CHAPTER TWENTY

Molecular Evolution

Getting Started—Formation of the Earth
The Early Atmosphere
Oparin's Theory of the Origin of Life
The Miller Experiment
Polymerization of Monomers to Give Macromolecules
Enzyme Activities of Random Proteinoids
Origin of Informational Macromolecules
Ribozymes and the RNA World
The First Cells
The Autotrophic Theory of the Origin of Metabolism
Evolution of DNA, RNA and Protein Sequences
Creating New Genes by Duplication
Paralogous and Orthologous Sequences
Creating New Genes by Shuffling
Different Proteins Evolve at Very Different Rates
Molecular Clocks to Track Evolution
Ribosomal RNA—A Slowly Ticking Clock
The Archaebacteria versus the Eubacteria
DNA Sequencing and Biological Classification
Mitochondrial DNA—A Rapidly Ticking Clock
The African Eve Hypothesis
Ancient DNA from Extinct Animals
Evolving Sideways: Horizontal Gene Transfer
Problems in Estimating Horizontal Gene Transfer

Getting Started—Formation of the Earth

The big bang is estimated to have happened about 20,000,000,000 years ago. About 15,000,000,000 years later, a cloud of interstellar dust and gas coalesced and condensed due to gravity into a large ball of gas that we call the sun, orbited by smaller spherical bodies of variable composition, called the planets. The universe consists mostly of the light molecular weight gases hydrogen and helium, which account for most of the material of the stars. The heavier elements together comprise only about 0.1 percent of the total and form the planets (Table 20.01).

As the earth formed, heat was released by the collapse due to gravity and also by the radioactivity of elements present in the original dust. During its first few hundred million years the Earth was too hot for water to liquefy and so H_2O was only present as steam. Later, as the Earth cooled off, the steam condensed to form oceans and lakes. Life is thought to have originated by means of chemical reactions occurring in the atmosphere followed by further reactions in the primeval oceans and lakes (the hydrosphere).

The Early Atmosphere

Life evolved under a reducing atmosphere lacking any free oxygen.

The Earth's original atmosphere, the **primary atmosphere**, consisted mostly of hydrogen and helium, but the Earth is too small a planet to hold such light gases and they floated away into space. The Earth then accumulated a **secondary atmosphere**, mostly by volcanic out-gassing. Volcanic activity was much greater on the hotter primitive Earth. Volcanic gas consists mostly of steam (95%) and variable amounts of CO_2, N_2, SO_2, H_2S, HCl, B_2O_3, elemental sulfur, and smaller quantities of H_2, CH_4, SO_3, NH_3 and HF but no O_2. Of all these, the concentration of CO_2 was in the second highest amount (about 4%). In addition, water vapor reacted with primeval minerals such as nitrides to give ammonia, with carbides to give methane and with sulfides to give hydrogen sulfide. There was no free oxygen.

Our present atmosphere, the **tertiary atmosphere**, is of biological origin. The methane, ammonia, and other reduced gases have been consumed and the inert components (nitrogen, traces of argon, xenon etc.) have remained largely unchanged. Substantial amounts of oxygen have been produced by photosynthesis. This could not

TABLE 20.01 Elemental Compositions in Atoms per 100,000

Element	Universe	Earth	Crust	Life
H	92,700	120	2,900	60,600
He	7,200	<0.1	<0.1	0
O	50	48,900	60,400	26,700
Ne	20	<0.1	<0.1	0
N	15	0.3	7	2,400
C	8	99	55	10,700
Si	2.3	14,000	20,500	<1
Mg	2.1	12,500	1,800	11
Fe	1.4	18,900	1,900	<1

primary atmosphere The original atmosphere of the earth consisting mostly of hydrogen and helium
secondary atmosphere The atmosphere of the earth after the light gases were lost and resulting mostly from volcanic out-gassing. It contained reduced gases but no oxygen
tertiary atmosphere The present atmosphere of the earth resulting from biological activity

TABLE 20.02 Approximate Evolutionary Time Scale

Millions of Years Ago	Major Events
20,000	Big Bang
5,000	Origin of planets and sun
3,500	Origin of life
3,000	Primitive bacteria start using solar energy
2,500	Advanced photosynthesis releases oxygen
1,500	First eukaryotic cells
1,000	Multicellular organisms
600	First skeletons give nice fossils
1.8	First true humans—*Homo erectus*
0.15	African Eve gives birth to modern Man
0.01	Domestication of crops and animals begins
0.002	Roman Empire
0.0001	Darwin's theory of evolution
0.00005	Double helix discovered by Watson and Crick
0.000003	Human genome sequenced

Oxygen in today's atmosphere is the result of photosynthesis.

occur until the cyanobacteria, the first true photosynthetic organisms, had evolved about 2.5 thousand million years ago (see Table 20.02 for timescale). As more and more photosynthetic organisms evolved, the oxygen content of the atmosphere increased. The oxygen content of the atmosphere reached 1 percent about 800 million years ago and 10 percent about 400 million years ago. Today it is about 20 percent.

Evidence for the increase in O_2 content of the atmosphere comes partly from the finding that rocks of different ages are oxidized to different extents. Thus rocks of age 1,800–2,500 million years sometimes contain UO_2, FeS, ZnS and PbS and FeO, all of which are unstable in the presence of even small amounts of gaseous O_2. Later rocks contain mostly Fe^{3+} rather than Fe^{2+}; and more oxidized ores of U, Zn and Pb.

Oparin's Theory of the Origin of Life

Reactions in the primeval atmosphere led to the accumulation of organic material in oceans and lakes—the primitive soup.

Ultraviolet radiation from the sun, together with lightning discharges, caused the gases in the primeval atmosphere to react, forming simple organic compounds. These dissolved in the primeval oceans and continued to react, forming what is sometimes referred to as the "**primitive soup**". The primitive soup contained amino acids, sugars, and nucleic acid bases among other randomly synthesized molecules (Fig. 20.01). Further reactions formed polymers and these associated, eventually forming globules. Ultimately, these evolved into the first primitive cells. This theory of the origin of life was put forward by the Russian biochemist Alexander Oparin in the 1920s. Charles Darwin himself had actually proposed that life might have started in a warm little pond provided with ammonia and other necessary chemicals. However, it was Oparin who outlined all the necessary steps and realized the critical point: life evolved before there was any oxygen in the air. Oxygen is highly reactive and would have reacted with the organic precursor molecules formed in the atmosphere, oxidizing them back to water and carbon dioxide.

primitive soup Mixture of random molecules, including amino acids, sugars, and nucleic acid bases, found in solution on the primeval earth

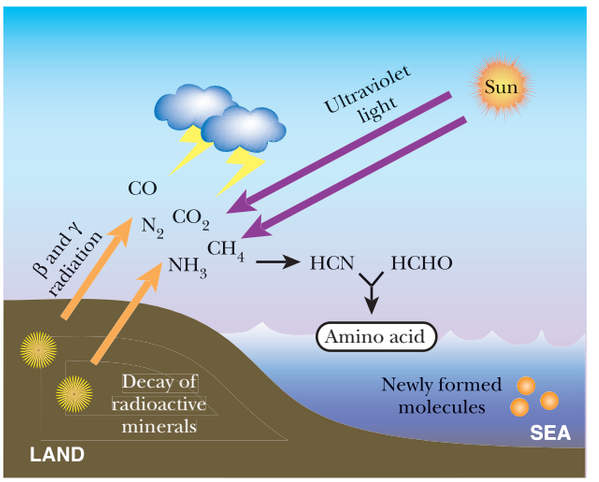


FIGURE 20.01 ***Formation of Primitive Soup***

According to Oparin's theory of the origin of life, conditions on planet earth were sufficient for forming early biological molecules. The atmosphere at this point contained CO, CO_2, CH_4, N_2, and NH_3. When energy was supplied by electrical discharge from lightning, ultraviolet radiation from the sun, and/or β and γ radiation from the earth, early organic compounds such as HCN and HCHO would form. These compounds would combine in the vapor phase and in the water to form amino acids. Dissolving in water protects the precursor molecules from being degraded again by the energy sources that triggered their formation.

The Miller Experiment

In the early 1950s, the biochemist Stanley Miller mimicked the reactions proposed to occur in the primitive atmosphere. An imitation atmosphere containing methane, ammonia and water vapor was subjected to a high voltage discharge (to simulate lightning) or to ultraviolet light (Fig. 20.02). The gases were circulated around the apparatus so that any organic compounds formed in the artificial atmosphere could dissolve in a flask of water, intended to represent the primeval ocean. These compounds could continue to react with each other in the water.

Experiments to mimic reactions in the primeval atmosphere have generated many of the metabolites and monomers found in modern cells.

There are many variants of this experiment (different gas mixtures, different energy sources, etc.). As long as oxygen is excluded, the results are similar. About 10 to 20 percent of the gas mixture is converted to soluble organic molecules and significantly more is converted to a non-analyzable organic tar. First, aldehydes and cyanides are formed, and then a large variety of other organic compounds (Table 20.03). Most of the naturally occurring amino acids, hydroxy-acids, purines, pyrimidines, and sugars have been produced in variants of the Miller experiment. These experiments produce all different isomers of the compounds; only a portion is biologically significant. For example, sarcosine and beta-alanine are both isomers of alanine and all three are generated in the Miller experiment. Many organic molecules, in particular sugars and amino acids, exist as two possible optical isomers, only one of which is normally found in biological macromolecules. In such cases the molecules formed in the Miller experiment consist of an equal mixture of the D- and L-isomers.

Since many chemical reactions are reversible, the same energy sources that produce organic molecules are also very effective at destroying them. The long-term build-up of organic material requires its protection from the energy sources that created it. This is the function of the imitation primeval ocean in the Miller experiment. Water shields molecules from ultraviolet radiation and from electric discharges. The survival of organic molecules on the primitive earth would have depended on

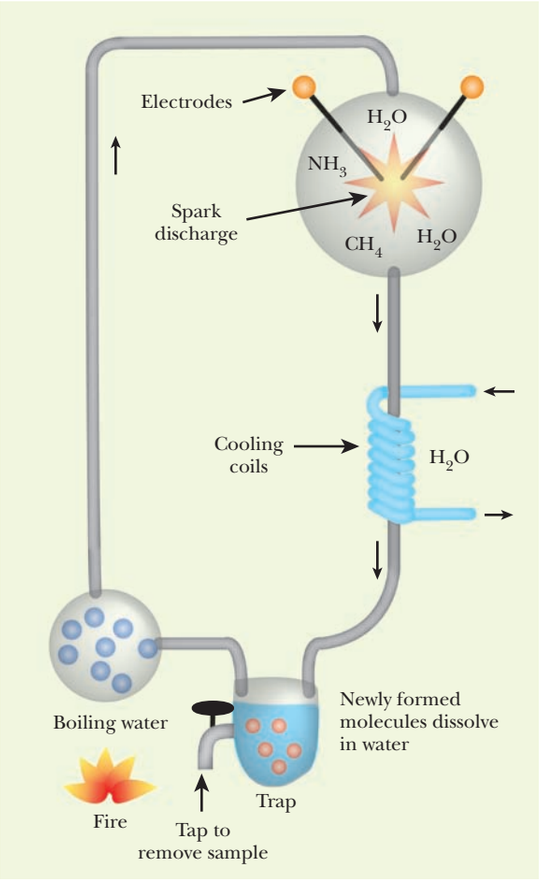


FIGURE 20.02 ***Miller's Experiment***

Miller's experiment used a closed system to simulate the primeval conditions on early planet earth. Water was boiled to make steam (lower left), which then mixed with NH_3 and CH_4 in another chamber. This upper chamber simulates earth's early atmosphere and was subjected to either electrical discharge (shown) or ultraviolet radiation (not shown). The resulting products were cooled and condensed as they passed coils filled with cold water. The condensed products dissolved in a flask of water, which simulated the early oceans. The newly formed molecules were analyzed at various times during the experiment.

TABLE 20.03 Typical Products from Miller's Experiment

Molecule	Name	Relative Yield
H-COOH	formic acid	1000
H_2N-CH_2-COOH	glycine	275
HO-CH_2-COOH	glycolic acid	240
H_2N-CH(CH_3)-COOH	alanine	150
HO-CH(CH_3)-COOH	lactic acid	135
H_2N-CH_2CH_2-COOH	beta-alanine	65
CH_3-COOH	acetic acid	65
CH_3-CH_2-COOH	propionic acid	55
CH_3-NH-CH_2-COOH	sarcosine	20
HOOC-CH_2CH_2-COOH	succinic acid	17
H_2N-CO-NH_2	urea	9
HOOC-CH_2CH_2CH(NH_2)-COOH	glutamic acid	2.5
HOOC-CH_2CH(NH_2)-COOH	aspartic acid	1.7

Water protects organic molecules from destruction by UV radiation or lightning.

their escape from UV radiation and lightning either by dissolving in seas or lakes or by sticking to minerals. Most organic molecules formed too far up in the sky would have been destroyed again very quickly, while those that reached the sea and dissolved would have survived. Note that organic acids, in particular amino acids, are water soluble and non-volatile. Once they are safely dissolved in water, there is little ten-

dency for such molecules to return to the atmosphere. Their precursors, the aldehydes and cyanides, are not only highly reactive but also volatile. Consequently, these molecules do not survive for long. Thus, even at this early stage, there was a form of natural selection between molecules.

Originally it was thought that the primitive secondary atmosphere contained mostly NH_3 and CH_4. However, it seems more likely that most of the atmospheric carbon was CO_2 with perhaps some CO and the nitrogen mostly as N_2. The two reasons for this are: volcanic gas has more CO_2, CO and N_2 than CH_4 and NH_3, and that UV radiation destroys NH_3 and CH_4, therefore, these molecules would have been short-lived. The UV destruction of CH_4 occurs in two steps. First, UV light photolyses H_2O to $H^{\bullet}$ and $^{\bullet}OH$ radicals. These then attack methane, giving eventually CO_2 and H_2, which would be lost into space. In Miller's experiment, gas mixtures containing CO, CO_2, N_2, etc. give much the same products as those containing CH_4 and NH_3 so long as there is no O_2. Since CO, CO_2 and N_2 do not supply H atoms; these come mostly from water vapor photolysis. In fact, in order to generate aromatic amino acids under primitive earth conditions it is necessary to use less hydrogen-rich gaseous mixtures. Most of the natural amino acids, hydroxy-acids, purines, pyrimidines, and sugars have been produced in variants of the Miller experiment.

Polymerization of Monomers to Give Macromolecules

Polymerization of monomers to give biological macromolecules usually requires the removal of H_2O. Clearly, water is in excess in the oceans and removal of H_2O from dissolved molecules is therefore unfavorable. Consequently, the assembly of macromolecules such as proteins and nucleic acids needs energy to form the linkages and/or remove the water. Before the high-energy phosphates used in modern cells were available, some other form of energy was needed.

