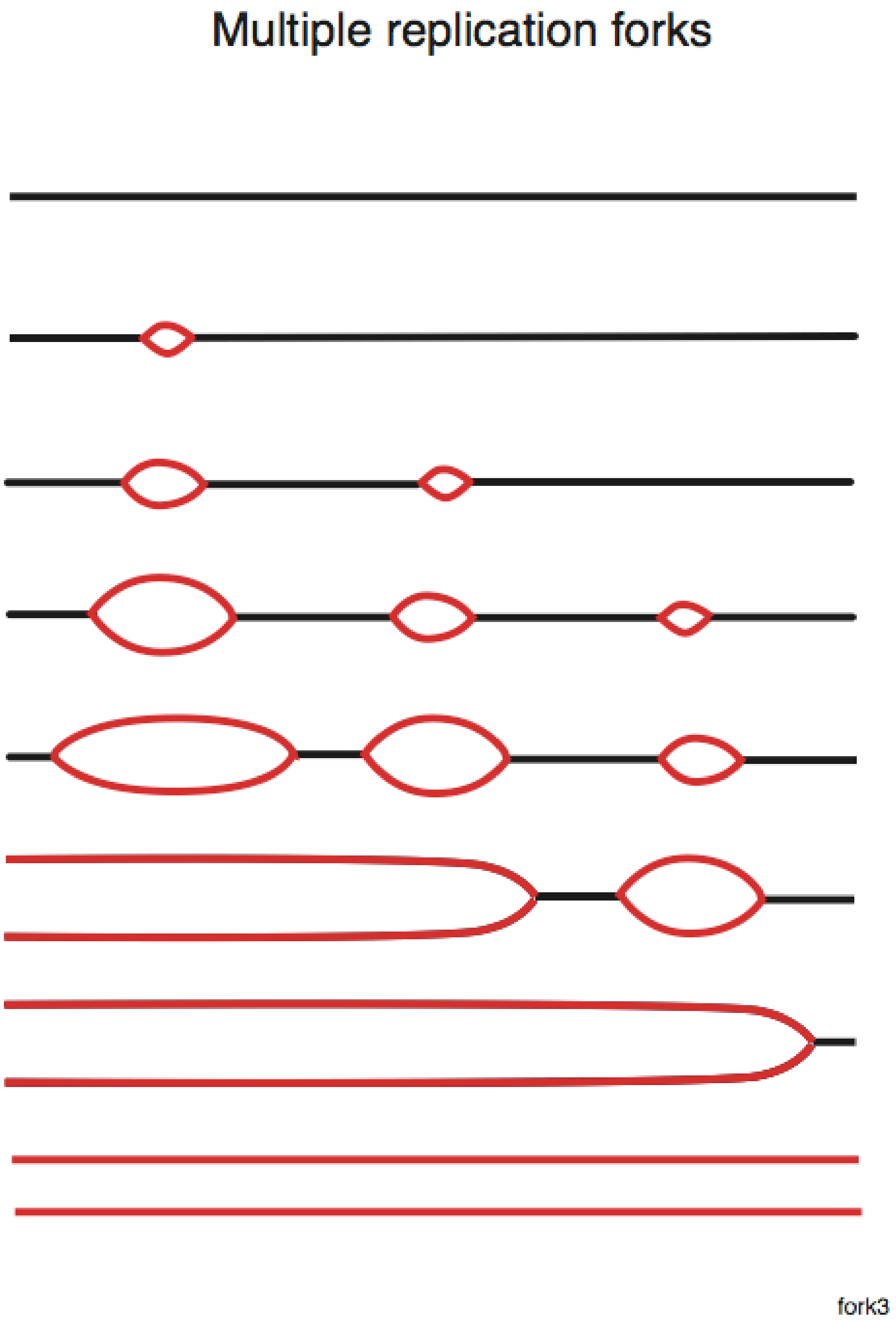


Figure fork3. Many organisms have multiple origins of replication for each DNA strand (or chromosome). Each origin of replication produces a “bubble” in this diagram showing the old (black) and new (red) strands. Each end of the bubble is a replication fork moving outward from the origin. Note that replication does not in general start at the same time for each origin. If it is difficult to increase the speed at which a DNA helix is replicated, an obvious advantage of multiple replication forks is a decrease is time required to replicate the whole chromosome. This is analogous to using multiple processors to increase the rate a computer can execute instructions.

The cell has evolved the same solution to this problem as has the computer designer who must use a slow CPU to do rapid computation: parallel processing. The computation is divided into multiple segments that can be simultaneously solved by multiple CPUs. DNA replication is well suited for parallel processing since replication of one segment can be done independently of replication of other segments. In the cell, two replication forks are assembled at a point of initiation, and each fork then moves out in opposite directions on the parental DNA helix. For humans, there are many thousands of unique points of initiation, defined by a group of specific DNA sequences, along the DNA of the cell. Initiation does not start at all these points at the same time; some are early while others start later in the process. The work of each fork ends when it runs into a fork working in the opposite direction, with the exception of forks that run into the physical end of a DNA helical chain, and that situation is handled by the telomerase system, as described in the previous section.

Error correction

Rates of error in DNA replication

Errors are made in any real communication system, and the storage and transfer of genetic information by DNA is no exception. One class of errors is the result of damage to DNA molecules, while another class is due to imperfect replication of DNA. There is a great variety of chemical and physical defects that result in incorrect or missing genetic information, and a corresponding large number of error correction mechanism have evolved to repair these defects. This contrasts with the Internet, where error correction is not typically dependent on the nature of the damage; in most cases the damage only needs to be detected and then a duplicate message is sent.

However, note that genetic errors, changes in transfer of information between parent and off spring, are a major mechanism for generating diversity. Diversity is necessary in order that selection can result in evolution of a species. The need for diversity is a fundamental difference between the Internet and life. The goal of the Internet is to transmit information faithfully. The goal of life is to survive, and faithful transmission of genetic information from parent to child is merely one means to this end.

In one human generation the average probability of a change is about 10^{-8} / nucleotide. Errors are more likely to occur in the information transmitted from the father than from the mother because sperm undergoes many more divisions per generation than the egg. In one cell division the error rate is about 10^{-11} / nucleotide[3]. These values are net error rates, e.g. the difference between the rate of error generation and the rate of error correction.

The rate of error generation is partially determined by chemical properties of the nucleotides, e.g. the relative proportion of keto and enol forms of the bases, but is also defined by the characteristics of the enzymes involved in DNA replication and the chemical environment within the cell. While some cellular environmental factors, e.g. temperature and pH, can be considered as fixed, there are many others, e.g. oxidants and other DNA reactants, are controlled by cellular enzyme activities. Finally, the effectiveness of the error corrections systems is defined by the nature of those systems. Thus, the rate of error in DNA replication, like all the other parameters of the living organism, are the result of evolution, and are presumably optimal values for the best survival probability of the species.

It is difficult to demonstrate that the rate of DNA replication errors in humans is the result of genetic selection. However, mutants of bacteria and bacterial viruses have been isolated that have both a lower and higher error rate than normal[4]. This demonstrates that DNA replication error rate is not a constant defined by fundamental chemical principles, but rather is determined by the specific nature of the enzymatic machinery that replicates and repairs DNA. Since these machines can be modified by mutation, the observed mutation rates found in organisms isolated “from nature” must be the result of evolution.

The ability to repair DNA damage caused by external agents varies from species to species, as would be expected if the synthesis and repair machinery is also variable from one species to the other. An extreme example is the bacterium *D. radiodurans* R1. This organism is 20 times more resistant to UV irradiation and 200 times more resistant to ionizing radiation as E. coli, the major species found in the gut of mammals. The nucleotide sequence of the genome of *D. radiodurans* R1 reveals many features that explain the ability to survive in harsh conditions. During rapid growth, multiple copies of the genome are present, and there are many sequence repeats within the genome. These two features allow recombination between genomes to rescue intact genes. There is a unique mechanism for transporting damaged nucleotides out of the cell, minimizing the chance they will be incorporated into DNA. There are additional features of the cellular machinery that contribute to the organisms resistance to damage, but the important point has been made; resistance to damage is determined by the specific nature of the repair machinery, and that is a characteristic of a species.

[3] These error rates are of such general interest that many estimates have been made. See “Rates of Spontaneous Mutation” by Drake et al., Genetics 148:1667 (1998) for typical values. A change in genetic information is called a mutation, and typically causes a loss or change in function of the gene produce.

[4] See “DNA Polymerase Fidelity: from Genetics Toward a Biochemical Understanding: by Goodman and Kuchnir, Genetics 148:1475 (1998) for an introduction to the literature on this area.

Proofreading

Normally during DNA replication the complementary nucleotide triphosphate makes a hydrogen-bonded pair with the nucleotide on the strand being copied. This base pair has a special geometry and charge distribution, which of course is the basis for it being selected and then incorporated. After the phosphodiester bond has been made, this special geometry results in a more stable structure in which the position of the new nucleotide is relatively immobile. However, if one of the three other nucleotides has been incorporated by mistake, it will not be stabilized by correct base pairing and will wiggle around in thermal motion. In its movement, it is likely to bind to another active site on the polymerase. This active site catalyzes the rearrangement of the bond that has just been made with one of the bonds in a water molecule; the phosphodiester bond is broken. The incorrect nucleotide then can diffuse away as a nucleotide monophosphate, and the original end of the DNA chain is exposed. Now the nucleotide addition process can be repeated, hopefully correctly this time. Many viral and bacterial mutants with abnormal mutation rates have been found to have a DNA polymerase with altered proofreading activities; generally inversely correlated with the mutation rate. This observation certainly suggests that proofreading is an important determinate of mutation rate.