Formation of biological polymers requires removal of water.

Imitation protein polymers, containing randomly linked amino acids, are known as "**proteinoids**." They can be formed by heating dry amino acid mixtures at around 150°C for a few hours (Fig. 20.03A). Whereas biological proteins are bonded using only the α-NH_2 and α-COOH groups of amino acids these "primeval polypeptides" contain substantial numbers of bonds involving side chain residues. They contain up to 250 amino acids and can sometimes perform primitive enzymatic activities. Such dry heat could have occurred near volcanoes or when pools left behind by a changing coastline evaporated. Much of the early work on proteinoids was done by Sydney Fox who proposed their thermal origin. However, another way to randomly polymerize amino acids is by using clay minerals with special binding properties (Fig. 20.03B). Binding of small molecules to the surface of catalytic minerals can promote many reactions. For example, certain clays, such as Montmorillonite, will condense amino acids to form polypeptides up to 200 residues long.

Water can be removed by moderate heating, by certain clay minerals or by chemical condensing agents.

Polymerization of amino acids may have also occurred in solution, but another component, a condensing agent, is required to withdraw water. Several possible primeval condensing agents have been proposed, including reactive cyanide derivatives and, more biologically relevant, **polyphosphates**. Inorganic polyphosphates would have been present in primeval times (formed by volcanic heat from phosphates for example). Polyphosphates can react with many organic molecules to give organic phosphates. Amino acids give two possible products (Fig. 20.04). **Acyl phosphates** have the phosphate group attached to the carboxyl group of the amino acid ($NH_2CHRCO\text{-}OPO_3H_2$) and **phosphoramidates** have the phosphate attached to the amino group of the amino acid ($H_2O_3P\text{-}NH\text{-}CHR\text{-}COOH$). Gentle heating or irradiation such derivatives will give polypeptides. Modern life uses acyl phosphate derivatives during protein

acyl phosphate Phosphate derivative in which the phosphate is attached to a carboxyl group
phosphoramidate Phosphate derivative in which the phosphate group is attached to an amino group
polyphosphate Compound consisting of multiple phosphate groups linked by high energy phosphate bonds
proteinoid Artificially synthesized polypeptide containing randomly linked amino acids

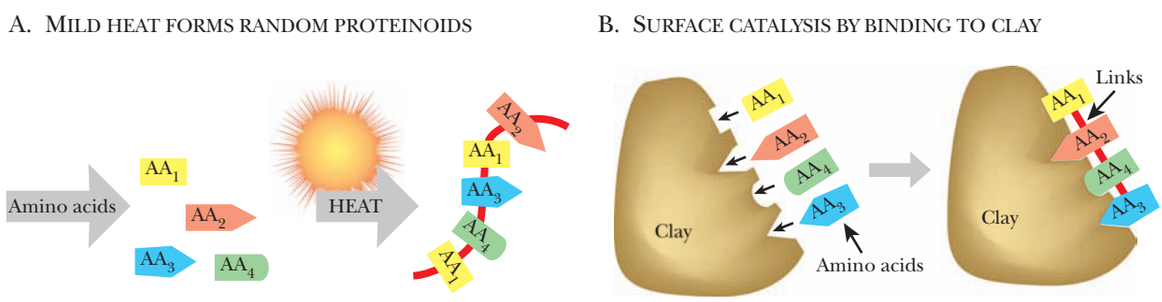


FIGURE 20.03 ***Formation of Proteinoids by Mild Heat or Clay Catalysis***

(A) A mixture of separate amino acids will form artificial polypeptide chains or "proteinoids" when subjected to heat in the absence of water for a few hours. (B) Amino acids can also form bonds when they bind to certain types of clay. The clay has binding sites for amino acids in close proximity, therefore, once bound the amino acids condense into a proteinoid.

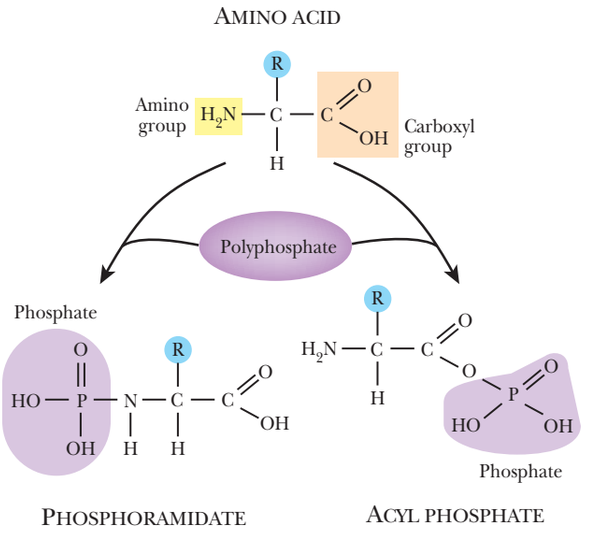


FIGURE 20.04 ***Formation of Acyl Phosphates and Phosphoramidates***

Condensing single amino acids into polypeptide chains could have occurred in solution as long as a condensing agent was present. One possible condensing agent was a polyphosphate that would react with the amino group or carboxyl group of individual amino acids. The two possible products, phosphoramidates and acyl phosphates, can form polypeptide chains by heating in solution.

synthesis, although from a chemical viewpoint, either derivative would work. In fact, laboratory synthesis of DNA does use phosphoramidates. Analogous reactions can produce AMP from adenine plus polyphosphate and polynucleotides can then form by polymerization (see below).

Enzyme Activities of Random Proteinoids

Artificial random proteinoids show inefficient but detectable enzyme activities.

Interestingly, random proteinoids stewed up in modern laboratories under fake primeval Earth conditions will carry out some simple enzyme reactions (Table 20.04). They are far slower and less accurate than enzymes made by real cells, but nonetheless they can perform recognizable enzymatic reactions. For example, random proteinoids can often remove carbon dioxide from molecules like pyruvate or oxaloacetate and split organic esters. About 50% of all modern enzymes contain metal ions as cofactors and the addition of metal ions greatly extends the enzyme activities of random proteinoids. The presence of traces of copper promotes reactions involving amino groups and iron mediates oxidation-reduction reactions. Incorporation of zinc allows the breakdown of ATP, which is used by modern cells, both as a precursor of nucleic acids and as an energy carrier. Most modern enzymes that process nucleic acids possess a zinc atom as cofactor.

TABLE 20.04 Enzyme Activities of Random Proteinoids

Reaction	Requirements	Substrate
esterase	histidine	p-nitrophenyl-phosphate
ATPase	Zn^{2+}	ATP
amination and deamination	Cu^{2+}	α-ketoglutarate ↔ glutamate
peroxidase and catalase	heme, basic proteinoids	H_2O_2 and H-donors e.g. hydroquinone, NADH
decarboxylation	basic proteinoids, or acidic proteinoids	oxaloacetate, pyruvate

Origin of Informational Macromolecules

Primeval synthesis of proteins or nucleic acids gives polymers with a mixture of natural and unnatural linkages.

Biological information is passed on by template-specific polymerization of nucleotides. A mixture of polyphosphate, purines and pyrimidines will produce random nucleic acid chains if ribose or deoxyribose is included. One problem, not yet solved, is that life uses 3′, 5′ linked nucleic acids whereas primeval syntheses give RNA molecules with a mixture of linkages, but mostly 2′, 5′. In contrast, deoxyribose has no 2′-OH and so cannot give 2′, 5′ links. However, it is generally thought that RNA probably provided the first informational molecule and that DNA is a later invention designed to store information in a more stable and accurate form.

Zn or Pb ions can catalyse non-enzymatic synthesis of RNA.

When an RNA template is incubated with a mixture of nucleotides, plus a primeval condensing agent, a complementary piece of RNA is synthesized. This non-enzymatic reaction is catalyzed by lead ions, with an error rate of about 1 wrong base in 10. With zinc ions, a great improvement is seen and lengths of up to 40 bases are produced, with an error rate of about 1 in 200. All modern day RNA and DNA polymerases contain zinc. If a 3′, 5′ linked RNA template is used about 75% of the newly formed RNA is 3′, 5′ linked. However, this does not surmount the problem that the original formation of random RNA type polymers favors the non-biological 2′, 5′ linkage very heavily.

The first strands of RNA to form act as templates for the assembly of later molecules.

If a mixture of nucleoside triphosphates (or nucleotides plus polyphosphate) is incubated under primeval conditions, using Zn as a catalyst, a molecule of single-stranded RNA with a random sequence will form. This original polymerization step is very slow. However, once an initial RNA polymer is present, it will act as a template for the assembly of a complementary strand. Template-directed synthesis is much more rapid, even in the absence of any enzyme. The complementary strand will in turn act as a template to generate more of the original RNA molecule. The net result is that once the first random sequence emerges, it will multiply rapidly and take over the incubation mixture (Fig. 20.05). The result will be a set of sequences with frequent mistakes but which are nonetheless clearly related (a molecular "**quasi-species**"). If a series of similar incubations are carried out, each individual sample will yield a "quasi-species" of related sequences. However, the sequence that takes over will be different for each incubation mixture.

Ribozymes and the RNA World

The ability of RNA to act as an enzyme as well as encode genetic information suggests that RNA came first.

Which came first the chicken or the egg?—Protein or nucleic acid? Since it is possible for random RNA molecules to be assembled and duplicate themselves under primeval conditions, it is generally held that nucleic acids appeared first. Furthermore, although most modern-day enzymes are indeed proteins, examples of RNA acting as enzyme and catalyzing reactions without help from proteins do exist. This scenario suggests

quasi-species A set of closely related sequences whose individual members vary from consensus by frequent errors or mutations

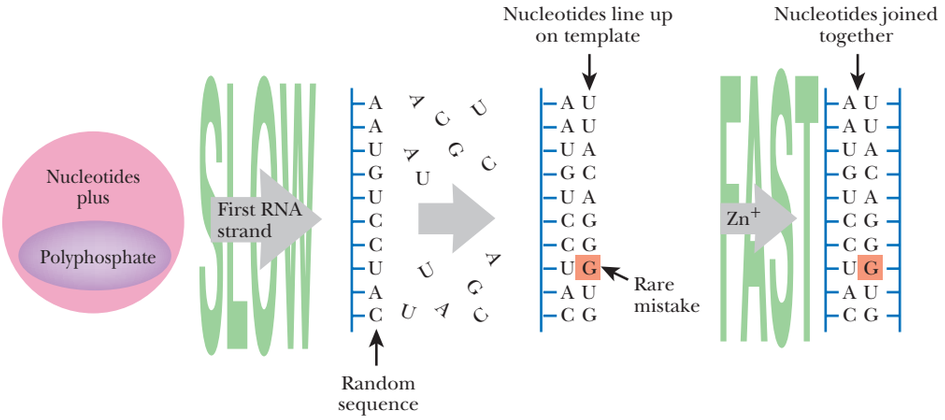


FIGURE 20.05 ***Assembly and Duplication of Random RNA***

A mixture of nucleosides and polyphosphates can form random stretches of RNA in the presence of zinc ions. The first strand of RNA forms very slowly. However, once the first strand is assembled, it may be used as a template to assemble complementary strands of RNA. Such template driven assembly of nucleotides is much faster than formation of the original random RNA strand. However, the non-enzymatic synthesis of RNA incorporates many wrongly paired bases.

that the primitive nucleic acid replicated alone and a protective protein coat was added later.

One rather extreme viewpoint is the idea that the earliest organisms had both genes and enzymes made of RNA and formed a so-called "**RNA world**." This idea was proposed by Walter Gilbert in 1986 and seeks to avoid the paradoxical problem that nucleic acids are needed to encode proteins, but that enzymes made of protein are needed to replicate nucleic acids. During the RNA world stage, RNA supposedly carried out both functions. Later, proteins infiltrated and took over the role of enzymes and DNA appeared to store the genetic information, leaving RNA as a mere intermediate between genes and enzymes.

Several examples illustrate the ability of RNA to perform enzymatic reactions as well as encode genetic information. These cases favor the primacy of RNA:

RNA molecules with enzyme activity include ribosomal RNA, ribonuclease P, self-splicing introns and viroids.

1. **Ribozymes** are single-use RNA molecules that are enzymatically active. A genuine enzyme processes large numbers of other molecules, and does not become altered during the reaction. Therefore, self-splicing RNA is not a true enzyme because it works only once. There is a growing list of known and suspected ribozymes. The most important is the ribosomal RNA of the large subunit, which is directly involved in the reactions of protein synthesis (see Ch. 8). One of the best-known ribozymes is **ribonuclease P**. This enzyme has both RNA and protein components and processes certain transfer RNA molecules. It is the RNA part of ribonuclease P that carries out the reaction. The protein serves only to hold together the ribozyme and the transfer RNA it operates on. In concentrated solution, the protein is not even necessary, and the RNA component will work on its own.
2. Self-splicing introns ("group I" introns) are an example of catalytic RNA. The genes of eukaryotic cells are often interrupted by non-coding regions (the introns), which must be removed from the messenger RNA before translation into protein. Normally, this is done by a spliceosome made up of several pro-

ribonuclease P A ribozyme found in many bacteria that processes certain transfer RNA molecules
ribozyme RNA molecule that is enzymatically active
RNA world The hypothetical stage of early life in which RNA encoded genetic information and carried out enzyme reactions without the need for either DNA or protein

teins and small RNA molecules. Occasionally, the intron RNA splices itself out without help from any protein. Such self-splicing is found in a few nuclear genes of some protozoans, in the mitochondria of fungal cells, and the chloroplasts of plant cells (see Ch. 12 for details).

3. Viroids are infectious RNA molecules that infect plants. As noted in Chapter 17, viroid RNA carries out a self-cleavage reaction during replication, i.e. viroids act as ribozymes.

Newly made nucleic acid molecules all start with a stretch of RNA.

4. DNA polymerase cannot initiate new strands but can only elongate pre-existing strands (see Ch. 5). Primers made of RNA must be used whenever new strands of DNA are started. RNA polymerase is capable both of initiation and elongation. This suggests that RNA polymerase may have evolved before DNA polymerase.
5. Small guide molecules of RNA are used in a variety of processes. These include the removal of introns, the modification and editing of messenger RNA (see Ch. 12 for details) and the extension of the ends of eukaryotic chromosomes by telomerase.
6. Riboswitches are recently discovered binding motifs in RNA that directly bind small molecules and so control gene expression in the absence of regulatory proteins (see Ch. 11 for details).