Figure repair_proof

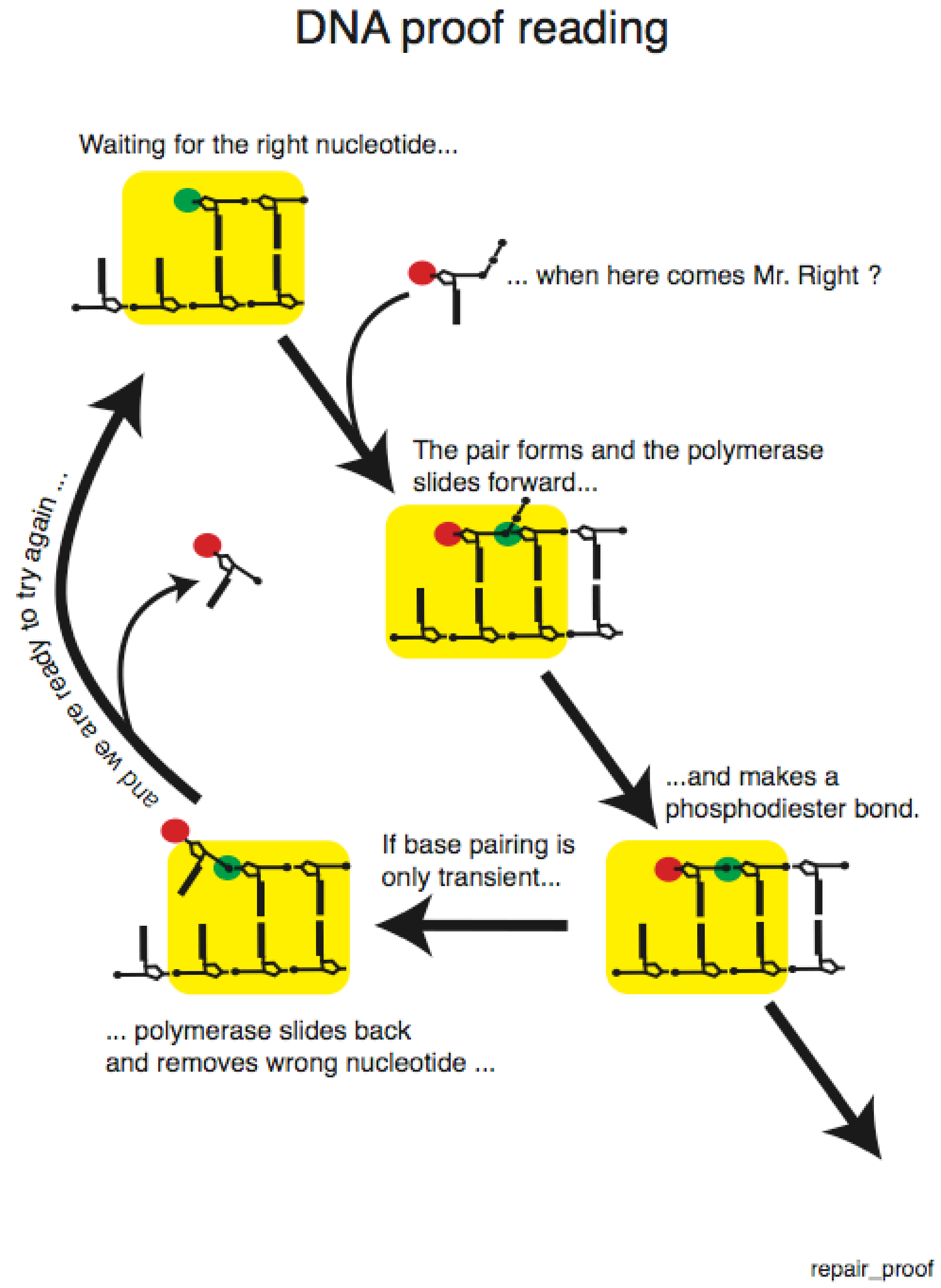

DNA proof reading
Waiting for the right nucleotide...
... when here comes Mr. Right ?
The pair forms and the polymerase slides forward...
...and makes a phosphodiester bond.
If base pairing is only transient...
... polymerase slides back and removes wrong nucleotide ...
and we are ready to try again ...
repair_proof

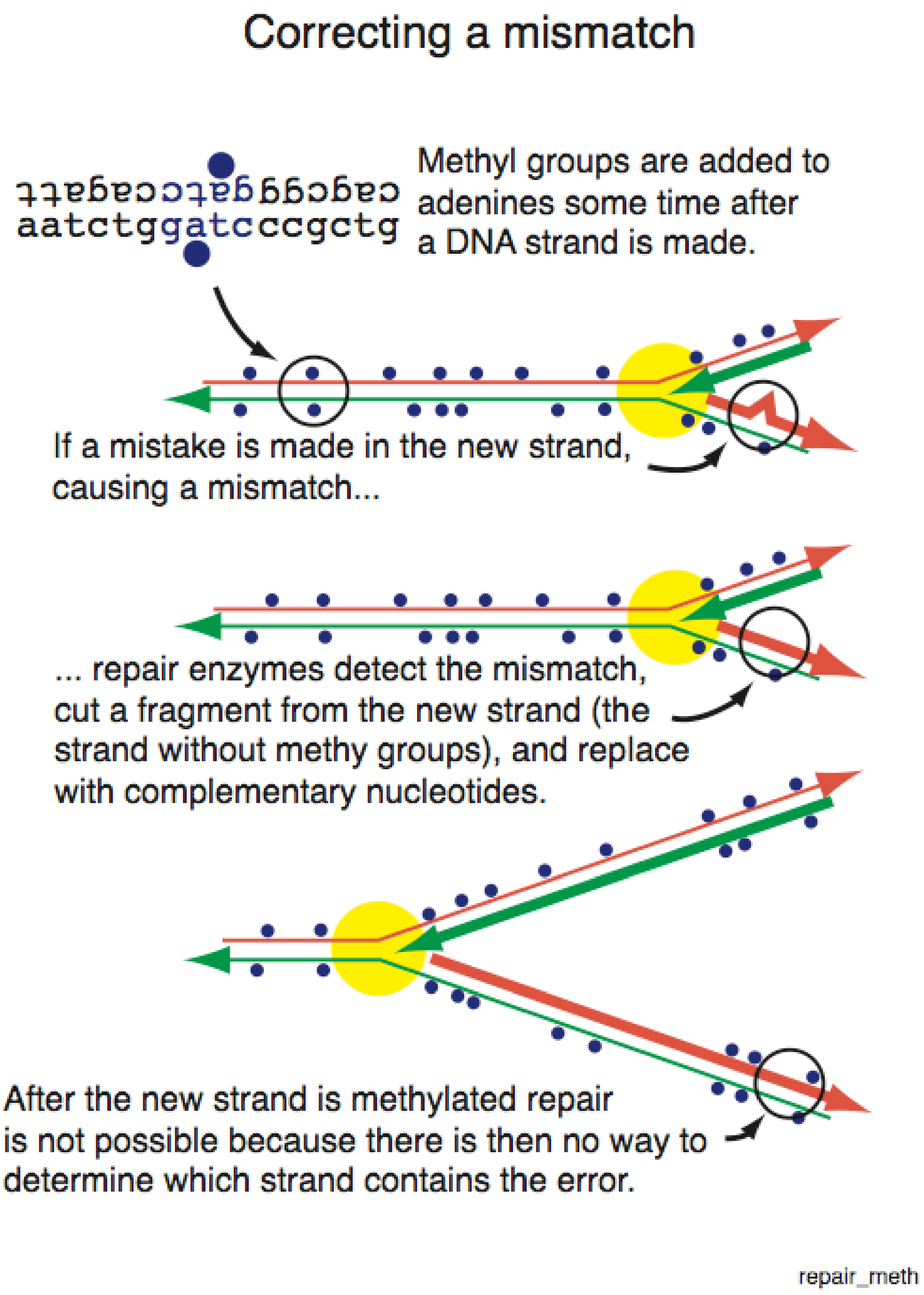
Figure repair_proof. DNA polymerase can both make and break the phosphodiester bond that holds the nucleotides together in a DNA strand. If the polymerase has added a nucleotide to the growing chain that does not form a stable hydrogen bonded pair with the template strand, the polymerase cleaves the bond it just made and the process of adding the next nucleotide starts all over again.

Of course proofreading requires energy; all corrections do. The discarded nucleotide monophosphate must be converted back to a nucleotide triphosphate before it can be used again. It is analogous to detecting that a packet has been corrupted because it doesn't have the correct check sum and discarding it. The packet must be transmitted again, which takes time and energy.

Mismatch repair

Even if a mismatched nucleotide has been added and the polymerase has moved on without proofreading, the bad nucleotide may be removed by another enzyme system that scans DNA for nucleotides that do not have the correct geometry. This enzyme cuts out the mismatched nucleotide, as well as some on either side, and the resulting gap is then filled in correctly by another DNA polymerase. It may not seem surprising that the repair enzymes can recognize a miss-paired base pair, but how does it recognize which base is the incorrect one?

Figure repair_meth



Correcting a mismatch

Methyl groups are added to adenines some time after a DNA strand is made.

If a mistake is made in the new strand, causing a mismatch...

... repair enzymes detect the mismatch, cut a fragment from the new strand (the strand without methy groups), and replace with complementary nucleotides.

After the new strand is methylated repair is not possible because there is then no way to determine which strand contains the error.

repair_meth

Figure repair_meth. Some time after the original DNA strands (thin lines) are copied by the polymerase to make complementary new strands (thick lines), an enzyme adds methyl groups to selected adenines. In the time lag before the methyl groups are added, the new strands can thus be recognized. Repair enzymes run along the unmethylated DNA strand looking for a base mismatch. If found it is removed and replaced with a base that pairs correctly with the old (methylated) strand.

Identification of the strand containing the mistake require another enzyme. This enzyme moves along the DNA helix after it has been replicated and attaches a methyl group to adenines that are part of the sequence **`gatc`**. These methyl groups are a signal to the repair system that the strand is old; at least old relative to regions that do not have methyl groups. If there is a mismatched base pair, the nucleotide on the younger strand, without the methyl group, must contain the incorrect nucleotide. The repair system then removes the mismatched base and inserts one that matches.

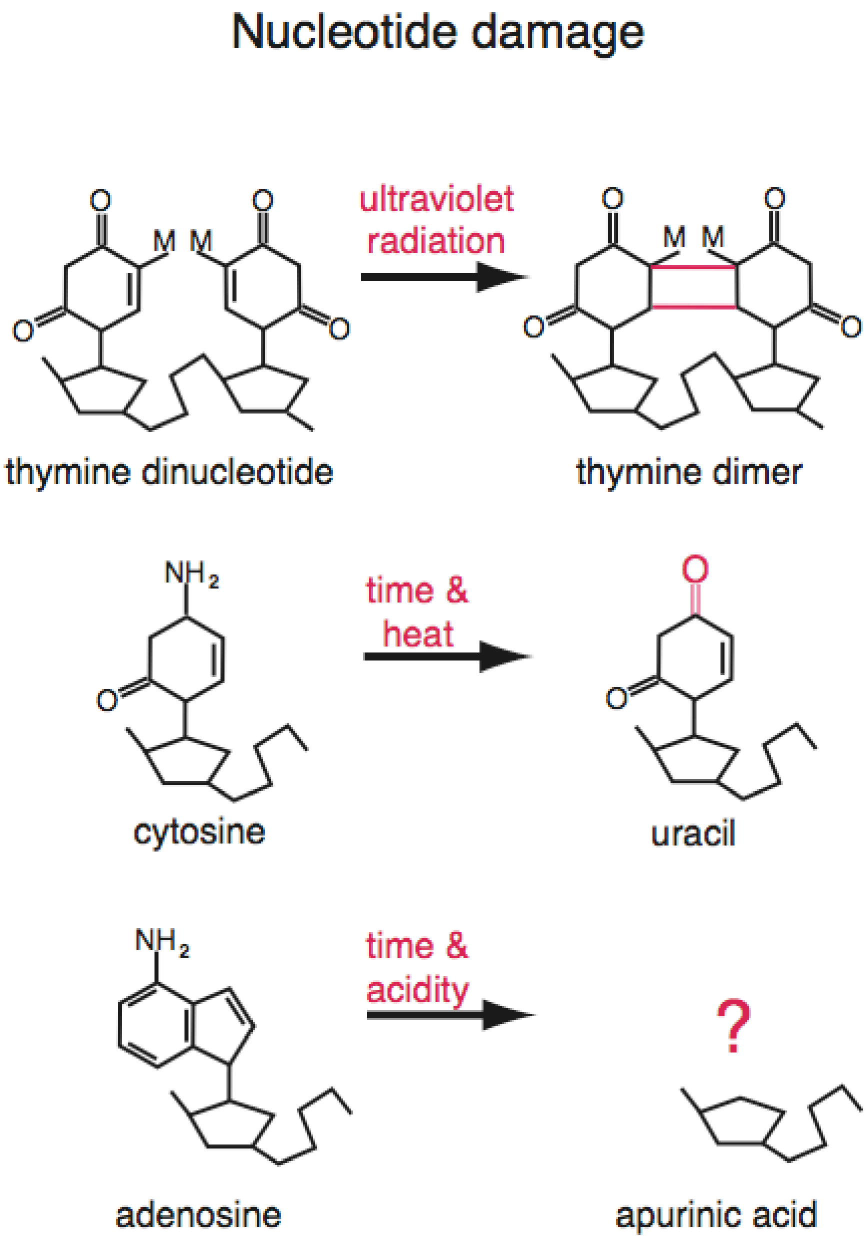
Damage repair

Even after DNA has been faithfully replicated, it may be subsequently damaged by chemical and physical agents. Ultra-violet radiation from the sun causes many unwanted chemical and physical changes to materials, as any farmer, sailor or roofer can tell you. DNA is particularly sensitive to UV irradiation due to the conjugated double bonds that absorb energy in exactly this wavelength. Adjacent thymine nucleotides on a DNA strand are a common site of UV damage. As seen in Figure repair_3, the two nucleotides become linked at the waist, so to speak. A specific repair enzyme system recognizes these thymine dimers, removes them, and a polymerase then fills in the gap by copying the complimentary strand.