In a way, the critical question is whether RNA can copy itself without the involvement of DNA or help from protein enzymes. Although no RNA polymerases that are ribozymes still exist, it has proven possible to generate them artificially. Altered RNA molecules can be selected by a form of Darwinian evolution at the molecular level. Pre-existing ribozymes can be used as starting materials. Alternately, some experimenters have used random pools of artificially generated RNA sequences. In one experiment, RNA molecules showing primitive RNA ligase activity were selected from a pool of random RNA sequences. Such artificial ribozymes can link together two chains of RNA by a typical ligase reaction just like protein enzymes in modern cells. The best of these RNA ligase ribozymes was then subjected to further rounds of mutation and selection. The result was a ribozyme of 189 bases that uses an RNA template to synthesize a complementary strand of RNA with about 96–99% accuracy. This ribozyme adds single nucleotides, one at a time, to an RNA primer using nucleoside triphosphates as substrates (Fig. 20.06). However, it is very slow and can only extend chains by around 14 nucleotides because it is not "processive". In other words, the ribozyme dissociates from the template after adding each nucleotide, whereas true polymerases remain attached and proceed along the template adding nucleotides in quick succession.

Artificial ribozymes can be isolated by screening pools of random RNA sequences for particular enzymatic reactions.

One problem with the "RNA world" concept is that RNA is more reactive than DNA. Although RNA would form more easily than DNA under primeval conditions, it would also be less stable. Thus DNA, though slower to form initially, might tend to accumulate under such conditions. Moreover, the primeval soup would contain a mixture of the sub-components of both types of nucleic acid as well as proteins, lipids and carbohydrates. So it seems perhaps more likely that an ill-defined mixture, perhaps even hybrid nucleic acid molecules with both RNA and DNA components, emerged first.

The First Cells

Forming biologically relevant molecules in the primitive earth would have been the first step on the road to forming the first primitive cells. Possibly random proteins and greasy lipid molecules collected around the primeval RNA (or DNA), so forming a microscopic membrane-covered organic blob. Eventually this proto-cell learned how to use RNA to code for its protein sequences. The lipids formed a membrane around the outside to keep the other components together. Early on, protein and RNA

The first cells probably used RNA in multiple roles some of which were later taken over by proteins or DNA.

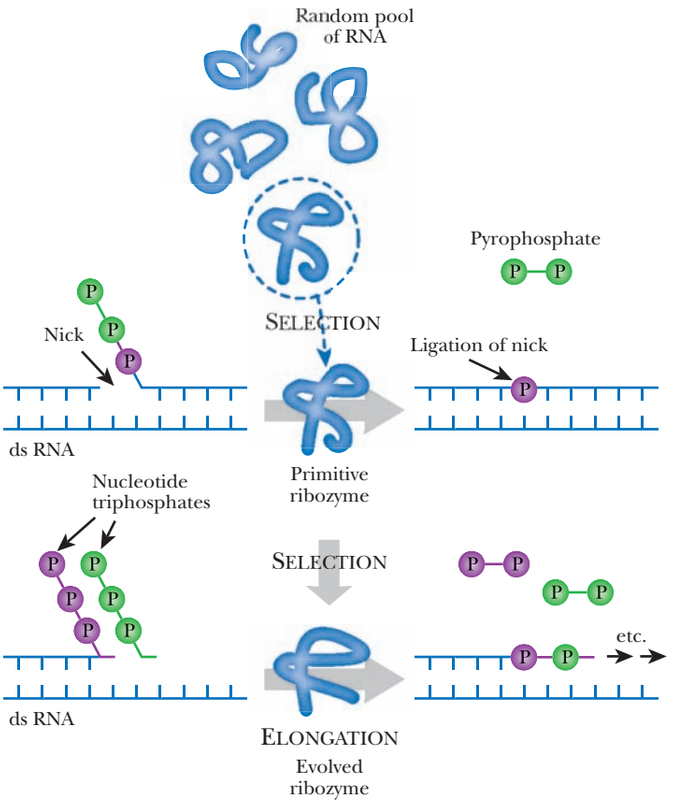


FIGURE 20.06 ***Artificial Evolution of Ribozyme RNA Polymerase***

A random pool of RNA was generated and screened for the ability to seal a nick in one strand of a broken piece of double-stranded RNA. Occasional random molecules of RNA that had this enzymatic activity were found and isolated. Through successive rounds of mutation and selection, this primitive ribozyme was altered just enough to actually catalyze the elongation of the RNA using a single-stranded template with a primer. Unlike true protein RNA polymerases, the RNA polymerizing ribozyme only added one nucleotide at a time, dissociating after each addition.

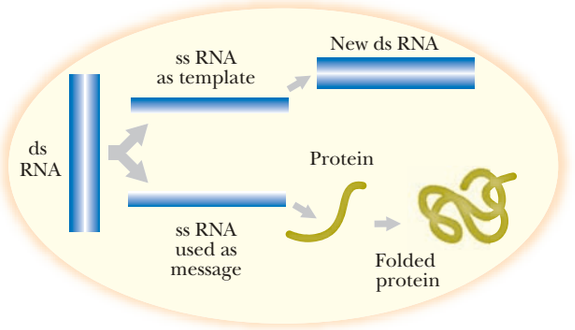


FIGURE 20.07 ***Emergence of the Proto-Cell***

An early primitive cell may have contained RNA molecules as the information coding material. RNA molecules would have acted as ribozymes to make copies of the RNA genetic material. Over time, the RNA would have evolved the ability to synthesize proteins to do much of the enzymatic work.

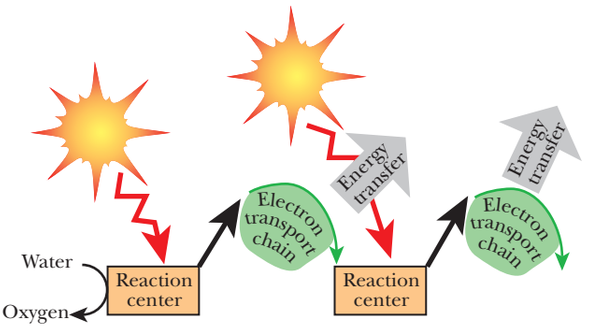
shared the enzymatic functions. Later, RNA lost most of its enzymatic roles as the more versatile proteins took these over (Fig. 20.07). It is generally thought that RNA was the first information storage molecule and that DNA was a later invention. Because DNA is more stable than RNA, it would store and transmit information with fewer errors.

This primitive cell vaguely resembles primitive bacteria, and lived off the organic compounds in the primitive soup. Eventually the supply of pre-made organic molecules was consumed. The proto-cell was forced to find a new source of energy and it turned to the sun (Fig. 20.08A). The earliest forms of photosynthesis probably used solar energy coupled to the use of sulfur compounds to provide reducing power. Later, more advanced photosynthesis used water instead of sulfur compounds. The water was split, releasing oxygen into the atmosphere.

Primitive photosynthesis probably used sulfur compounds as a source of electrons, more advanced photosynthesis resulted in the release of oxygen from water.

Before this, the atmosphere had been void of oxygen. The addition of oxygen completely altered the primitive earth. Once oxygen became available, respiration could

A. PHOTOSYNTHETIC APPARATUS



B. RESPIRATION RELIES ON OXYGEN FROM PHOTOSYNTHESIS

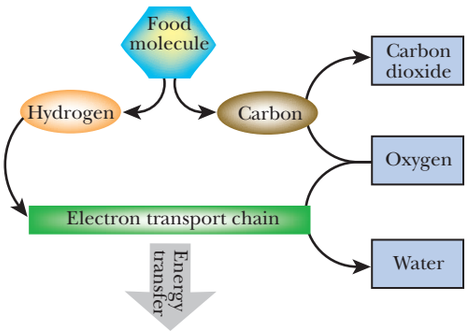


FIGURE 20.08
Development of photosynthesis sets the stage for respiration

(A) When early primitive cells ran out of pre-made organic molecules to supply energy, the cells started using the energy of the sun. The light energy was harvested by reaction centers that used water to provide the reducing power and released molecular oxygen into the atmosphere. Electrons carrying energy derived from the sunlight passed down electron transport chains. This allowed the conversion of the energy to a form used by the cell. (B) The accumulation of oxygen in the atmosphere allowed some of the proto-cells to change their biochemistry. Instead of using water and sunlight, these cells started to use oxygen to oxidize organic matter. This provided the cell with energy and released carbon dioxide and water into the environment.

be developed. Cells reorganized components from the photosynthetic machinery to release energy by oxidizing food molecules with oxygen (Fig. 20.08B). Photosynthesis emits oxygen and consumes carbon dioxide, whereas respiration does the reverse. The overall result is an ecosystem where plants and animals complement each other biochemically.

The Autotrophic Theory of the Origin of Metabolism

There is an alternative theory for the chemical origin of life. According to this view, the first proto-cells were not heterotrophic scavengers of organic molecules but were autotrophic and fixed carbon dioxide into organic matter themselves. An autotroph is defined as any organism that uses an inorganic source of carbon and makes its own organic matter as opposed to a heterotroph, which uses pre-made organic matter. The most familiar autotrophs are plants that use energy from sunlight to convert carbon dioxide into sugar derivatives. However, a variety of bacteria exist that fix carbon dioxide without light but instead rely on other sources of energy. Furthermore, the pathways of carbon dioxide fixation vary. In particular, some autotrophic bacteria incorporate carbon dioxide into carboxylic acids rather than generating sugar derivatives like plants.

An alternative theory suggests that energy released by the reaction of sulfur compounds powered the earliest life forms.

The autotrophic theory of the origin of life postulates the chemical oxidation of readily available iron compounds as the primeval energy source. In particular, the conversion of ferrous sulfide (FeS) to pyrite (FeS_2) by hydrogen sulfide (H_2S) releases energy and provides H atoms to reduce carbon dioxide to organic matter. [Anaerobic bacteria are found today that generate energy by the oxidation of iron Fe^{2+} compounds to Fe^{3+}, as well as others that generate energy by oxidizing sulfur compounds. Thus a primeval metabolism based on iron and sulfur seems reasonable.]

Several possible schemes have been suggested for the first carbon dioxide fixation reactions. One scheme involves iron catalyzed insertion of CO_2 into sulfur derivatives of those carboxylic acids still found today as metabolic intermediates (e.g., acetic acid, pyruvic acid, succinic acid, etc.). These early reactions would have occurred on the surface of iron sulfide minerals buried underground, rather than in a primeval soup. This leaves open the question of where such organic acids came from originally. One possibility is that they resulted from a Miller type synthesis, as described above. More radical is the suggestion that the first organic molecules were derived directly from carbon monoxide plus hydrogen sulfide. It has been demonstrated that a mixed FeS/NiS catalyst can convert carbon monoxide (CO) plus methane thiol (CH_3SH) into a thioester (CH_3-CO-SCH_3), which then hydrolyses into acetic acid. Inclusion of catalytic amounts of selenium allows conversion of CO plus H_2S alone to CH_3SH (and then to the thioester and acetic acid).

Organic matter can be generated from simple gas molecules using metallic catalysts.

Recently it was shown that carbon monoxide (CO) activated by the same mixed FeS/NiS catalyst can also drive the formation of peptide bonds between alpha-amino acids in hot aqueous solution. Not surprisingly, this system will also hydrolyze polypeptides.

Evolution of DNA, RNA and Protein Sequences

Consider the genes of an ancient ancestral organism. Over millions of years, mutations will occur in the DNA sequences of its genes at a slow but steady rate (see Ch. 13). Most mutations will be selected against because they are detrimental, but some will survive. Most mutations that are incorporated permanently into the genes will be neutral mutations with no harmful or beneficial effects on the organism. Occasionally, mutations that improve the function of a gene and/or the protein encoded by it will occur, although these are relatively rare. Sometimes a mutation that was originally harmful may turn out to be beneficial under new environmental conditions.

Due to mutation, the sequences of DNA and the encoded proteins will gradually change over long periods of time.

The actual function of the protein matters the most, not the exact sequence of the gene. If the protein can still operate normally, a mutation in the gene may be acceptable. Many of the amino acids making up a protein chain can be varied, within reasonable limits, without damaging the function of the protein too much. Replacement of one amino acid by a similar one (i.e. a conservative substitution—see Chapter 13) will rarely abolish the function of a protein. If we compare the sequences of the same protein taken from many different modern-day organisms we will find that the sequences can be lined up and are very similar. For example, the α chain of hemoglobin is identical in humans and chimpanzees. Yet, 13% of the hemoglobin amino acids are different in pigs compared with humans, 25% are different in chickens, and 50% are different in fish. This divergence in sequence is much what is expected from other estimates of evolutionary relatedness. Fig. 20.09 has an example of one of these alignments. It shows a highly conserved iron binding site found in a related family of enzymes—a group of alcohol dehydrogenases found in microorganisms—that use iron in their active site mechanism.

Related organisms contain genes and proteins with related sequences. These may be used to construct evolutionary trees.

It is possible, then, to construct an evolutionary tree using a set of sequences for a protein as long as it is found in all the creatures being compared. The α chain of hemoglobin is only found in our blood relatives. In contrast, cytochrome c is a protein involved in energy generation in all higher organisms, including plants and fungi. It even has recognizable relatives in many bacteria. A cytochrome c tree is shown in Figure 20.10. Humans and fish differ in amino acid sequence by only 18% for cytochrome c. From humans to either plants or fungi gives about 45% divergence. However plants and fungi also differ by 45%, which tells us that, by this measure, plants have diverged as far from fungi as animals have from plants.