The amino group on the cytosine ring is comparatively labile, and can be spontaneously converted to a keto group, generating the base uracil. Since uracil is not found in DNA, the repair enzyme for this defect has no difficulty detecting this damage and then repairing it.

When the link between the sugar and base is broken, which occurs most frequently in the adenine and guanine nucleotides, the identity of the base is lost. But DNA is a double helix, and thus the information is contained in the other strand. A segment of the damaged strand is cut out and a new segment added by a polymerase.

Figure repair_3



repair_3

Figure repair_3. Three kinds of DNA damage are seem in this Figure: formation of a thymine dimmer, conversion of cytosine to uracil, and loss of a adenine or guanine base. Each of these can be recognized by enzymes and repaired.

There are other types of chemical damage that can occur to DNA and there are corresponding enzyme repair systems; but you get the general picture by now. As mentioned before, on the Internet the exact nature of the damage to the digital signal is not usually important; if it is damaged it is just sent again.

Extensive DNA damage halts replication

If DNA in a cell has been damaged, e.g. by UV exposure, products generated by this damage cause repair enzyme systems to be activated. Severe DNA damage will halt DNA replication throughout the cell, allowing the repair enzymes to do their work before the damage is copied into progeny cells. If the system that halts DNA replication in response to damage is not functional, information in the genome will be garbled badly. A cell with garbled DNA might die, which could be the best outcome. A more serious result would be a damaged cell that did not respond to the normal signals for proliferation, e.g. a cancer cell.

Apoptosis

If DNA damage is so extensive that it can not be possibly repaired, there is one final remedy; the cell commits suicide. This process is not just the failure to live, it is an active decision by the cell and requires a specific enzyme system. It is possible for a cell to die in a more passive way, perhaps induced by starvation, a rapid rise in temperature or mechanical trauma. However, apoptosis can be distinguished from these more non-specific catastrophes by the morphological changes before death, as well as a characteristic profile of enzymes that are produced to kill and digest the cell.

Summary of replication layer sites

Figure ReplicationLayer summarizes the relative positions of DNA sites that are utilized during DNA replication (the function of the centromere is best understood in relation to the cell and entire genome, and is thus described in the next chapter). Unlike the messages sent on the Internet, error correction code is not localized to a specific location. Essentially all error correction relies on the duplication of information in the two complementary strands of the DNA chain. Error correction can be considered a second phase of replication.

Figure ReplicationLayer

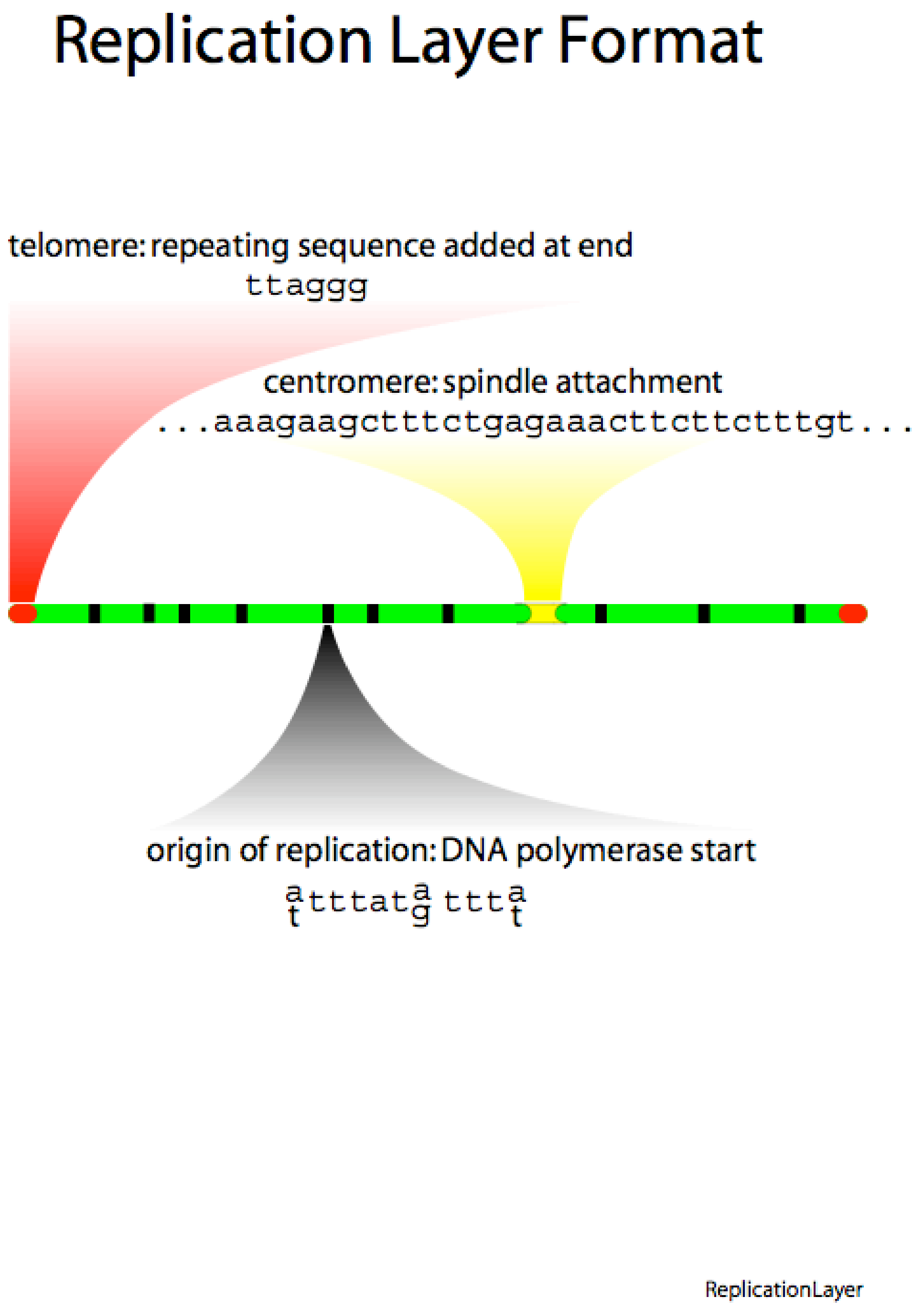

Replication Layer Format
telomere: repeating sequence added at end
ttaggg
centromere: spindle attachment
...aaagaagctttctgagaaacttcttctttgt...
origin of replication: DNA polymerase start
a tttat a ttt a
t g t
ReplicationLayer

Figure ReplicationLayer. Two polymerase complexes start replication at a number of sites scattered along the parental DNA helix, and proceeds in both directions. In some species there appear to be fairly specific nucleotide sequences at which replication starts{Gilbert, 2001 #55}, in others however, replication seems to start a almost any sequence. Special telomere sequences{Bailey, 2001 #54} at the ends of the DNA enable the special telomerase polymerase to extend the chain. The single centromere region in the interior of the molecule is the attachment point for the spindle fiber that pulls the newly replicated daughter chains apart. The centromere region is quite large, and the sequence varies between species and chromosomes in a species. In vertebrates it consists of hundreds of repeats of a slightly variable, 171 nucleotide long, AT rich “alpha satellite” sequence {Schueler, 2001 #56}. A portion of a consensus sequence is shown here{Koch, 2000 #57}.

Reading DNA: transcription and translation

DNA is the archive for genetic information, but for this information to be used it must be converted into a function. The many functions needed by the cell include structure, motion, and chemical conversion (catalysis). The detailed mechanisms for implementing these functions are best understood at the molecular level, although the roles they play in the life of the organism are often better understood at much lower magnification. The functions encoded in DNA ultimately generate macroscopic objects, e.g. you and me.

All the DNA in one organism is called its genome. The genome may be one molecule (for some bacteria), 46 plus molecules (for humans), or even more molecules for other organisms. These long DNA molecules consist of regions, or genes, which code for RNA molecules. The DNA dependent RNA polymerases are enzymes that copy a stand of DNA into a RNA molecule. RNA has essentially the same "alphabet" as DNA, thus copying the information of DNA into RNA is called transcription.

Some of these RNA molecules carry out functions directly, or are part of complexes that are functional, e.g. the telomerase. However, many RNA molecules are just intermediates in the synthesis of proteins which then become the functional molecules. Since proteins are polymers of 20 amino acids that have no structural relation to the 4 nucleotides that make up RNA, the process of converting the information in RNA to protein is called translation, a conversion of one language into another.

The newly made amino acid polymer folds into a specific shape to form the final protein molecule. This folding is determined by the amino acid sequence of the protein. In some cases folding can be demonstrated in vitro (in the test tube). In other cases correct folding occurs only in the special environment of the living cell.

Figure readingPreview

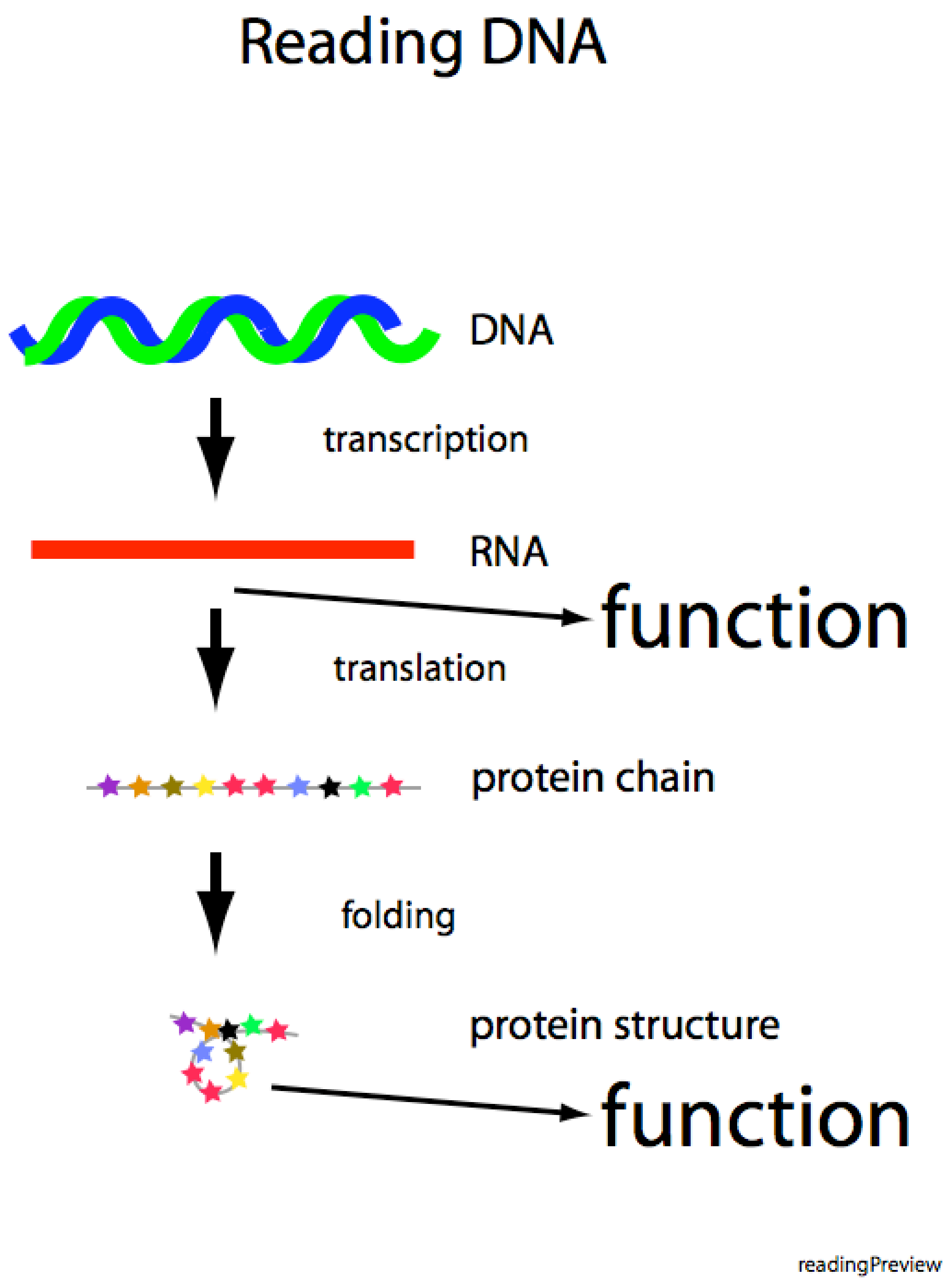


Figure readingPreview. A segment from one strand of the DNA helix is copied (transcribed) into an RNA molecule. In some cases the RNA itself performs a function, e.g. the RNA in telomerase, in most cases the information in RNA is converted to a protein molecule which performs a function.

RNA

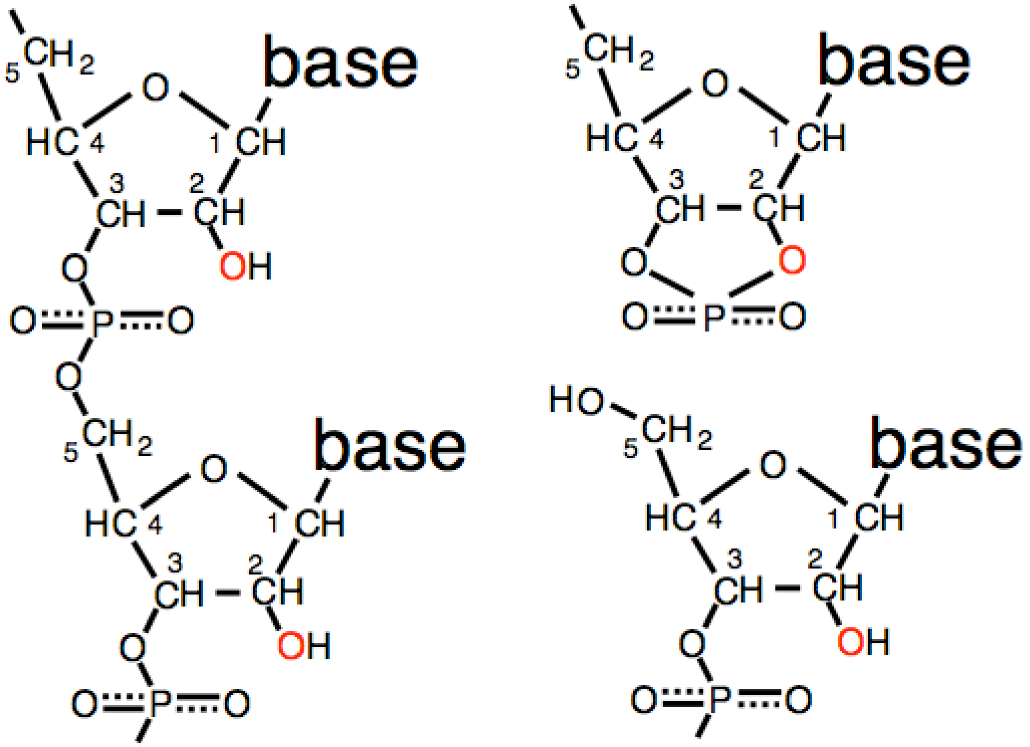
Ribose nucleic acid, RNA, is similar in structure to DNA, a polymer of four kinds of nucleotides. Three RNA nucleotide bases are identical to three of DNA, but the RNA base uracil takes the place of thymine found in DNA. The sugars in the RNA molecule contain oxygen as a hydroxyl group on carbon 2, while the sugars in DNA do not (that's what the "deoxy" in deoxyribonucleic acid means).

RNA differs from DNA in a few important ways

The "extra" hydroxyl group in RNA causes RNA chains to be more susceptible to breakage by a spontaneous chemical reaction. Specifically, ribose can form a cyclic phosphodiester using this hydroxyl group, which results in a break in the chain.

Figure RNA_break

RNA chain breaks



intact

broken

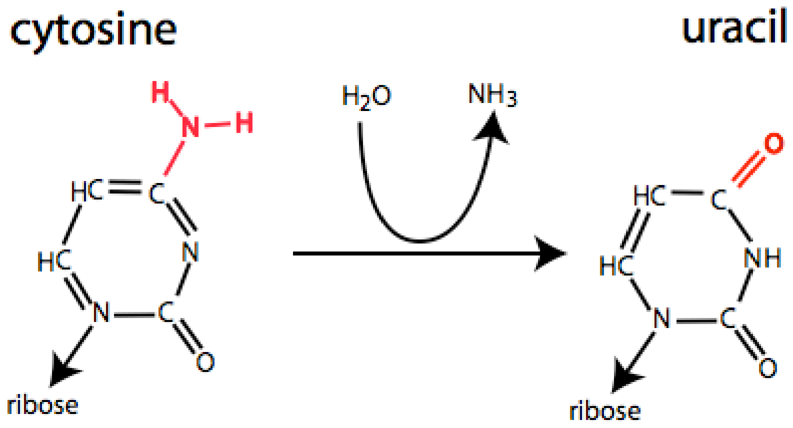
RNA_break

Figure RNA_break. The oxygen on the 2′ carbon of the ribose (red) is a liability to an RNA chain since it can form a bond to the phosphate, displacing the phosphodiester bond to the adjacent nucleotide and thus breaking the chain.

Occasionally the amino group on the ring of the cytosine base is spontaneously converted (by hydrolysis) to an oxygen containing keto group. This event converts the base to uracil. If the newly created uracil is part of an RNA chain, the message carried by the chain is altered. However, if the same process occurs on a DNA chain, the uracil is a "foreign" base, and will be recognized by a repair enzyme system and converted back to the original thymine.

Figure cytosine_uracil

Deamination of cytosine



correct base

in DNA it is recognized as an error (since it is not normally present) but it is normal in RNA, and thus isn't corrected

cytosine_uracil

Figure cytosine_uracil. The amino group on the cytosine ring is occasionally converted to a keto group, which creates uracil. Since uracil is a normal base in RNA it is not recognized as an error. DNA contains thymine in place of uracil, so any uracil in DNA must be an error, and is converted back to thymine.

DNA is generally double stranded, the mechanism for its replication insures this. However, RNA is made as a copy of only one strand of the DNA, and thus is mostly single stranded. However, short segments of RNA chains may contain self complementary sequences, and these segments thus form double stranded, helical structures. Double stranded regions are particularly important for RNA species that must have a specific structure (as opposed to a sequence) to carry out a function. Prominent examples are transfer RNA and the RNA chains that are part of the ribosome.

Making an RNA copy

A RNA copy of a segment of a DNA chain is made by an enzyme complex that binds to a specific site on the DNA, then opens up the double stranded helix, and polymerizes ribonucleotide triphosphates into an RNA chain which is complementary to one strand of the DNA helix. This is thus similar to DNA replication, except that the copy is an RNA molecule, and only one strand of the DNA is copied. The enzyme in the complex that actually makes the bonds of the RNA chain is called DNA dependent RNA polymerase (or just RNA polymerase).