Individual mutations may revert and restore the ancestral sequence of a gene or protein at a particular location. However, genes almost never mutate backwards to resemble the ancestors they diverged from many mutations ago. This is essentially a matter of probability. There is nothing forbidding any particular mutation to revert to

FIGURE 20.09 ***Alignment of Related Sequences***

Amino acid sequences of related polypeptides are given in one letter code. The conserved iron binding site is shown in bold. In particular, the iron atom is bound by the two conserved histidine (H) residues. The glycerol dyhydrogenase from *Bacillus* is related to the other members of this protein family but no longer uses iron and, as can be seen, the iron binding sequence has diverged and both histidines have been replaced by other amino acids.

```
ECO    NPVARERVHS  AATIAGIAFA  NAFL GVCHSM  AHKLGSQFHI  PHG LANALLI  CHNVIRYNAND
SALM   NPVARERLHS  AAYIAPIAFA  NAFL GVCHWM  AHKLPAQLHI  PHG PFNARYR  HSVRR   AQS
CLOE   NEKAREKMAH  ASTMAGMASA  NAFL GLCHSM  AIKLSSEHNI  PSG IANALLI  EEVIRKFNAVD
CLOB   E  AREQMHY  AQCLAGMAFS  NALL GICHSM  AHKTGAVFHI  PHG CANAIYL  PYVIKFNSKT
ZYMMO  DMPAREAMAY  AQFLAGMAFN  NASL GYVHAM  AHQLGGYYNL  PHG VCNAVLL  PHVLAYNASV
ENTH   DLEAREKMHN  AATIAGMAFA  SAFL GMDHSM  AHKVGAAFHL  PHG RCVAVLL  PYHVIRYNGQ
YEAST  DKKARTDMCY  AEYLAGMAFN  NASL GYVHAL  AHQLGGFYHL  PHG VCNAVLL  PHVQ  EANM
PRD    D  AGEEMAL  GQYVAGMGFS  NVGL GLVHGM  AHPLGAFYNT  PHG VANAILL  PHVMRYNADF
GLD    EKCEQTLFKY  GK            LAYESVK  AKVVTPALE    AVVEANTL  LSGLGFESGG

                                IRON BINDING SITE
```

ECO = *E. coli* alcohol dehydrogenase
SALM = *Salmonella typhimurium* alcohol dehydrogenase
CLOB = *Clostridium acetobutylicum* butanol dehydrogenase
CLOE = *C. acetobutylicum* alcohol dehydrogenase
ZYMMO = *Zymomonas mobilis* alcohol dehydrogenase II
ENTH = *Entamoeba histolytica* alcohol dehydrogenase
YEAST = Yeast alcohol dehydrogenase IV
PRD = *E. coli* propanediol dehydrogenase
GLD = *Bacillus* glycerol dehydrogenase (a zinc enzyme)

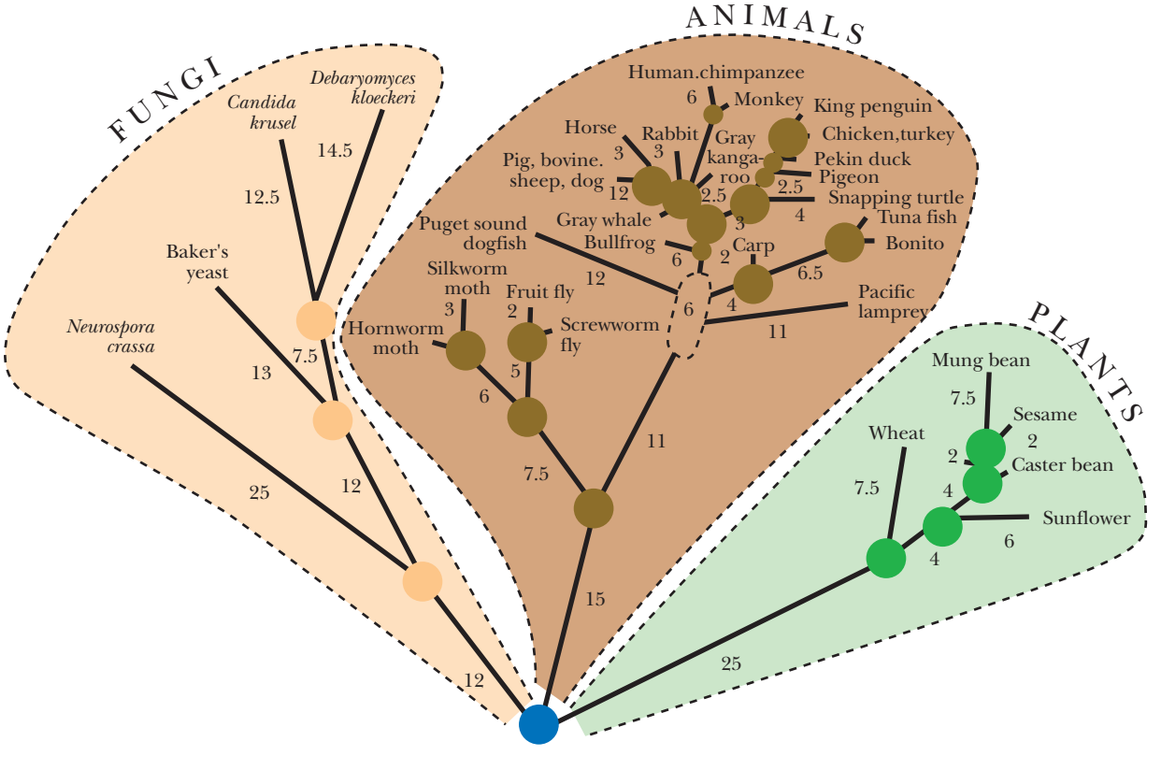


FIGURE 20.10 ***Evolutionary Tree based on Cytochrome c***

All three kingdoms, fungi, animals, and plants, have the gene for cytochrome c, which is involved in production of energy. This phylogenetic tree was constructed by comparing the amino acid sequence of cytochrome c from each of these organisms. The closer the branches are together, the more similar the sequences. The numbers along the branches represent the number of amino acid differences.

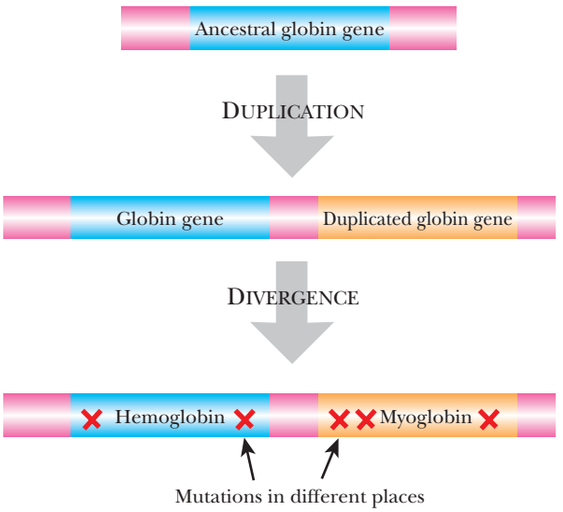


FIGURE 20.11 ***Duplication Creates New Genes***

During evolution, the entire ancestral globin gene was duplicated. The two copies were then able to diverge independently of each other. The first gene developed into hemoglobin, which is only found in red blood cells. The second copy developed into myoglobin, which is found in muscle tissue. Both proteins still carry oxygen, but they have tissue specificity.

the original sequence, but the likelihood of precisely reversing each of many mutations is infinitesimally small.

Creating New Genes by Duplication

Duplication of segments of DNA creates a supply of new genes.

How do new genes arise? The standard way is via gene duplication. As discussed in chapter 13, mutations may cause the duplication of a segment of DNA that carries a whole gene or several genes. The original copy must be kept for its original function but the extra copy is free to mutate and may be extensively altered. In most cases the mutations that accumulate will inactivate the duplicate copy. Less often, the extra copy will remain active and be altered so as to perform a related but different function from the original copy.

The steady accumulation of mutations in duplicated genes results in families of genes with related sequences.

Multiple duplication followed by sequence divergence may result in a family of related genes that carry out related functions. One of the best examples is the **globin** family of genes. Hemoglobin carries oxygen in the blood, whereas myoglobin carries it in muscle. These two proteins have much the same function, have similar 3-D shapes, and their sequences are related. After the ancestral globin gene duplicated, the two genes for hemoglobin and myoglobin slowly diverged as they specialized to operate in different tissues (Fig. 20.11).

The actual hemoglobin of mammalian blood has two alpha (α)-globin and two beta (β)-globin chains forming an α2/β2 tetramer, unlike myoglobin, which is a monomer of a single polypeptide chain. The α-globin and β-globin were derived by further duplication of the ancestral hemoglobin gene. In addition, the ancestral α-globin gene split again, to give modern α-globin and zeta (ζ)-globin. The ancestral β-globin gene split again, twice, to give modern β-globin and the gamma (γ)-, delta (δ)-, and epsilon (ε)-globins (Fig. 20.12).

These globin variants are used during different stages of development. At each stage, the hemoglobin tetramer consists of two α-type and two β-type chains. The ζ-globin and ε-globin chains appear in early embryos, which possess ζ2/ε2 hemoglobin. In the fetus, the ε-chain is replaced by the γ-chain and the ζ-chain is replaced by the α-chain, so giving α2/γ2 hemoglobin. A fetus needs to attract oxygen away from the mother's blood, so the α2/γ2 hemoglobin binds oxygen better than the adult α2/β2 hemoglobin (Fig. 20.12).

globins Family of related proteins, including hemoglobin and myoglobin, that carry oxygen in the blood and tissues of animals

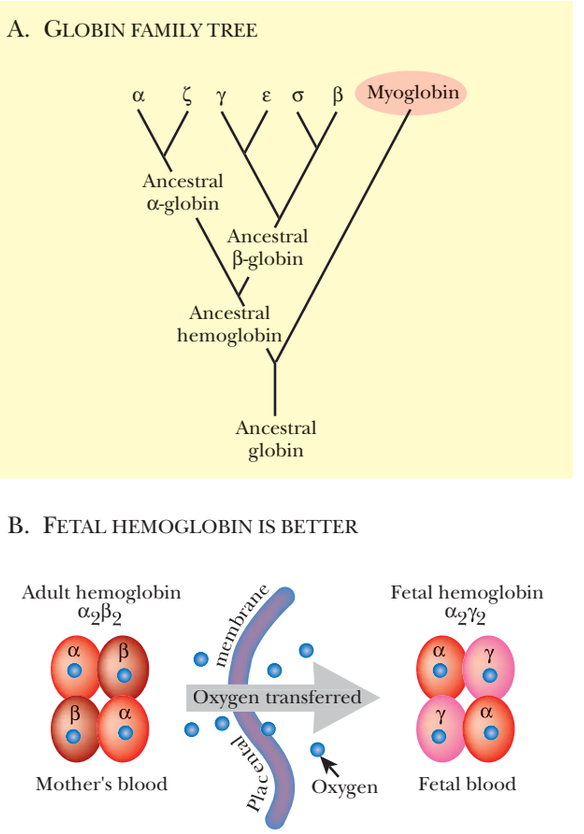


FIGURE 20.12 ***Origin of Globin Family of Genes***

(A) Over the course of evolution, a variety of gene duplication and divergence events gave rise to a family of closely related genes. The first ancestral globin gene was duplicated giving hemoglobin and myoglobin. After another duplication, the hemoglobin gene diverged into the ancestral α-globin and ancestral β-globin genes. Continued duplication and divergence created the entire family of globin genes. (B) The different members of the hemoglobin family are adapted for specific functions during development. Thus the fetus uses two α chains and two γ chains to form its hemoglobin tetramer. This form is able to extract the oxygen from the mother's blood because it has a higher affinity for oxygen than the adult form of hemoglobin.

The globin genes are an example of a **gene family**, a group of closely related genes that arose by successive duplication. The individual members are obviously related in their sequences and carry out similar roles. During evolution, continued gene duplication may give rise to multiple new genes whose functions steadily diverge until their ancestry may be difficult to recognize; this gives a **gene superfamily**. The genes of the immune system provide good examples of gene families and superfamilies.

Reverse transcriptase may generate duplicate genes that lack introns and promoters and are located far away from the original copy.

In eukaryotes, retro-elements that encode reverse transcriptase are relatively common (see Ch. 15). Consequently, occasional reverse transcription of cellular mRNA molecules may occur. This gives a complementary DNA copy that may be integrated into the genome. This results in a duplicate copy of the gene, although this lacks the introns and promoter of the original gene. Such inactive copies are known as pseudogenes and usually accumulate mutations that inactivate the coding sequence. Rarely, a pseudogene may end up next to a functional promoter and be expressed. This gives a duplicate functional copy of the original gene that may be altered by mutation as already discussed.

Occasionally whole genomes may be duplicated.

Rare mistakes during cell division may result in the whole genome being duplicated. In particular, errors in meiosis may give diploid gametes. Fusion of two diploid gametes would give a tetraploid zygote and hence a tetraploid individual. More often, a triploid individual forms by fusion of one mutant diploid gamete plus one normal haploid gamete. Most triploids are sterile, as they give gametes with incorrect numbers of chromosomes. But occasionally triploids will be able to generate tetraploid progeny. Aberrant ploidy levels are fairly common in plants. Around 5 in 1000 plant gametes

gene family Group of closely related genes that arose by successive duplication and perform similar roles
gene superfamily Group of related genes that arose by several stages of successive duplication. Members of a superfamily have often diverged so far that their ancestry may be difficult to recognize

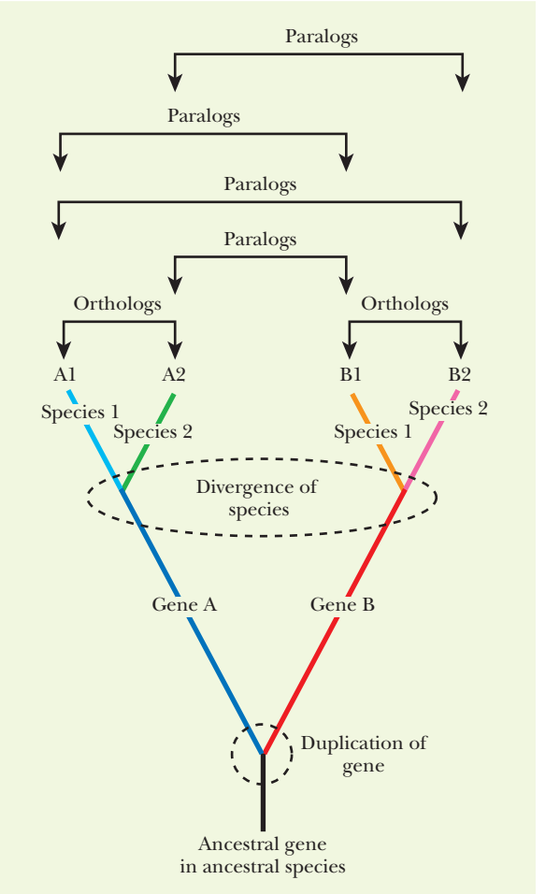


FIGURE 20.13 ***Paralogous and Orthologous Sequences***

In this example, an ancestral gene duplicated and diverged into genes A and B, which by definition are paralogs. These two genes were both present when the ancestral species diverged into species 1 and species 2. Thus both species 1 and species 2 have genes for A and B (referred to as A1 & B1 and A2 & B2 respectively). Each such pair are still paralogs. However, since species 1 and 2 are now two separate species, the A1 and A2 genes are orthologs, and the B1 and B2 genes are also orthologs.

are diploid. Therefore, in a cross between two different parents, approximately 2.5 in 10^{-5} zygotes will be tetraploid. Over time, the duplicate copies of genes in a tetraploid organism will gradually diverge. Eventually, once the duplicate copies diverged far enough to be distinct and have assumed new functions, the organism will effectively become "diploid" again.

Paralogous and Orthologous Sequences

A gene may be related to similar genes in other organisms and also to other genes of similar sequence within the same organism.

Sequences are said to be **homologous** when they share a common ancestral sequence. If several organisms each contain single copies of a particular gene that all derive from the same common ancestor then sequence comparison should give an accurate evolutionary tree. However, gene duplication may result in multiple copies of the same gene within a single organism. These alternatives are illustrated in Fig. 20.13. **Orthologous** genes are those found in separate species and which diverged when the organisms containing them diverged. **Paralogous** genes are multiple copies located within the same organism due to gene duplication.

To generate an accurate evolutionary tree orthologous genes must be compared. For example, we must compare the sequences of α-globin from one animal with the orthologous α-globin from another and not with the paralogous β-globin. Since paralogous sets of genes have similar sequences, these may cause confusion unless their

homologous Related in sequence to an extent that implies common genetic ancestry
orthologous genes Homologous genes that are found in separate species and which diverged when the organisms containing them diverged
paralogous genes Homologous genes that are located within the same organism due to gene duplication

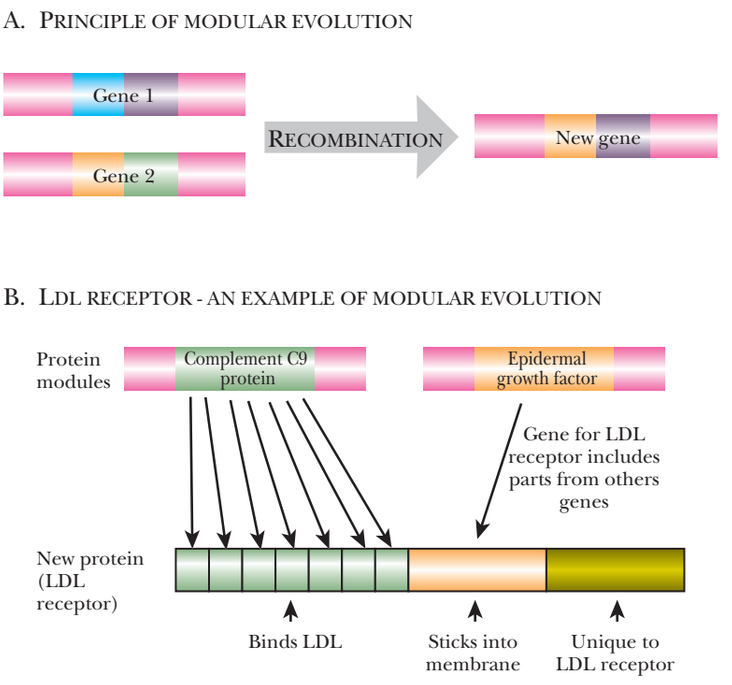


FIGURE 20.14 ***Novel Gene Derived from Pre-Existing Modules***

(A) Modular evolution of a new gene may involve the fusion of separate gene modules, or functional units. For example, the orange module of gene 1 may provide a function that complements the purple module of gene 2. If these two domains fuse they might then form a novel but functional gene. (B) The LDL receptor has domains found in several other proteins. The first part of the LDL receptor gene has seven repeated modules or functional units also found in the C9 complement factor of the immune system. The next module allows the protein to bind the cell membrane. This module is very similar to a portion of the epidermal growth factor receptor. Finally, the LDL receptor has a module that is unique.

correct ancestry is known. In particular, we need to know whether or not an organism contains multiple sequences derived from the same ancestor. Suppose that, due to partial information, we only knew of α-globin from pigs and β-globin from dogs. If we were unaware of the presence of other members of the globin family in these organisms we might compare these two sequences as if they were orthologous. This would lead to an incorrect relationship.

Creating New Genes by Shuffling

New genes may be made by shuffling of modular segments of DNA.

Another way to create new genes is by using pre-made modules. Segments from two or more genes may be fused together by DNA rearrangements so generating a novel gene consisting of regions derived from several sources (Fig. 20.14A). An example of the formation of a new gene from several diverse components is the LDL receptor (Fig. 20.14B). LDL is low-density lipoprotein that carries cholesterol around in the blood. The LDL receptor is found on the surface of cells that take up LDL. The gene for the LDL receptor consists of several regions, two of which are derived from other genes. Towards the front are seven repeats of a sequence also appearing in the C9 factor of complement, an immune system protein. Farther along is a segment related to part of epidermal growth factor (a hormone). When such a gene mosaic is transcribed and translated, we get a patchwork protein consisting of several different domains.

Different Proteins Evolve at Very Different Rates

The sequences of different genes and proteins evolve at different rates. Less critical sequences are free to evolve faster.

Obviously, we should not rely on a single protein to build an evolutionary tree. If we make trees for several proteins, we often get rather similar evolutionary relationships. However, different proteins evolve at different speeds. As noted above, humans and fish differ by 50% in the α chain of hemoglobin but by less than 20% in their cytochrome c. If we plot the number of amino acid changes versus the evolutionary time scale (Fig. 20.15), we can see this easily for cytochrome c (slow), hemoglobin (both α and β chains evolve at medium speed), and fibrinopeptides A and B (rapid evolution).

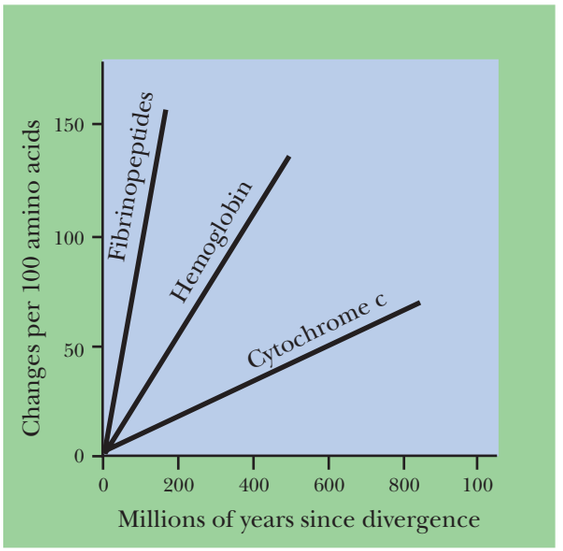


FIGURE 20.15 ***Rates of Protein Evolution***

During the course of evolution, some proteins accumulate more mutations than others. The cytochrome c gene is very stable, and only 50 changes/100 amino acids have occurred in 800 million years. Fibrinopeptides A and B, on the other hand, have accumulated 50 changes/100 amino acids in less than 100 million years.

TABLE 20.05 Rates of Evolution for Different Proteins

Protein	Rate of Evolution
Neurotoxins	110–125
Immunoglobulins	100–140
Fibrinopeptide B	91
Fibrinopeptide A	59
Insulin C peptide	53
Lysozyme	40
Hemoglobin α chain	27
Hemoglobin β chain	30
Somatotropin	25
Insulin	7.1
Cytochrome c	6.7
Histone H2	1.7
Histone H4	0.25

The rate of evolution is given as the number of mutations per 100 amino acid residues per 100 million years.

Table 20.05 gives the evolutionary rates for an assortment of proteins. Fibrinopeptides are involved in the blood clotting process. They need an arginine at the end and must be mildly acidic overall. Apart from this they can vary widely as there are so few constraints on what is needed. In contrast, histones bind to DNA and are responsible for its correct folding. Almost all changes to a histone would be lethal for the cell, so they evolve extremely slowly.

Cytochrome c is an enzyme whose function depends most critically on a few amino acid residues at the active site, which bind to its heme cofactor. Consequently, these active site residues rarely vary, even though amino acids around them change. Of 104 residues, only Cys-17, His-18 and Met-80 are totally invariant. In other places variation is low; large, nonpolar, amino acid residues always fill positions 35 and 36. Several cytochrome c molecules have been examined by X-ray crystallography and all have the same 3-D structure. Although cytochrome c molecules may vary by as many

as 88% of their residues, they retain the same 3-D conformation. Thus, little variation is seen with the amino acids that are essential to the function or structure of cytochrome c.

Insulin is a hormone that evolves at much the same rate as cytochrome c. Insulin consists of two protein chains (A and B) encoded by a single insulin gene. During protein synthesis, a long pro-insulin molecule is made. This has the middle, the C-peptide, cut out and discarded. Disulfide bonds hold the A and B chains together. Since the C chain is not part of the final hormone, it is free to evolve much faster and it changes at almost 10 times the rate of the A and B chains. Notice that all these proteins maintain the critical residues throughout evolution. It is important to note that mutations are random. A mutation is just as likely to occur in the A, B, and C portion of the insulin gene. The mutations that do occur in the A and B portions may be detrimental to the organism, therefore, these mutations never get passed to a new generation. On the other hand mutations in the C portion occur, do not harm the function of the protein, and get passed onto the progeny.

Molecular Clocks to Track Evolution

A rapidly evolving protein will eventually become so altered in sequence between diverging organisms that the relationship will no longer be recognizable. Conversely, a protein that evolves very slowly will show little or no difference between two different organisms. Therefore, we need to use slowly changing sequences to work out distant evolutionary relationships and fast-evolving sequences for closely related organisms.

Rapidly evolving sequences reveal the relationships between closely related organisms. Slowly changing sequences are needed to compare distantly related organisms.

Most human proteins have identical sequences to those of the closely related chimpanzee. Even if we examine the rapidly evolving fibrinopeptides, humans and chimpanzees end up on the same branch of the evolutionary tree. So how can we tell people apart from chimps? Mutations that do not affect the sequence of proteins accumulate much faster during evolution, since they have little or no detrimental effect. So if instead of protein sequences we look at the DNA sequences of very closely related organisms we find many more differences. These are found mainly in non-coding sequences and in the third codon position. As discussed in Chapter 8, changing the third base of most codons does not alter the encoded amino acid. So changing the DNA sequence at the third base of most codons leaves the encoded protein unaltered (Fig. 20.16).

Introns are non-coding sequences that are spliced out of the primary transcript and so do not appear in the messenger RNA (see Ch. 12). Intron sequences are therefore not represented in the final protein. Apart from the intron boundaries and splice recognition sites, the DNA sequence of an intron is free to mutate extensively. Other non-coding sequences exist between genes and, if not involved in regulation, they are also relatively free to mutate.

The early data on cytochrome c, hemoglobin, etc., were obtained by direct sequencing of proteins. Since DNA sequencing is much easier and more accurate, most of the recently discovered protein sequences are deduced from the DNA sequence. Hence, we have a lot of DNA information on closely related animals. Using this data will help fine-tune the evolutionary relationship between animals such as humans and chimpanzees.

Ribosomal RNA—A Slowly Ticking Clock

A major problem is how to construct an evolutionary tree that includes all living organisms and shows the relationships between all the main groups of organisms. To achieve this, first we need a molecule that is present in all organisms. Second, the chosen molecule must evolve slowly so as to still be recognizable in all the major groups of living things.

Ribosomal RNA sequences have been used to construct a global evolutionary tree.

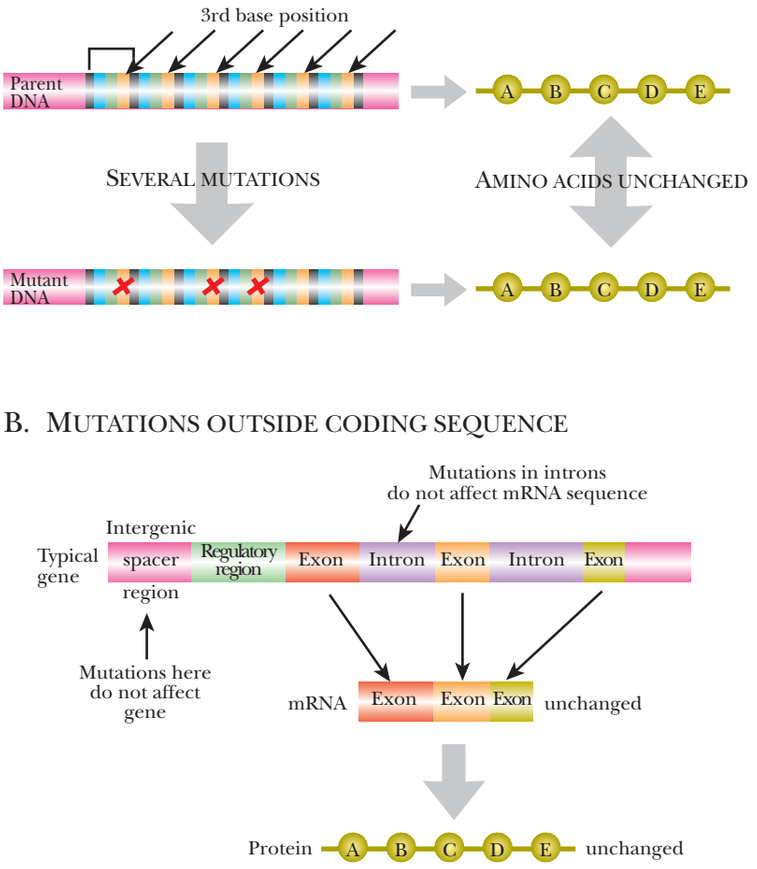


FIGURE 20.16 ***Non-Coding DNA Evolves Faster***

(A) During evolution mutations in the third position of the triplet codon rarely alter the amino acid sequence of the protein; therefore they are rarely deleterious. (B) Non-coding DNA regions often have no obvious function and so many mutations accumulate within these regions. Mutations in intragenic spacer regions or in introns have no effect on protein sequence or function.

Although histones evolve very slowly, only eukaryotic cells possess them; they are missing from bacteria. The solution is to use ribosomal RNA. In practice, the DNA of the genes that encode the RNA of the small ribosomal subunit (16S or 18S rRNA) is sequenced and the rRNA sequence is then deduced. All living organisms have to make proteins and they all have ribosomes. Furthermore, since protein synthesis is so vital, ribosomal components are highly constrained and evolve slowly. The only group excluded is the viruses, which have no ribosomes. (Whether viruses are truly alive is debatable and their evolutionary origins are still controversial; see Ch. 17.)

Sequence analysis indicates that fungi are closer to being immobile animals than non-photosynthetic plants.

Use of relationships based on ribosomal RNA has allowed the creation of large-scale evolutionary trees encompassing all the major groups of organisms. Higher organisms consist of three main groups—plants, animals and fungi (Fig. 20.17). Analysis of rRNA indicates that the ancestral fungus was never photosynthetic but split off from the plant ancestor before the capture of the chloroplast. Despite traditionally being studied by botanists, fungi are actually more closely related to animals than plants. A variety of single celled organisms sprout off the eukaryotic tree near the bottom and do not fall into any of the three major kingdoms.

Sequencing of rRNA reveals that chloroplasts and mitochondria are related to the bacteria.

As discussed in Chapter 19, most eukaryotic cells contain mitochondria and, in addition, plant cells contain chloroplasts. These organelles are derived from symbiotic bacteria and contain their own ribosomes. The rRNA sequences of mitochondria and chloroplasts reveal their relationship to the bacteria. Relationships among eukaryotes, like that shown in Fig. 20.17, are therefore made by using the rRNA of the ribosomes found in the cytoplasm of eukaryotic cells. These ribosomes have their ribosomal RNA coded for by genes in the cell nucleus.