Figure transcription

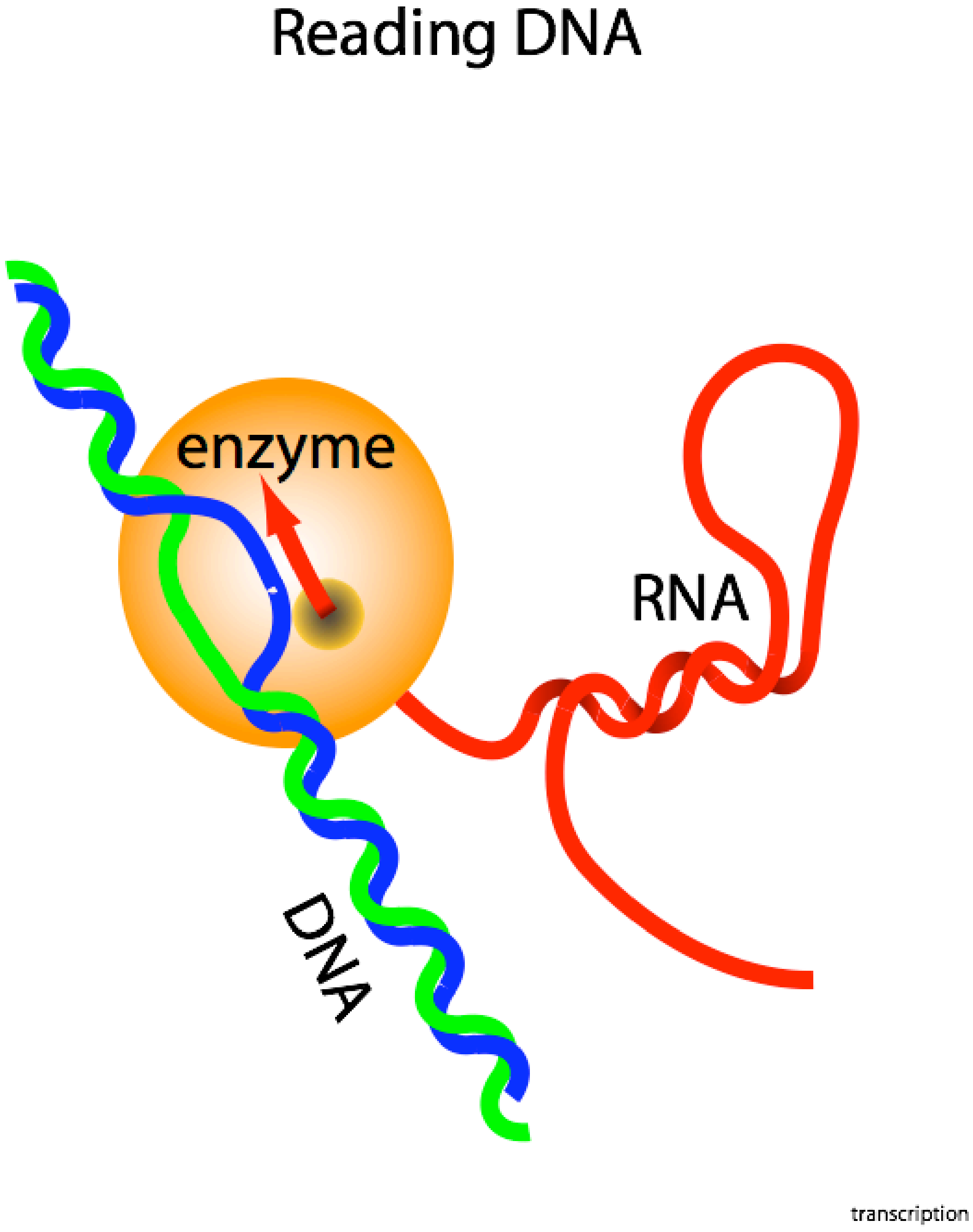


Figure transcription. The enzyme DNA dependent RNA polymerase moves along the DNA helix and copies one strand into a complementary RNA chain. To copy the DNA strand the DNA helix must be transiently unwound to form a bubble on the enzyme surface and exposing the base pairs to RNA nucleotide triphosphates that form the growing RNA chain. The new RNA may have short segments that are complementary to other nearby segments in the RNA, and thus "hairpin" structures can form.

The RNA polymerase molecule is actually a rather large molecular machine, consisting of several protein subunits and other proteins that are transiently required to start the process of transcribing the DNA into RNA. The newly made RNA chain is funneled through a tunnel in the RNA polymerase so it is removed from the region of the transcription bubble, allowing the two strands of DNA to reform the original helix.

Controlling the rate of transcription

Making RNA from DNA requires several control signals on the DNA. The RNA polymerase must know where to bind to the DNA helix and which strand to copy (this also corresponds to knowing which direction to move while copying). There must also be a signal causing the RNA polymerase to stop making an RNA copy and then fall off the DNA.

However, the rate of transcription from a particular segment of DNA is of also of great importance. All cells in our body have the same DNA (except a few that have no DNA, e.g. red blood cells) but the spectrum of RNA and protein that a cell must produce is very diverse. There are three major reasons for the need for variation in gene expression.

- As an organism develops from one fertilized egg cell to an adult different RNA and protein species are required.

- Cells in different tissues and organs must produce different products, even though they all have the same DNA. Only cells in the liver produce albumin, the major protein in the blood. Only specific cells in the pancreas produce insulin.

- Cells must alter production of proteins in response to changes in the environment. If the oxygen concentration decreases and remains low for days the production of red blood cells increases to compensate by increasing the concentration of red cells in the blood. This requires an increase in production of several proteins by cells in the bone marrow, the primary one being hemoglobin.

The rate at which a single RNA chain is made, the rate of chain elongation, is relatively constant. Thus, control of RNA production is at the initiation stage, once an RNA chain starts it is made at a standard rate until it is complete. The frequency of RNA chain initiation is modified by a collection of proteins that bind to the RNA polymerase at the promoter. Different sets of transcription enhancing proteins bind to the promoter regions of different genes. Some of these enhancers must also bind to DNA sequences thousands of base pairs upstream of the promoter, and some bind to sequences downstream. Some transcription control proteins, especially in bacteria, repress rather than enhance the rate of gene expression.

The genes that code for proteins that control the rate of transcription represent a sizable fraction of all the genes of a complex organism. The number of combinations of protein-protein and protein-DNA interactions in the network that controls gene expression must be immense. Note that while the digital information in the genes for these control proteins defines their structure, the proteins function as analog machines. The 3-D structures of the control proteins do not bind to DNA sequences in the digital base pairing scheme that controls DNA replication or transcription, and of course the protein-protein interactions do not either.

Analog control of transcription rate seems reasonable for the organism, it is analogous to analog control of the voltage output of a power supply in a computer. However, analog control means that there is no digital code for the scientist to discover. Rather there are 3-D interactions between the charged surfaces of complex proteins, and the 3-D structures of most of these proteins are not known.

Posttranscriptional modification

After an RNA molecule is made in a eukaryotic cell, several modifications are typically required for it to function. The beginning end of the chain is “capped” with a G using an unusually link and a string of A’s is added to the end of the RNA chain. In addition, most mRNA chains have specific segments cut from the middle, with the resulting ends rejoined to make an intact chain. While the removed fragments (introns) must be defined by nucleotide sequences, these are not entirely clustered at the break points. The distributed nature of this information has prevented the break point code from being completely broken (as of the year 2005).

An initially surprising nature of intron excision was that it was catalyzed by RNA, since prior to that time all chemical reactions in the cell were catalyzed by protein enzymes. To top off the surprise, in some cases the RNA was the RNA itself!

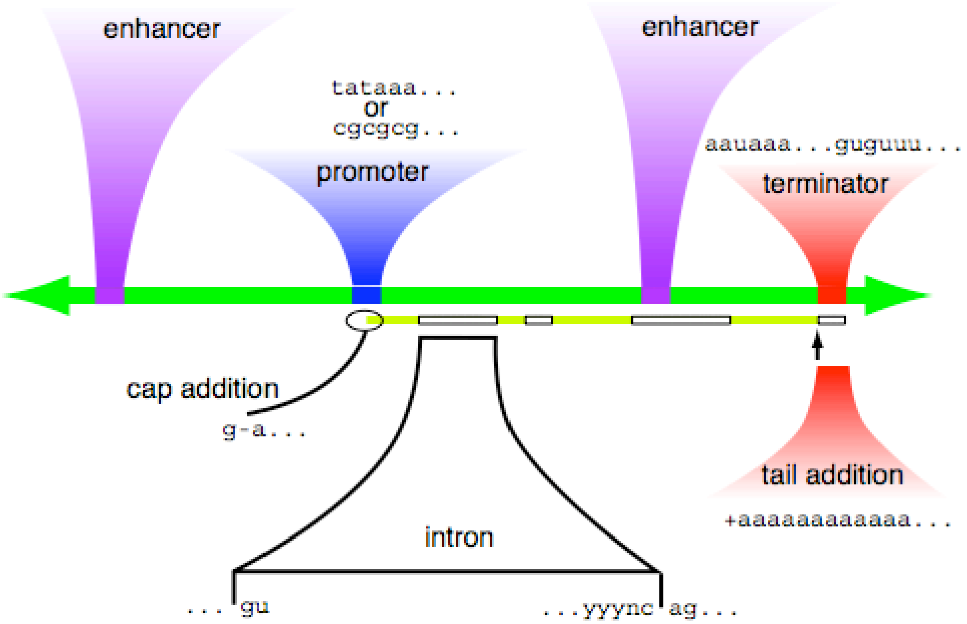
Summary of transcription layer sites

As with the replication layer there are two groups of processes. First a specific region of DNA is copied into RNA, with the number of RNAs made per unit time controlled by multiple signals. Second, the RNA chains are modified to produce a mature molecule which is transported from the nucleus to the cytoplasm.

Figure TranscriptionLayer

Transcription Layer Format

step 1. synthesis of pre-m RNA



step 2. modification to produce m RNA

TranscriptionLayer

Figure TranscriptionLayer. For each gene (transcription unit) the start of transcription is defined by a promoter sequence and the end by a termination sequence. Enhancer sequences ahead, within, and behind the gene bind proteins that in turn bind to the RNA polymerase complex to stimulate the rate of transcription. After the RNA chain is made it is modified before being exported from the nucleus. Note that the known constant sequence at the beginning and end of the intron segment is not large enough to actually specify its position.

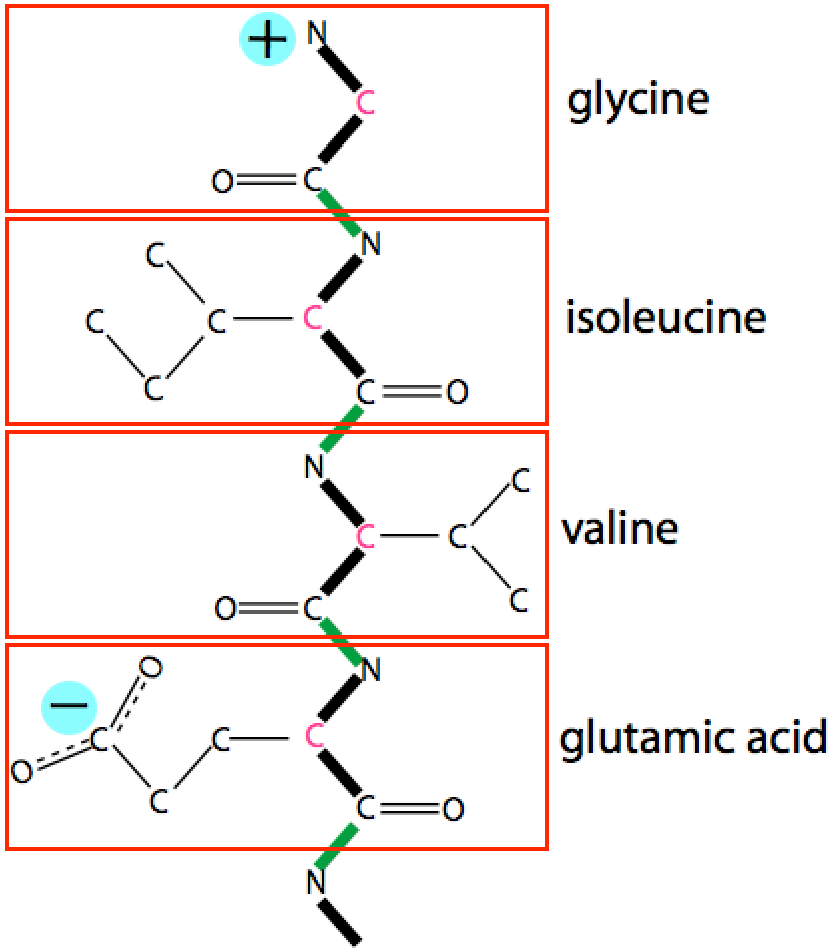
Protein

Much of the structure and most of the chemical machinery in the living cell is made of protein. Thus, most of the RNA is an intermediate messenger RNA (m-RNA), carrying information that will be translated into protein. However, some RNA species represent a final product. Most of these RNA species are part of the machinery that either processes messenger RNA or translates the information into protein.

Proteins are polymers of amino acids. There are 20 common amino acids, and each has the same three atom backbone when linked together to form a protein: an amino nitrogen attached to an alpha carbon attached to a carboxyl carbon. The protein polymer is thus formed by linking the carboxyl carbon of one amino acid to the amino nitrogen of the next amino acid. This carbon-nitrogen bond is called a peptide bond, and proteins (especially short ones) are sometimes called polypeptides. The first four amino acids in one of the polypeptide chains of the hormone insulin are shown in Figure peptideChain.