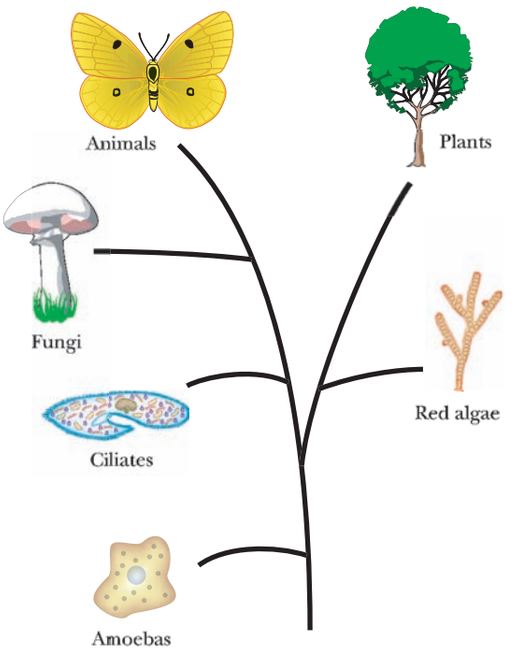


FIGURE 20.17 ***Eukaryote Kingdoms Based on Ribosomal RNA Sequences***

Comparing the ribosomal RNA sequences has allowed scientists to deduce how closely the major divisions of organisms are related. Protozoans such as amoebas were the earliest groups to branch from the ancestral eukaryotic lineage. A division between photosynthetic and non-photosynthetic organisms occurred next. The two branches evolved separately, the photosynthetic branch formed algae and higher plants, whereas the non-photosynthetic branch developed into ciliates, fungi and animals.

Life consists of three domains—the eubacteria, the archaebacteria and the eukaryotes.

When an rRNA-based tree is made that includes both prokaryotes and eukaryotes it turns out that life on Earth consists of three lineages (Fig. 20.18). These three domains of life are the **eubacteria** ("true" bacteria, including the organelles), the **archaea** or **archaebacteria** ("ancient" bacteria) and the eukaryotes. There is as much difference between the two genetically distinct types of prokaryote as between prokaryotes and eukaryotes. Sequencing of organelle rRNA indicates that mitochondria and chloroplasts belong to the eubacterial lineage.

One bizarre aspect of classifying life forms by ribosomal RNA is that the organism itself is not needed. A sample of DNA containing the genes for 16S rRNA is sufficient. Although many microorganisms present in the sea or in soil have never been successfully cultured, DNA can be extracted from the soil or seawater directly. Using PCR (see Ch. 23) it is possible to amplify the DNA from a single cell and get enough of the 16S rRNA gene to obtain a sequence. Several new groups of bacteria that branched off very early from the archaebacterial lineage have been discovered by this method, despite not being successfully cultured.

The Archaebacteria versus the Eubacteria

The archaebacteria are somewhat closer to the eukaryotes than to the eubacteria.

Although most common bacteria are eubacteria, there is another group, the archaebacteria or archaea as they have recently been renamed. Both types of bacteria have microscopic cells without a nucleus. They both have single circular chromosomes and divide in two by simple binary fission. In short, they both conform to the definition of a prokaryotic cell and there was no obvious reason to suspect from their superficial structure that they were so radically different. However, sequence analysis of ribosomal RNA indicates that there is about as much genetic difference between the eubacteria and archaea as between either of these two groups and eukaryotic cells.

Of the two groups of prokaryotes, the archaea are probably slightly more closely related to the **urkaryote**, the primeval ancestor of the eukaryotic nucleus. Some

archaebacteria One of the three domains of life comprising the "ancient" bacteria
archaea New name for archaebacteria, one of the three domains of life
eubacteria One of the three domains of life comprising the "true" bacteria, including the organelles
urkaryote The hypothetical primeval ancestor of the eukaryotic nucleus

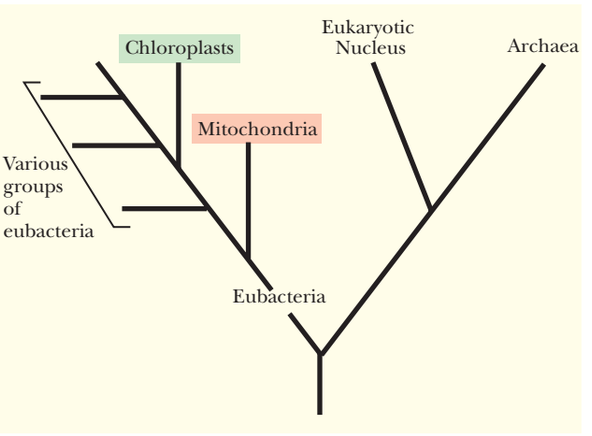


FIGURE 20.18 ***The Three Domains of Life***

All of today's organisms belong to one of three main divisions based on relationships among ribosomal RNA: the eubacteria, the archaebacteria (or archaea) and the eukaryotes. Mitochondria and chloroplasts have rRNA that most closely resembles eubacteria.

archaea have their DNA packaged by histone like proteins that show some sequence homology to the true histones of higher organisms. In addition, the details of protein synthesis and the translation factors of archaea resemble those of eukaryotes, rather than eubacteria. These similarities have led to the suggestion that the primeval eukaryote evolved from an archaeal ancestor.

Archaea differ biochemically from eubacteria in several other major respects. Archaea have no peptidoglycan and their cytoplasmic membrane contains unusual lipids, which are made up from C5 isoprenoid units rather than C2 units as with normal fatty acids (Fig. 20.21). Moreover, the isoprenoid chains are attached to glycerol by ether linkages instead of esters. Some double-length isoprenoid hydrocarbon chains stretch across the whole membrane.

Archaea tend to be found in bizarre environments and many of them are adapted to extreme conditions. They are found in hot sulfur springs, thermal vents in the ocean floor, in the super salty Dead Sea and Great Salt Lake and also in the intestines of cows (and other animals) where they make methane. Some examples of archaea are:

Halobacteria: These are extremely salt-tolerant and grow in up to 5M NaCl but will not grow below 2.5M NaCl (sea water is only 0.6M). They trap energy from sunlight by using bacterio-rhodopsin, a molecule related to the rhodopsin pigment used as a photo-detector in animal eyes.

Methanogens (methane producing bacteria): These are obligate anaerobes and are very sensitive to oxygen. They convert H_2 plus CO_2 to CH_4 (methane). Their metabolism is unique—they contain coenzymes found in no other living organisms but have no typical flavins or quinones.

Sulfolobus: Lives in geothermal springs and grows best at a pH optimum of 2–3 and a temperature of 70–80°C. These archaea oxidize sulfur to sulfuric acid. Many other sulfur-metabolizing archaea are also found living under a variety of extreme conditions.

DNA Sequencing and Biological Classification

Before the sequencing of DNA became routine, animals and plants were classified reasonably well, fungi and other primitive eukaryotes were classified poorly, and bacterial classification was a lost cause due to lack of observable characters. Using gene sequences for classification was developed for bacteria and has since spread to other types of organism. Nowadays ancestries may be traced by comparing the sequences of DNA, RNA or proteins that are more representative of fundamental genetic relationships than are many superficial characteristics. Furthermore, in situations where division into species, genera, families, etc., is arbitrary, sequence data can provide

Instant Evolution of Ribosomal RNA

Consider an essential molecule that evolves slowly, such as histones or ribosomal RNA. It is possible that certain combinations of two mutations might yield a functional molecule, but that either alone would be lethal. For example, a mutation from G to C that destroyed a critical GC base pair in a stem loop structure might be fatal in 16S rRNA. However, replacing the GC with a CG base pair might well allow function (Fig. 20.19). During normal evolution, this replacement is highly unlikely since either single mutation is lethal and the likelihood of simultaneous mutations in just these two bases is very low.

Consequently, a CG base pair in this particular position will probably be very rare among the 16S rRNA sequences of existing life forms. To fully analyze the relationship of structure and function in a molecule such as rRNA, several artificial mutations must be introduced simultaneously. This may be done by a procedure known as "instant evolution" that was developed in the laboratory of Dr. Philip R. Cunningham at Wayne State University. In this approach, 16S rRNA is mutated and those mutations that prevent protein synthesis are isolated. Next, suppressor mutations that restore protein synthesis are selected. Alternatively, multiple random mutations may be simultaneously introduced into a short region of the rRNA that is suspected to play an important role in protein synthesis. In either case, most of these mutations in rRNA would be lethal under normal circumstances; to avoid killing the bacteria they must be manipulated so that the mutant form of the 16S rRNA does not interfere with normal cellular protein synthesis.

The following technology was developed to prevent the mutated form of rRNA from affecting the normal bacterial functions (Fig. 20.20):

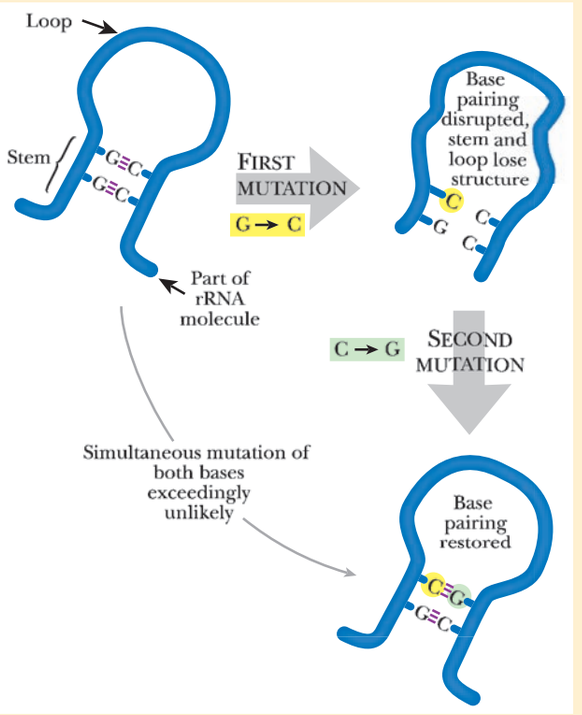


FIGURE 20.19 ***Single Change Lethal, Two Changes Functional***

Mutations in ribosomal RNA are lethal since they alter the stem-loop structure of the molecule. In this example, mutating the guanine to cytosine prevents a critical base pair from completely the stem portion. If a second mutation were to change the cytosine to a guanine, the base pair would reform and the ribosomal RNA would be functional again. Although changing both positions simultaneously is extremely unlikely, this would not be detrimental to the organism and the mutation would be passed onto successive generations.

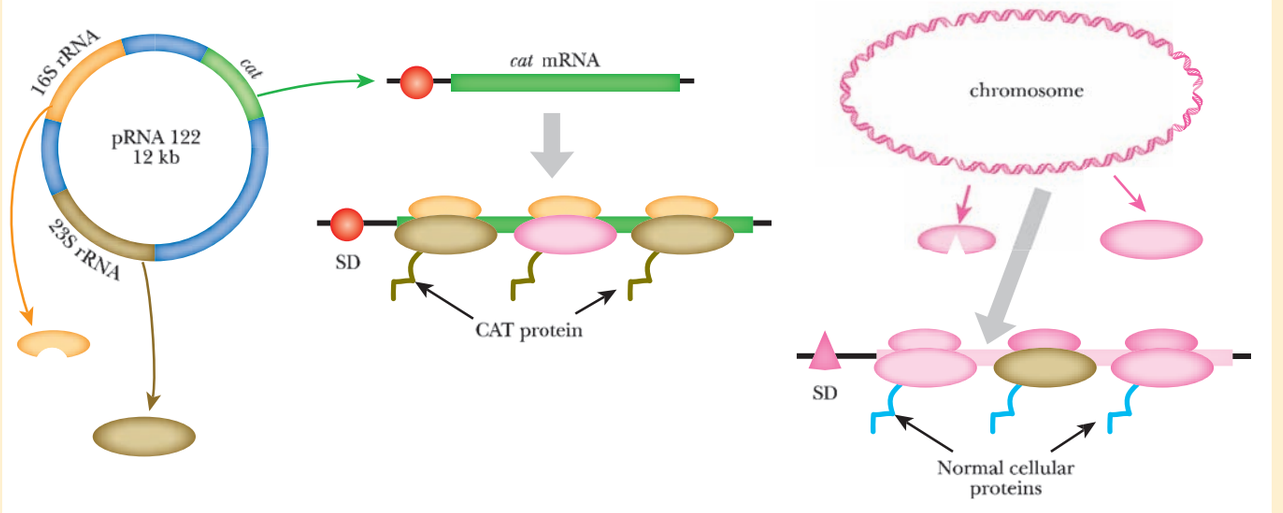


FIGURE 20.20 ***Instant Evolution of Ribosomal RNA***

The plasmid pRNA122 carries the sequence for altered 16S rRNA and the reporter gene (in this example, CAT). Separation of chromosomal and plasmid-directed protein synthesis occurs primarily due to the changes in the Shine-Dalgarno (SD) sequence. The 16S rRNA encoded by the plasmid cannot recognize the SD sequence of normal cellular mRNA, but does recognize the SD sequence upstream of the reporter gene (*cat*). If a mutation in the 16S rRNA prevents it from functioning, the reporter gene is not translated into CAT protein, and therefore the bacteria are not chloramphenicol resistant. Translation of normal cellular proteins occurs without interruption because the chromosomal copy of the 16S rRNA (small pink oval) is used. The chromosomal copy of the 16S rRNA does not recognize the SD sequence of the *cat* mRNA, and so does not allow synthesis of CAT protein.

a. A copy of the 16S rRNA gene was put on a plasmid and mutated. Since the genomic copy of the 16S rRNA is still functional, most of the cell's ribosomes will be normal. Only a fraction of the ribosomes will have the mutant 16S rRNA.

b. The anti-Shine-Dalgarno sequence of the plasmid 16S rRNA is altered so that it no longer recognizes normal cellular mRNA, thus lethal mutations in this 16S rRNA copy will not interfere with normal protein synthesis.

c. A reporter gene is designed with an altered Shine-Dalgarno sequence that matches the plasmid or mutated form of the 16S rRNA. Thus only the translation of the mRNA from the reporter gene responds to the mutations in the plasmid-borne copy of the 16S rRNA. Two different reporter genes were used, chloramphenicol acetyl transferase (CAT), which gives chloramphenicol resistance to the bacteria, and green fluorescent protein (GFP), which causes the bacteria to turn green under fluorescent light. The mutant 16S rRNA is therefore functionally isolated from the rest of the cell and can be analyzed by monitoring the expression of the two proteins CAT and GFP. Lethal mutations in 16S rRNA merely prevent expression of CAT and GFP without affecting the normal protein synthesis of the bacteria.

In these experiments nearly 60,000 different SD-anti-SD combinations were tried but only 13 were found to be functional without killing the cell. Researchers in the Cunningham laboratory showed that almost any change in the nucleotide sequence of the anti-SD was lethal to bacteria-probably because it disrupted the balance of protein synthesis. Since its development, instant evolution has been used by several researchers throughout the world to study the role of ribosomal RNA in protein synthesis. This technology may also allow the development of new antibiotics targeted against critical regions of the ribosome and that are not susceptible to the development of drug resistance.