Figure peptideChain

Start of a protein chain

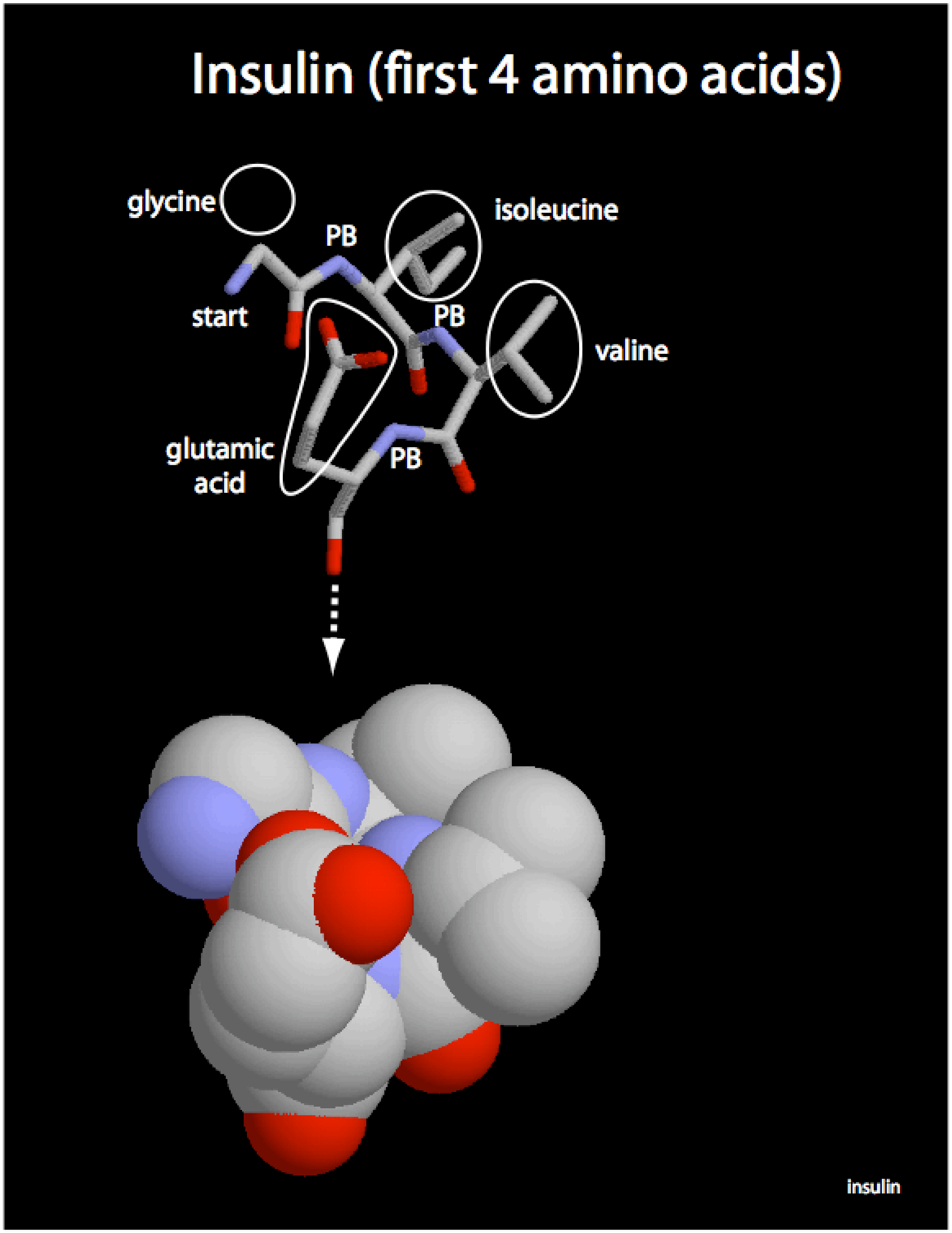


peptideChain

Figure peptideChain. A simplified stick model of the beginning of the insulin molecule reveals fundamental features of all proteins. The amino acids (identified by the red rectangles) that make up proteins are linked together by peptide bonds (green). Each amino acid contains an alpha carbon (red C) which is linked to a chemical group that makes the amino acid different from the others.

As with the large biomolecules we have described previously, the atoms in insulin do not actually lie on a flat plane but have a specific three dimensional structure, and they occupy enough space to make an essentially solid object. The stick model in the next Figure represents the three dimensional locations of the atoms, while the space filling model in the same Figure is the most realistic model of the structure. Unfortunately it is almost impossible to follow the polypeptide chain in the last model; we gain realism at the expense of clarity.

Figure insulin



Insulin (first 4 amino acids)

Figure insulin. The three dimensional structure of the first four amino acids of insulin. The stick model at the top shows how the amino acids are arranged in space. The characteristic side groups for each amino acid are circled in white, and the peptide bonds (purple indicates nitrogen) are labeled "PB". The bottom space filling model shows that the amino acids form a compact structure, which in fact prevents us from seeing much interior detail in this view.

While the 20 common amino acids have the same backbone, they come in different sizes and chemical properties because they have different groups of atoms attached to the alpha carbon. The amino acid glycine is one extreme, with only two small hydrogen atoms attached to the alpha carbon. The amino acids lysine and arganine have several carbons and positively charged basic amino groups attached to the alpha carbon. The amino acids glutamic and aspartic acid have several atoms and a negatively charged acidic group attached to the alpha carbon. The amino acid tyrptophan has several carbon rings attached to the alpha carbon. The amino acid leucine has a long uncharged carbon chain attached to the alpha carbon.

Thus, while DNA and RNA were each composed of a sequence of four nucleotides, with two sizes and four patterns of partial charges which enabled base pairing, proteins are made up of amino acids that have very different sizes, charges, and hydrophobic properties. The very different properties of the 20 amino acids result in amino acid polymers folding up into a huge variety of specific three dimensional shapes. The proteins represent the result of translation of digital to analog information[5].

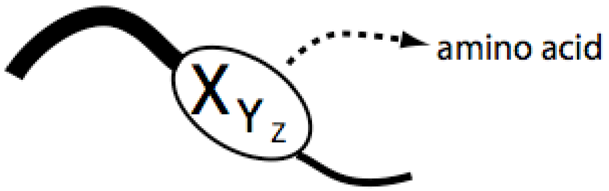
The nucleotide to amino acid code

Each amino acid code word is three nucleotides long. There are thus 4 x 4 x 4 = 64 possible triplets, with most amino acids are represented by several triplets. Three of the 64 triplets do not code for amino acids, but rather signal termination of the protein chain. The third nucleotide in each code word is the least significant in specifying the amino acid, and for some amino acids it is effectively ignored. For most other amino acids others it is only significant for the third nucleotide to be either a purine (A or G) or a pyrimidine (U or C).

[5] Since proteins are a sequence of amino acids they are digital in a literal sense. However, unlike DNA, the function of the protein is defined by its three dimensional shape and chemical properties, not the sequence of amino acids per se.

Figure aaCode

The nucleotide (RNA) to amino acid (protein) code



	U	C	A	G	
U	phenylalanine	serine	tyrosine	cysteine	U
	phenylalanine	serine	tyrosine	cysteine	C
	leucine	serine	STOP	STOP	A
	leucine	serine	STOP	tryptophan	G
C	leucine	proline	histidine	arganine	U
	leucine	proline	histidine	arganine	C
	leucine	proline	glutamine	arganine	A
	leucine	proline	glutamine	arganine	G
A	isoleucine	threonine	asparagine	serine	U
	isoleucine	threonine	asparagine	serine	C
	isoleucine	threonine	lysine	arganine	A
	methione	threonine	lysine	arganine	G
G	valine	alanine	aspartate	glysine	U
	valine	alanine	aspartate	glysine	C
	valine	alanine	glutamate	glysine	A
	valine	alanine	glutamate	glysine	G

🟥	negative charge
🟦	positive charge
⬛	hydrophilic
🟧	hydrophobic

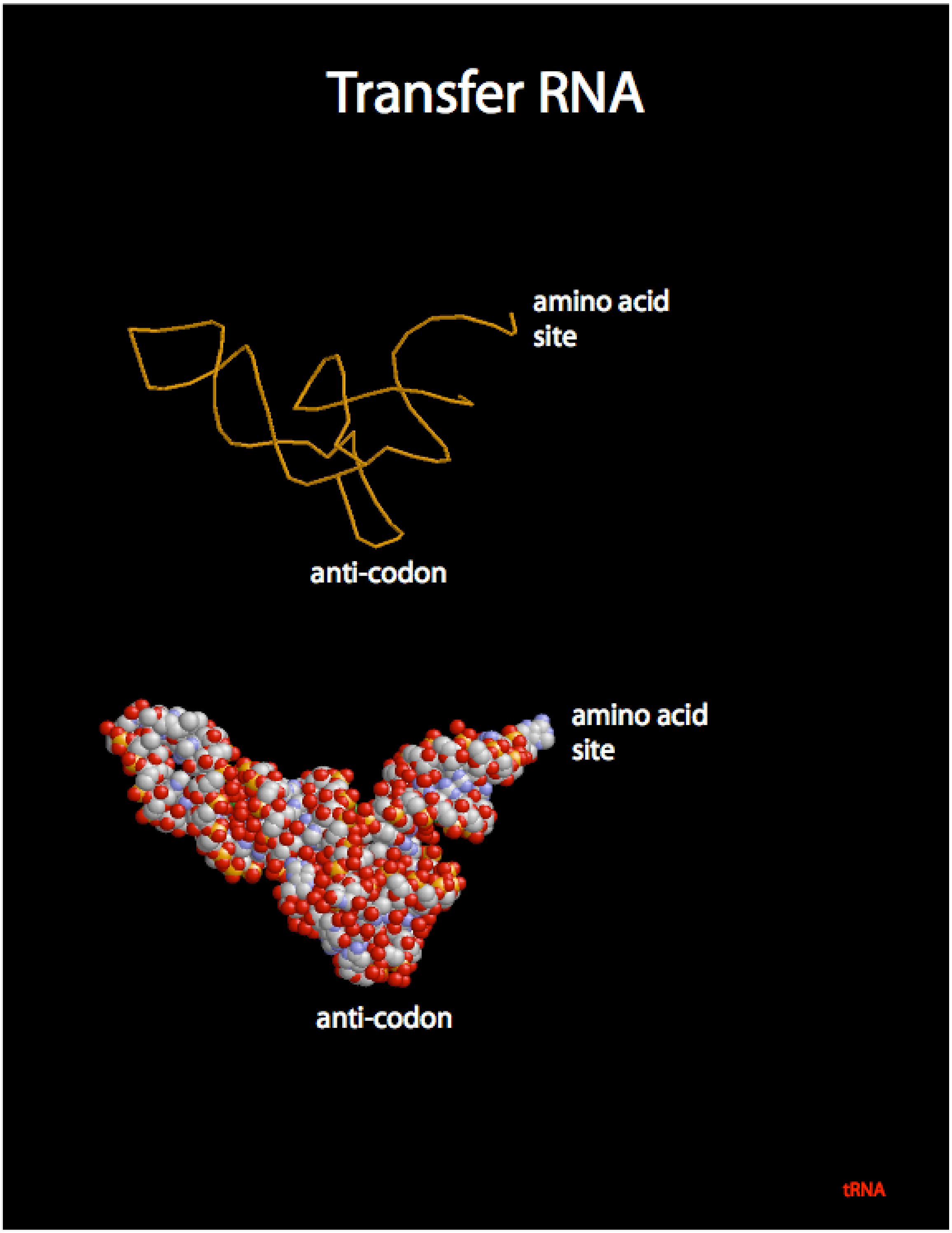
aaCode

Figure aaCode. The first nucleotide is indicated by the largest sized letters (along the left side of the table). The second nucleotide is middle sized (along the top of the table). The third nucleotide is represented by the four entries in each table cell. The 20 amino acids have very different physio-chemical properties, as indicated by the four colors.

Using the code: translating sequence to shape

Implementing the code requires converting a nucleotide sequence to an amino acid sequence. This is done using a collection of "adaptors", small RNA molecules (transfer or t-RNA) that contain a three nucleotide sequence (the anti–codon) complementary to a triplet code word in the messenger RNA. The specific shape of each species of t-RNA is due to nucleotide sequences in the RNA that are complementary to other sequences in the same chain. The base pairing of these regions result in small double stranded helical segments and specific loops in each t-RNA species which result in a fairly rigid and characteristic shape. The differences in shape between each species of t–RNA are further amplified by enzymes that modify the structure of specific nucleotides by adding methyl or other groups of atoms to the t–RNA. Although each t–RNA has a specific shape, as a group they have a common profile: a chain of about 75 nucleotides that coil up to form an L shape with the code triplet exposed at the bend in the L and the amino acid linked to one end.

Figure tRNA



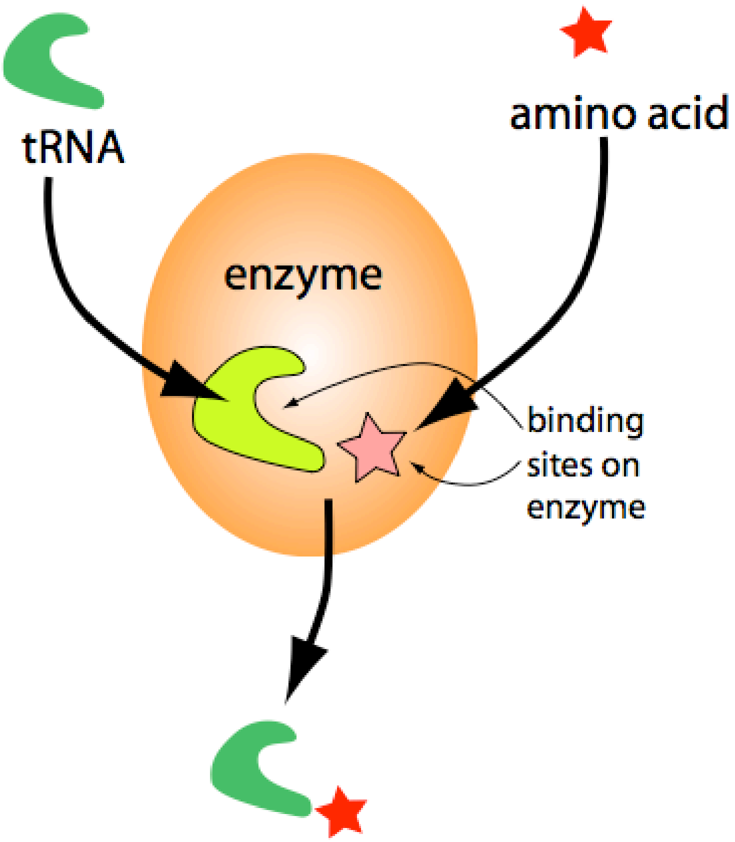
Transfer RNA

Figure tRNA. There are about 50 different t-RNA molecules in a cell, because some can bind to more than one of the 61 three nucleotide codon. All the t-RNA species have two short double stranded regions which form an L-shape, with the anti-codon at the bend and the amino acid at one end.

However, the specific shape of a t-RNA is not recognized by the amino acid itself, but rather by a group of much larger enzymes, the aminoacyl-t–RNA synthetases, which specifically bind both the amino acid and t–RNA and attaches the amino acid to the end of the t-RNA. The amino acid–t–RNA is then released and diffuses to bind to the proper triplet on the messenger RNA, and the amino acid is added to the growing protein chain.

Figure charge_tRNA

Linking an amino acid
to its tRNA



charge_tRNA

Figure charge_tRNA. The aminoacyl-t-RNA synthetase enzyme (orange egg) recognizes a specific amino acid and one or more of the t-RNA species that are specific for that amino acid. The t–RNA is recognized not by its anti-codon, but by its shape. Once the correct amino acid has be attached to a t-RNA with the correct codon, it is possible to translate the nucleotide code to the amino acid code.

As you know by now, rearrangements of chemical bonds in the cell require an enzyme, and the transfer of an amino acid from t-RNA to the end of a growing protein chain is no exception; in fact this task requires many enzymes and associated molecules that are part of a huge molecular machine. Perhaps the size of this structure is not surprising since two large molecules, the m-RNA being translated, and the growing protein chain, must be held in precise alignment and periodically moved relative to each other by precisely three nucleotides.

The ribosome is the translation machine. It is composed of two subunits which bind together at the start of translation of the m-RNA. Each of the ribosome subunits is made up of dozens of proteins and one or two RNA molecules. In addition to the ribosome, several other smaller enzymes diffuse into the reaction zone and carry out functions necessary for the addition of the amino acid. The process is illustrated in the Figure ribosome.

Figure ribosome

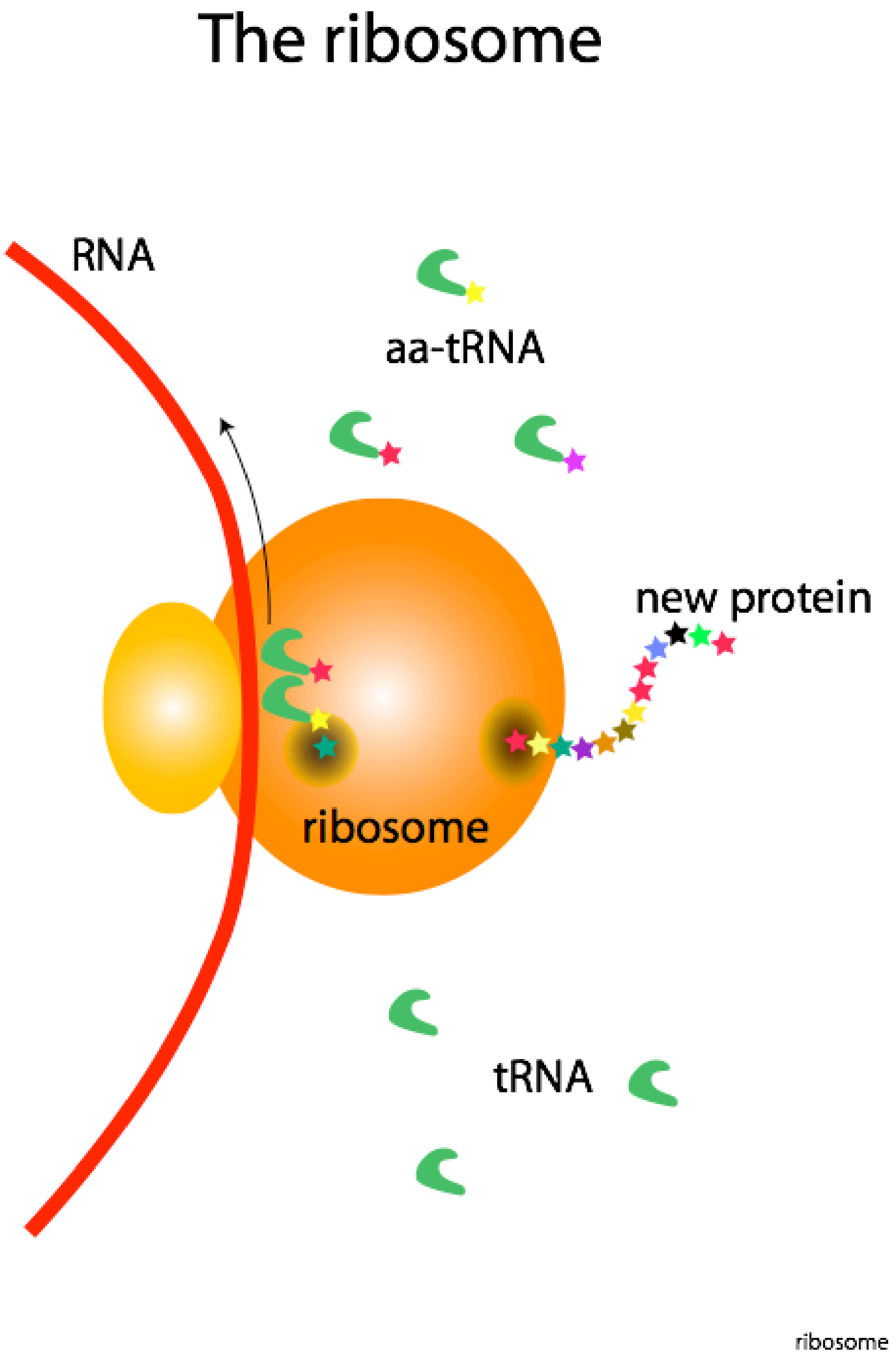


Figure ribosome. The ribosome is composed of a small (yellow egg) and large (orange egg) subunit which come together at the start of translation of the messenger RNA (red ribbon). The ribosome moves along the RNA as each t-RNA (green L shape) binds to successive three nucleotide codons and the new amino acids are added to the growing protein (multicolored stars). The new protein chain moves through a tunnel in the ribosome and then begins the process of folding into its characteristic three dimensional structure.

From the Figure you can see that the growing protein chain is not just flapping around. As it is made it passes through a tunnel in the ribosome, and is thus kept from interfering with the action occurring along the m-RNA. This tunnel thus plays a role similar to the tunnel in the RNA polymerase.

However, the ribosome is not just a scaffold. The actual formation of each peptide bond of the protein is catalyzed by a integral component of the ribosome; an RNA molecule (no relation to any of the t-RNAs). Thus while most enzymes are proteins, some of the most central reactions in the cell are carried out by RNA molecules. The important role of RNA molecules as enzymes was only gradually realized, but it has an important implication for the origin of life.

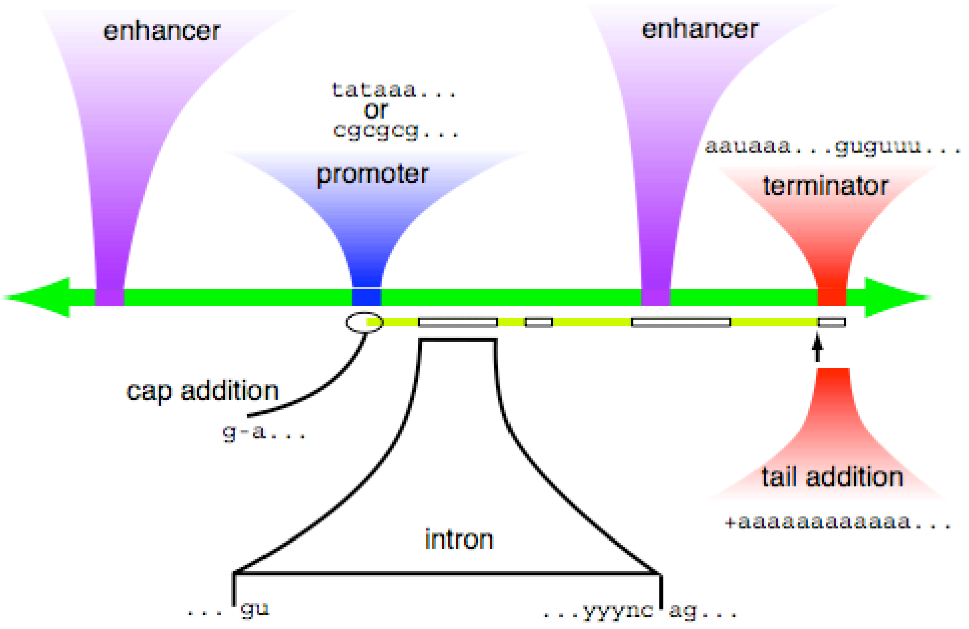
Summary of translation layer sites

All proteins begin with the amino acid methionine, coded by the nucleotide triplet AUG (this part of the protein is usually removed at a later stage). The nucleotides triplets are then translated into a protein until one of the three stop triplets is encountered.