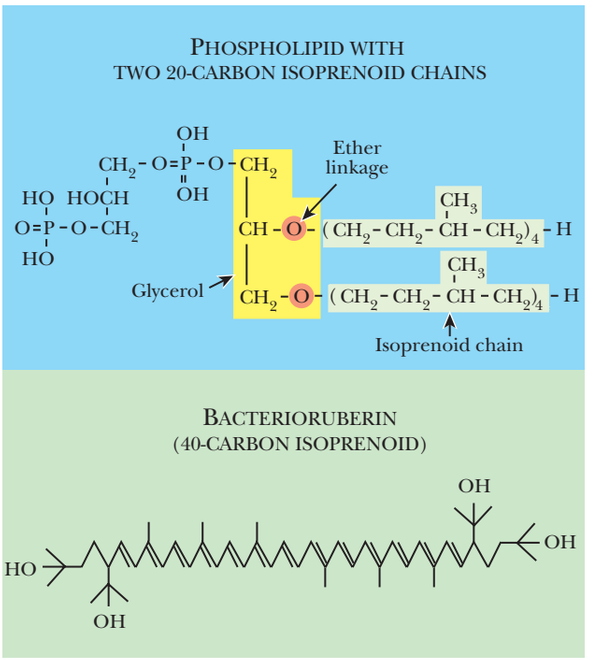


FIGURE 20.21 ***Unusual Lipids of Archaea***

The archaebacteria have lipid chains made of five carbon isoprenoid units rather than two carbon units as seen in eubacteria. The isoprenoid chains are linked to a glycerol via an ether link rather than an ester link. In some instances the isoprenoid lipid chains may contain 40 carbons (bacterioruberin, for example). These longer lipids span the whole membrane of the archaea.

quantitative measurements of genetic relatedness. Even if we cannot unambiguously define a species, we can be consistent in how much sequence divergence is needed to allocate organisms to different species or families.

Originally ribosomal RNA sequences were used for classification. However as ever more sequence data is obtained, including whole genomes, it is possible to take an increasing number of other genes into account. Computer programs exist for calculating the relative divergence of the sequences and can generate trees such as that in Figure 20.22. Here we have four bacteria, all in different genera but belonging to the same family, the Enterobacteria. To root such a tree correctly we also need the sequence from an organism in an "out-group;" in this case we have used the bacterium *Pseudomonas*, which is only distantly related to enteric bacteria. The nodes in Figure 20.22 represent the deduced common ancestors. The branch lengths are often scaled to represent the number of mutations needed and the numbers indicate how many base changes are needed to convert the sequence at each branch point into the next. (The total length of the 16s rRNA of Enteric bacteria is 1542 bases.)

Evolutionary trees are often constructed that show the number of sequence differences in rRNA.

As discussed in Ch. 19, parasites have many adaptations and quirks that developed due to the unusual environment they inhabit. Establishing phylogenetic relationships among parasites is very difficult based on simple trait analysis. Fortunately, gene sequences can often be used to trace the ancestry of parasitic or aberrant life forms. Aberrant development due to unusual environments is not only seen in parasites. Animals such as moles have adapted to living underground or in caves, and have lost their eyes in the process since these organs are not useful. Sometimes vestigial remnants of structures remain even though the animal has no use for the structure. Whales have the atrophied remains of hind limbs, which implies that whales are not real fish, but mammals that have become fish-like in general form as they have adapted to life in the ocean. Until gene sequencing emerged, it remained unknown which mammals are the whale's closest relatives. It now appears that whales are related to the artiodactyls, hoofed mammals such as hippos, giraffes, pigs and camels.

Sequences are especially useful in classifying aberrant organisms that have lost structural characters normally used for comparison.

One major problem with sequence comparison is that base changes can revert. Although statistical comparison of multiple sequences with many altered sites is often sufficient to establish a lineage, ambiguity sometimes remains. A useful way to help resolve ambiguities is by using conserved insertions or deletions—known as signature sequences or "indels". Although a single base insertion or deletion might possible revert, the likelihood that an insertion or deletion of several bases might revert so as

The ancestries of sequences can be confirmed if they share major insertions or deletions.

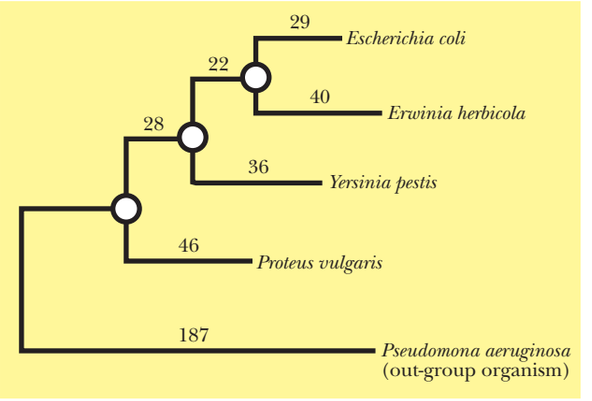


FIGURE 20.22 ***Phylogenetic Tree for Enteric Bacteria***

The phylogenetic relationship between bacteria can be deduced by comparing ribosomal RNA sequences. Here the sequences of the 16S rRNA genes are compared for the four enteric bacteria, *E. coli*, *Erwinia herbicola*, *Yersinia pestis*, and *Proteus vulgaris*. The relatively unrelated bacterium *Pseudomona aeruginaosa*, is used as an outgroup organism to provide the base or root of the tree. From these comparisons, it can be deduced that *P. vulgaris* was the first to branch from the primitive ancestor and that *E. coli* and *E. herbicola* were the latest.

to exactly restore the original length and sequence is vanishingly small. Consequently, if a subgroup of a family of related sequences all contain an indel of defined length and sequence at the same location, they must all have been derived from the same ancestral sequence.

Mitochondrial DNA—A Rapidly Ticking Clock

Mitochondrial DNA changes fast enough to be used to classify subgroups within the same species.

Although mitochondria contain circular molecules of DNA reminiscent of bacterial chromosomes, the mitochondrial genome is much smaller. The mitochondrial DNA codes for a few of the proteins and ribosomal RNA of the mitochondrion, but most components are now encoded by the eukaryotic nucleus as already discussed in Chapter 19. What concerns us here is that the mitochondrial DNA of animals accumulates mutations much faster than the nuclear genes. In particular, mutations accumulate rapidly in the third codon position of structural genes and even faster in the intergenic regulatory regions. This means that mitochondrial DNA can be used to study the relationships of closely related species or of races within the same species. Most of the variability in human mitochondrial DNA occurs within the D-loop segment of the regulatory region. Sequencing this segment allows us to distinguish between people of different racial groups.

One apparent drawback to using mitochondrial DNA is that mitochondria are all inherited from the mother. Although sperm cells do contain mitochondria, these are not released during fertilization of the egg cell and are not passed on to the descendants. On the other hand, analysis of mitochondria gives an unambiguous female ancestry, as complications due to recombination may be ignored. Furthermore, a eukaryotic cell contains only one nucleus but has many mitochondria so there are often thousands of copies of the mitochondrial DNA. This makes extraction and sequencing of mitochondrial DNA easier from a technical viewpoint.

Mitochondrial DNA can sometimes be obtained from museum samples and extinct animals. Mitochondrial DNA extracted from frozen mammoths found in Siberia differed in four to five bases out of 350 from both Indian elephants and African elephants. The DNA analysis supports the three-way split proposed based on anatomical relationship. The quagga is an extinct animal, similar to the zebra. It grazed the plains of Southern Africa only a little over a hundred years ago. A pelt preserved in a

German museum has yielded muscle fragments from which DNA has been extracted and sequenced. The two gene fragments used were from the quagga mitochondrial DNA. The DNA from the quagga differed in about 5 percent of its bases from the modern zebra. The quagga and mountain zebra are estimated from this to have had a common ancestor about three million years ago.

DNA has also been successfully extracted from Egyptian mummies. Although the amounts of DNA obtained are only 5 percent or so of those from fresh, modern, human tissue, DNA sequences have been obtained from a mummy 2,400 years old. Although several thousand base pairs were sequenced, no actual human genes were identified. Since the DNA of higher animals consists mostly of non-coding sequences, this is hardly surprising. Nonetheless, the mummy DNA did contain Alu elements that are characteristic of human DNA.

The African Eve Hypothesis

Attempts to sort out human evolution from skulls and other bones led to two alternative schemes. The multi-regional model proposes that *Homo erectus* evolved gradually into *Homo sapiens* simultaneously throughout Africa, Asia and Europe. The Noah's Ark model proposes that most branches of the human family became extinct and were replaced, relatively recently, by descendants from only one local sub-group (Fig. 20.23). Although anthropologists take both theories seriously, few geneticists regard the multi-regional model as plausible. This model implies continuous genetic exchange between widespread and relatively isolated tribes over a long period of prehistory. Not surprisingly, recent molecular analysis has tended to support the Noah's Ark model.

Mitochondrial sequence analysis suggests all modern humans are derived from a small group of ancestors who lived in Africa around 100,000 years ago.

Although mitochondria evolve fast, the overall variation among people of different races is surprisingly small. Calculations based on the observed divergence and the estimated rates suggest that our common ancestor lived in Africa between 100,000 and 200,000 years ago. Since mitochondria are inherited maternally, this ancestor has been named "**African Eve**". This African origin is supported by the deeper "genetic roots" of modern-day African populations. In other words, different sub-groups of Africans branched off from each other before the other races branched off from the Africans as a whole (Fig. 20.24).

The ancestors of today's Europeans split off from their Euro-Asian forebears and wandered into Europe via the Middle East around 40,000 to 50,000 years ago (Fig. 20.25). American Indians appear to derive from two major migrations originating from mainland Asian populations. The earlier Paleo-Indians (around 30,000 years ago) populated the whole American continent, while the more recent migration (less than 10,000 years ago) produced the Na-Dene peoples who are mostly North American Indians.

Besides using mitochondrial DNA, sequences of microsatellite regions of the chromosomes have been compared among the different races. The phylogenetic results are very similar. They also give a primary African—non-African split, and if anything, they suggest an even more recent date for the common ancestor, nearer 100,000 years ago.

But what about Adam, or "**Y-guy**" as he is sometimes called by molecular biologists? The shorter human Y chromosome does not recombine with its longer partner, the X chromosome over most of its length. This allows us to follow the male lineage without complications due to recombination. For example, the *ZFY* gene on the Y chromosome is handed on from father to son and is involved in sperm maturation. The sequence data for *ZFY* suggest a split between humans and chimps about 5 million years ago and a common male ancestor for modern mankind about 250,000 years ago. However, recent data from a much larger number of genetic markers on the Y chro-

African Eve Hypothetical female human ancestor thought to have lived in Africa around 100,000–200,000 years ago
Y-guy Hypothetical male human ancestor thought to have lived in Africa around 100,000–200,000 years ago

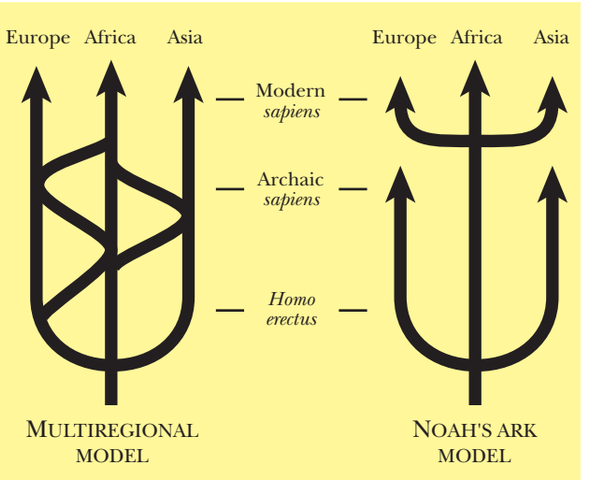


FIGURE 20.23 ***Multi-Regional and Noah's Ark Models of Human Evolution***

The multiregional model of human evolution (left) suggests that *Homo sapiens* developed from multiple interactions between several ancestral lines. The early *Homo erectus* ancestor branched and migrated from Africa to Asia and Europe. Traits developing in each branch were transmitted to the other branches implying genetic exchanges between the three branches, even though many thousands of miles separated the early ancestral groups. The Noah's Ark model (right) seems more plausible based on genetic analysis. The model suggests that modern *Homo sapiens* developed from one ancestral group in Africa. Other branches of archaic *sapiens* did develop and inhabited different regions in Europe and Asia for a while before dying out. The modern *sapiens* branch has then evolved into several branches from a relatively recent African ancestor.

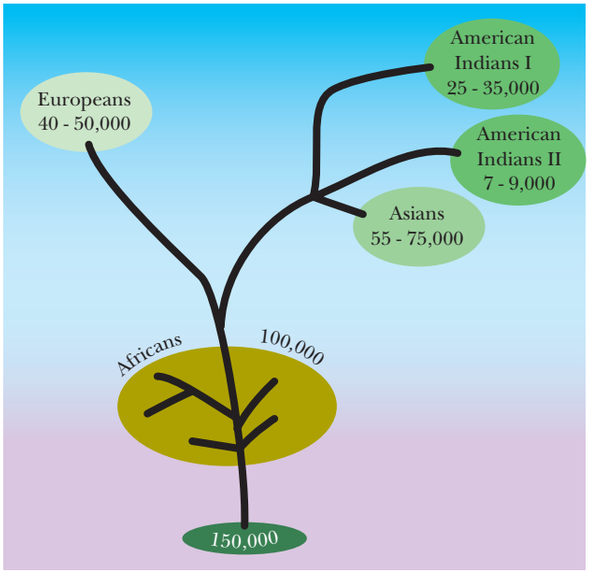


FIGURE 20.24 ***African Eve Hypothesis I—DNA***

This phylogenetic relationship was deduced by comparing mitochondrial DNA sequences from living humans. Numbers shown are estimated years before the present (BP). According to the African Eve theory, early humans developed in Africa about 150,000 years BP and diverged into many different tribal groups, most of which remained in Africa. The European and Asian races are derived from those relatively few groups of African ancestors who emigrated into Eurasia via the Middle East.

Y chromosome sequences confirm the recent African origin of humans.

mosome dates Y-guy to somewhat less than 100,000 years ago. Recent analyses of clusters of mutations on the Y chromosome in are incompatible with the multi-regional model and confirm the recent African origin of modern humans.