Figure TranslationLayer

Transcription Layer Format

step 1. synthesis of pre-m RNA



step 2. modification to produce m RNA

TranscriptionLayer

Figure TranslationLayer. The first amino acid is always methionine, and at the beginning of a translated sequence its triplet, AUG, is preceded by ACC. Translation proceeds one triplet at a time until one of the three nonsense triplets is encountered.

Protein folding

At this stage of information transfer the processes are no longer digital in nature, although the information that defines and guides these processes is of course encoded in the digital nucleotide sequence of the DNA .

Specific folding of the amino acid chain into a compact protein molecule is required for the amino acid chain to carry out its function. Thus, folding is required for the information in DNA to be converted into function. Even before it was known that DNA contained the genetic information it was understood that the three dimensional structure of proteins was essential to their correct function. When it was shown that folding could be destroyed by heat or treatment with urea, and the polypeptide chain would spontaneously refold into the original structure when conditions were restored to physiological, it was clear that information for folding was contained in the amino acid sequence itself.

However, although we have known for more than 50 years that protein folding is central to a complete understanding of information transfer in life, we still don't have a complete or even robust theory. That is certainly not to say that nothing is known about folding, just that a scientist may not be able to make a very good prediction for the structure of a protein when given an unusual amino acid sequence (as of 2005). I use the term unusual because three major advances in the study of proteins over the years enable fairly good guesses for the structures of most proteins.

First we now know the three dimensional structures of about 10,000 proteins. Many of these structures were painstakingly obtained by analysis of x-ray diffraction of crystals of the proteins. The structures of proteins of modest size can now be obtained more easily by NMR.

Second, several common motifs, 10-100 amino acids long, appear repeatedly in different proteins. The first to be discovered was the alpha-helix, a tight single stranded coil held together by hydrogen bonds between peptide groups on one loop and carbonyl oxygens on the loop below (or above, depending on how you visualize the helix). Another common motif is the beta-sheet, peptide loops linked by similar hydrogen bonds, which form a planar sheet with a slight twist. Some proteins are mostly segments of alpha-helixes while other proteins have no alpha helix component, and the same is true of beta-sheets. It is sometimes possible to predict that a amino acid sequence will form an alpha-helix in a protein, but sometimes the prediction fails; the same is true of beta-sheets.

Third, the amino acid sequences of hundreds of thousands of proteins have been deduced from DNA sequences. These advances have revealed that almost all proteins can be grouped into related families. As discussed in the next chapter this grouping is not surprising since essentially all living organisms and all genes, are related to each other because they evolved from common ancestors. The data base of known structures and the existence of protein families means that any "new" amino acid sequence is likely to be fairly closely related to a sequence with a known structure. This homology plus our modest knowledge of the fundamental chemical basis of protein structure usually enables a fairly good prediction for the structure of the "new" protein.

Thus the study of protein folding is a work in progress. A fundamental problem is that the ensemble of possible foldings for a peptide chain of even modest size is so immense that it is impossible to find the most stable structure (the one with the strongest internal bonding energy) by a brute force search of all possible structures. In fact, in some cases the correct protein structure may be the result of kinetics, e.g. the fact that it is sequentially elongated from amino to carboxyl end. The correct structure may also sometimes be dependent on the molecular environment of cellular synthesis, e.g. the fact that the most recently segment of the chain is protected by the ribosome.

Protein transport

Many proteins must be transported to a specific area of the cell in order to function. This address is encoded in the amino acid sequence at the N-terminal (beginning) end of the protein, however the code is only partially understood, and the amino acid sequence is interpreted in the context of the structure of the protein. There are not a large number of possible destinations, and thus the transport coding scheme doesn't need to contain a large number of bits of information.

Protein glycosylation

Some proteins contain groups of sugar molecules attached to the surface of the molecule. Proteins on the surface of cells, or proteins that are secreted are typically glycosylated. Sugars can represent more than 20 percent of the protein by weight, and form complex branched chains of several types of sugars attached to specific amino acids. Sugars can have a big effect on the solubility and stability of the protein. In addition they often are the major antigenic determinate, which means they define the protein as "self" or "foreign" for the immune system. Blood groups are defined by glycosylation, and transfusion of mismatched blood can be fatal due to destruction by the recipients immune system. Glycosylation can only occur at a few specific amino acid sequences, but the three dimensional structure and location also determine the glycosylation pattern.

Figure ProteinXLayer

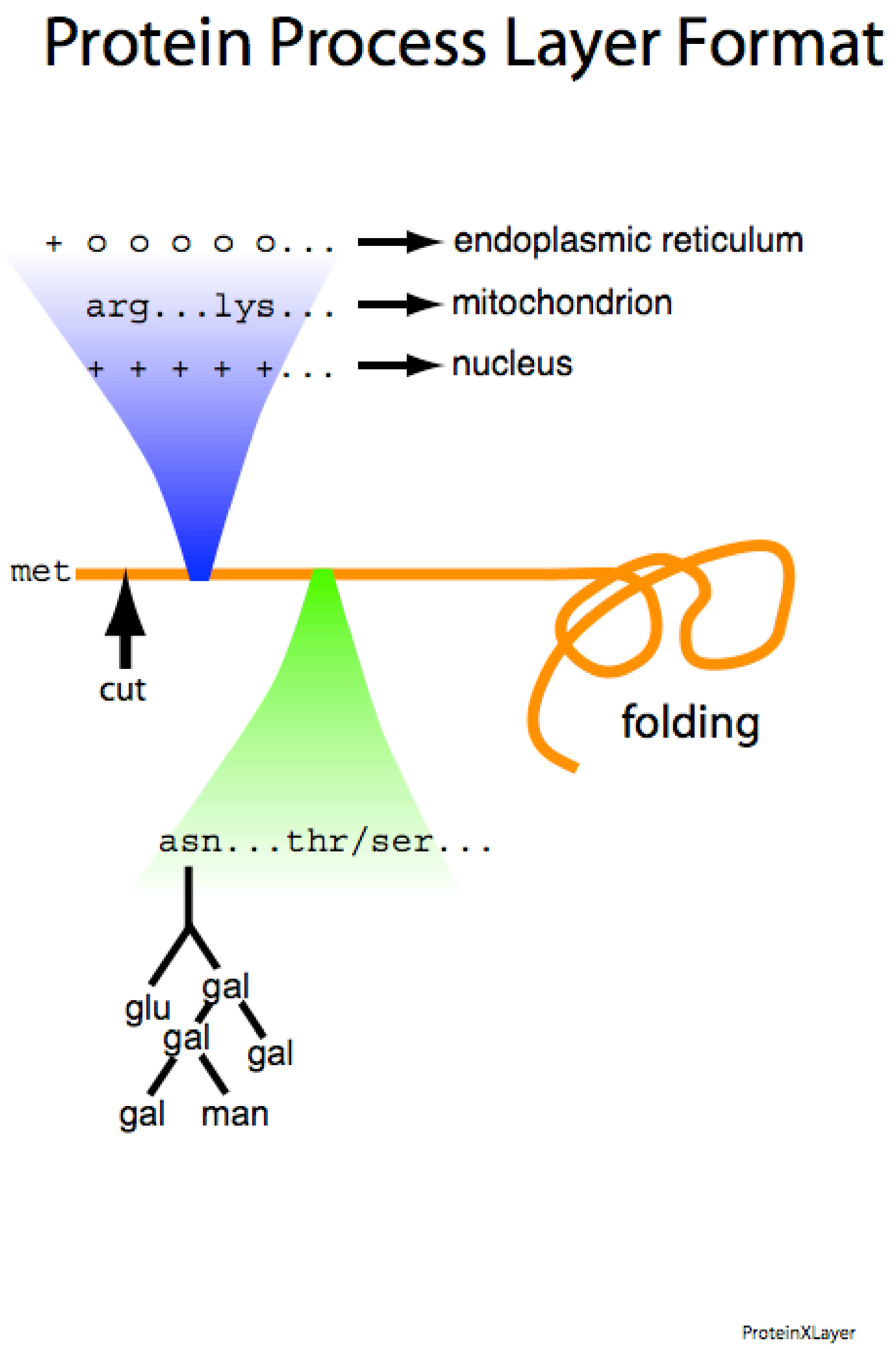


Figure ProteinXlayer. Three major types of modifications to proteins which occur after translation are illustrated here. While the first amino acid of all proteins is methionine, proteases typically remove it and several subsequent amino acids. The specificity of this trimming appears to be mostly determined by the three dimensional structure of the protein. The amino acids at the beginning of the protein sometimes determine where the molecule is transported. A positively charged amino acid followed by neutral amino acids is the signal for transport to the endoplasmic reticulum, a series of arginines and lysines guide the protein to the mitochondrion, while a series of positively charged amino acids results in the protein being transported to the nucleus of the cell. Sugars are added to some proteins. The three dimensional shape of the protein and its location in the cell determine the structure of the sugars.

Chapter summary

A DNA strand is copied by making a second DNA strand that has a base sequence that is complementary to it. The new chain is made by sequentially adding nucleotides, with the formation of each new chemical bond catalyzed by the enzyme DNA polymerase.

Each new nucleotide precursor is actually a nucleotide triphosphate, and when it is added to the growing DNA chain a pyrophosphate group is released; a new chemical bond is made as old chemical bonds are broken. The net result is a decrease in energy; all chemical reactions must result in a decrease in energy. The great majority of chemical reactions that occur in the cell require an enzyme to form a transient intermediate during the reaction. Most enzymes are proteins, but some are RNA.

The DNA molecules in living organisms consist of two complementary strands twisted into a helix. During replication both strands are copied at a replication fork which contains many other enzymes in addition to the DNA polymerase. Thus, as the fork moves along, two new double stranded helixes are created.

Replication of DNA is not always perfect, and spontaneous chemical reactions occur which damage DNA. However, there are a series of enzymes that correct most of the changes that would represent corruption of the information carried by the DNA.

In order to use the information in DNA, a complementary chain of RNA is made by the enzyme DNA dependent RNA polymerase. Some RNA molecules are part of the chemical machinery of the cell, while other RNA chains are copied into protein chains.

Proteins are polymers of 20 different amino acids, and are thus very different from DNA and RNA. The conversion of the genetic information from DNA to protein is a translation of one language into another. The protein chains fold to make a specific three dimensional structure characteristic of each protein species, and this structure is essential for the function of the protein. Thus, translation is a conversion of digital to analogue information.