A final sad note concerns Neanderthal Man. Although Neanderthal Man survived to live alongside the modern races of *Homo sapiens* in Europe and the Middle East until relatively recently (approximately 30,000 years ago), sequence analysis suggests that the Neanderthals came to a dead end. Comparison of DNA sequences suggests

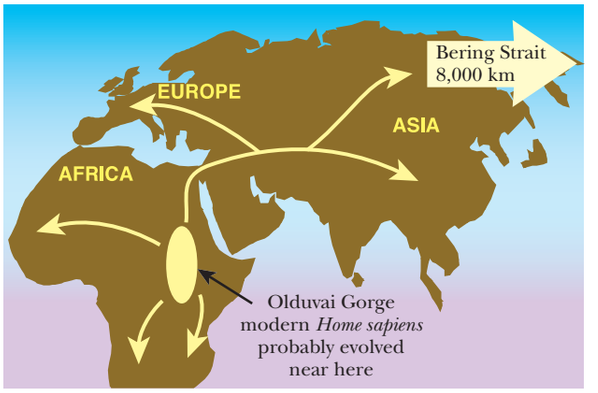


FIGURE 20.25 ***African Eve Hypothesis II—Migrations***

The divergence of the African ancestor into the modern African, European and Asian races included migration into different parts of the world. Scientists believe that modern *Homo sapiens* evolved in eastern Africa, around the Olduvai Gorge. Descendents of these early ancestors migrated to Europe and Asia as well as other areas in Africa. Descendents of some Asian groups crossed the Bering Strait to inhabit the American continent. Once isolated, these various groups evolved independently.

that the Neanderthals did not interbreed with modern Man or contribute significantly to the present-day human gene pool.

Ancient DNA from Extinct Animals

Small stretches of DNA sequence have been rescued from extinct life forms.

Apart from the occasional mummy or mammoth, DNA sequences from still living creatures are normally used to construct evolutionary schemes. However, ancient DNA extracted from the fossilized remains of extinct creatures can provide a valuable check on estimated evolutionary rates. The oldest available DNA so far successfully analyzed comes from amber. Amber is a polymerized, hardened resin produced by extinct trees that has gradually solidified to a glassy consistency over millions of years. Sometimes small animals were stuck in the resin when it oozed out of the trees and have been preserved there ever since (Fig. 20.27). Most of the trapped animals are insects, but occasionally worms, snails, and even small lizards have also been found. Amber acts as a preservative and the internal structure of individual cells from trapped insects can still be seen with an electron microscope. It has proven possible to recover DNA that is 25 to 125 million years old from some insects and some of this has been amplified by PCR and sequenced.

The largest chunks of amber are no more than 6 inches across, so larger animals such as dinosaurs cannot be preserved. Nonetheless, a few blood cells preserved in the gut of a blood-sucking insect could, in theory, provide the complete DNA sequence of a large animal. This was the scenario for Michael Crichton's high-tech thriller, Jurassic Park, where dinosaurs were resurrected by having their DNA inserted into amphibian eggs. In real life, such ancient dinosaur DNA would be severely damaged and only short segments would be readable. Nonetheless, the possibility of someday obtaining a few short fragments of some *Tyrannosaurus rex* genes is no longer a total fantasy.

Although DNA has indeed been isolated from samples that are several million years old it is so severely degraded that identification has not been possible. To date, the oldest identified animal DNA is approximately 50,000 years old and comes from mammoths preserved in the permafrost of Siberia. The permafrost has also yielded identifiable plant DNA from grasses and shrubs around 300,000–400,000 years old.

Microorganisms may also be trapped in amber and in such cases it may be possible to revive the whole organism, not merely obtain DNA samples. In particular,

Genghis Khan's Y Chromosome

Large-scale surveys have shown that about 1 in 12 men in Asia carry a variant of the Y chromosome that originated in Mongolia roughly 1,000 years ago. Around 30 natural genetic markers were surveyed in several thousand men. The markers included deletions and insertions, sequence polymorphisms and repetitive sequences. Most men carry Y chromosomes with more or less unique combinations of such DNA markers. However, about 8% of Asian males carry Y chromosomes with the same (or almost the same) combination of genetic markers. This phenomenon was not seen among men from other continents. Furthermore, the Asian men with the special "Mongol cluster" of genetic markers were found only among those populations who formed part of the Mongol Empire of Genghis Khan. For example, the "Mongol cluster" was absent from Japan and southern China, which were not incorporated into the Mongolian Empire, but was present in 15 different populations throughout the area of Mongolian domination (Fig. 20.26). In addition, although very few Pakistanis have the "Mongol cluster", about 30% of a small tribal group known as the Hazara do possess it. The Hazara are known to be of Mongolian origin and claim to be direct descendents of Genghis Khan. Since present-day Pakistan is outside Genghis Khan's area of conquest they presumably migrated to their present location later.

This particular variant has therefore been proposed to be the Y chromosome of Genghis Khan the great Mongolian conqueror. About 800 years ago the warlord Temujin united the Mongols and in 1206 assumed the title of Genghis Khan ("Lord of Lords"). The Mongols massacred many of the males and impregnated many of the women in areas they conquered. The present day distribution of Y chromosomes apparently reflects these practices. Whether this special variant of the Y chromosome was present in Genghis Khan himself or just frequent among his Mongol warriors cannot be known for certain. Nonetheless, it is more likely than not that Genghis Khan himself had this Y chromosome, as all the warriors in such tribes were usually closely related.

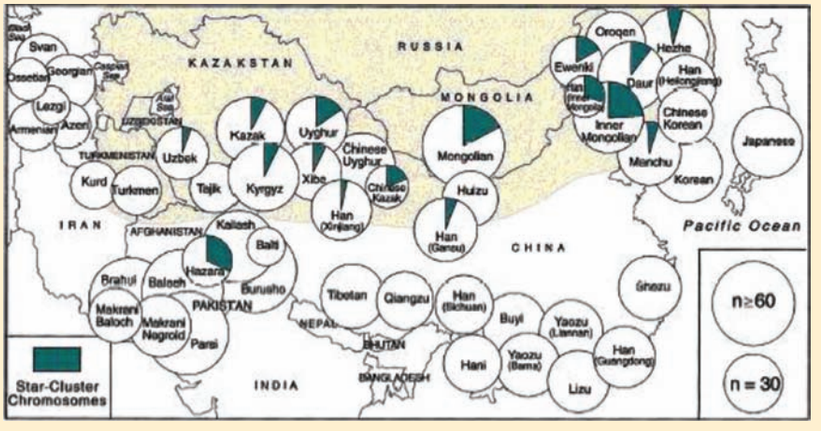


FIGURE 20.26 ***Genghis Khan's Empire and Y Chromosome***

The relative proportion of Mongol cluster chromosomes at various geographical locations is represented by the green segments in the circles. The size of the circles indicates sample size. From Zerjal and Tyler-Smith, The genetic legacy of the Mongols, *American Journal of Human Genetics* (2003) 72:717–721.

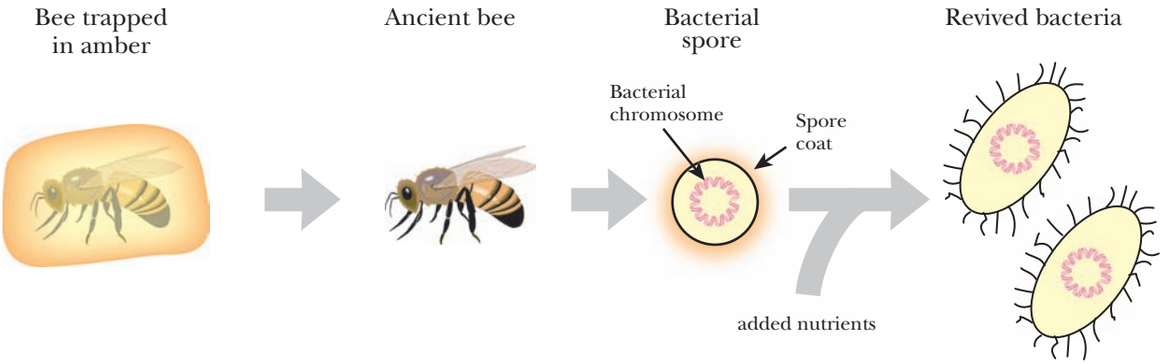


FIGURE 20.27 ***Ancient DNA Preserved in Amber***

At some point millions of years ago, a bee was trapped in sap from a tree. The sap gradually hardened and solidified into solid clear yellowish material—amber. The bee was completely preserved together with the bacterial spores that it carried. After extraction from the amber resin, occasional spores are capable of growth given the right nutrients and environmental conditions.

spores, covered by a protective coat are formed by some bacteria to survive bad conditions and may survive for extremely long periods. Some 30 million-year-old bacterial spores have been found inside bees trapped in pieces of amber. When provided with nutrients, the spores grew into bacterial colonies. These reawakened bacteria were identified as *Bacillus sphaericus*, which is found today in association with bees. DNA from the ancient *Bacillus sphaericus* was similar in sequence to its modern relative, but not identical, as it would have been if the ancient bacteria were just contaminants. More recently, spores of another bacillus species were isolated and revived from a brine inclusion within a 250-million year old salt crystal.

Evolving Sideways: Horizontal Gene Transfer

Genetic information may be passed "vertically" from an organism to its direct descendents or "horizontally" to other organisms that are not descendents.

Standard Darwinian evolution involves alterations in genetic information passed on from one generation to its descendants. However, it is also possible for genetic information to be passed "sideways" from one organism to another that is not one of its descendents or even a near relative. The term **vertical gene transfer** refers to gene transmission from the parental generation to its direct descendants. Vertical transmission thus includes gene transmission by all forms of cell division and reproduction that create a new copy of the genome, whether sexual or not. This contrasts with "**horizontal gene transfer**" (also known as "**lateral gene transfer**") in which genetic information is passed sideways, from a donor organism to another that is not its direct descendent.

For example, when antibiotic resistance genes are carried on plasmids they can be passed between unrelated types of bacteria (see Ch. 16). Since genes carried on plasmids are sometimes incorporated into the chromosome, a gene can move from the genome of one organism to an unrelated one in a couple of steps. The complete genomes of many bacteria have now been fully sequenced. Estimates using this data suggest that about 5–6% of the genes in an average prokaryotic genome have been acquired by horizontal transfer. The effects of horizontal transfer are especially noticeable in a clinical context. Both virulence factors and antibiotic resistance are commonly carried on transmissible bacterial plasmids.

horizontal gene transfer Transfer of genetic information "sideways" from one organism to another that is not directly related
lateral gene transfer Movement of genes sideways between unrelated organisms. Same as horizontal gene transfer
vertical gene transfer Transfer of genetic information from an organism to its descendents

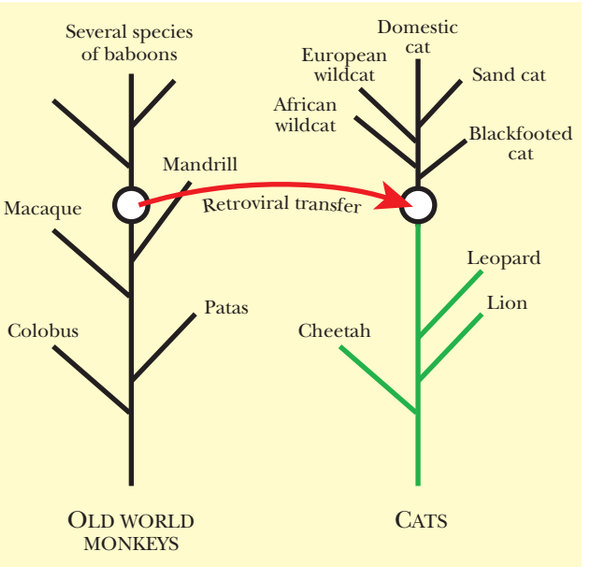


FIGURE 20.28 ***Horizontal Transfer of Type-C Virogene in Mammals***

The type-C virogene was present during evolution of old world monkeys from their common ancestor. Surprisingly, a version of this gene closely related to the one in baboons was identified in North African and European cats. Since baboons and cats are not closely related, the gene must have moved from one group to another via horizontal transfer. Further supporting the idea of horizontal transfer, the gene is not found in cats like the lion or cheetah, which developed before the North African and European cats branched off.

Horizontal gene transfer usually involves viruses, plasmids or transposons.

Such horizontal transfer may occur between members of the same species (e.g. the transfer of a plasmid between two closely related strains of *Escherichia coli*) or over major taxonomic distances (e.g. the transfer of a Ti-plasmid from bacteria to plant cells). Horizontal gene transfer over long distances depends on carriers that cross the boundaries from one species to another. Viruses, plasmids and transposons are all involved in such sideways movement of genes and have been discussed in their own chapters (see Chs. 15–17). Retroviruses, in particular, are capable of inserting themselves into the chromosomes of animals, picking up genes and moving them into another animal species.

One well-described example of horizontal transfer in animals concerns the type-C virogene shared by baboons and all other Old World monkeys. The type-C virogene was present in the common ancestor of these monkeys, about 30 million years ago, and since then has diverged in sequence just like any other normal monkey gene. Related sequences are also found in a few species of cats. Only the smaller cats of North Africa and Europe possess the baboon type-C virogene. American, Asian and Sub-Saharan African cats all lack this sequence. Therefore, the original cat ancestor did not have this type-C virogene. Furthermore, the sequence found in North African cats resembles that of baboons more closely than the sequences in monkeys closer to the ancestral stem (see Fig. 20.28). This suggests that about 5–10 million years ago a retrovirus carried the type-C virogene horizontally from the ancestor of modern baboons to the ancestor of small North African cats. The domestic European pussycat originally came from Egypt, so it also carries the type-C virogene. However, other cats that diverged more than 10 million years ago lack these sequences.

Problems in Estimating Horizontal Gene Transfer

When the human genome was sequenced, several hundred human genes were at first attributed to horizontal transfer from bacteria. However, later analyses indicated that very few of these were genuine cases of horizontal transfer (see Ch. 24). Several factors have contributed to such over-estimates of horizontal transfer, both for the human genome and in other cases.

a. Sampling Bias. Relatively few eukaryotic genomes have been fully sequenced whereas hundreds of bacterial genomes have been sequenced. Thus the absence of sequences homologous to a human gene from a handful of other eukaryotes

is insufficient evidence for an external (bacterial) origin. As more eukaryotic sequence data has become available many genes supposedly of "bacterial" origin have in fact been found in other eukaryotes.

b. The loss of homologs in related lineages may suggest that a gene originated externally to the group of organisms that retain it. As in the related case (a) above, the solution to this artifact is the collection of more sequence data from many related lineages.

c. Gene duplication followed by rapid divergence may give rise to apparently novel genes that are missing from the direct vertical ancestor of a group of organisms.

d. Intense evolutionary selection for a particular gene may result in a greatly increased rate of sequence alteration. Those genes that evolve faster than normal will tend to be misplaced when sequence comparison is used to construct evolutionary trees.

e. The ease of horizontal transfer of genetic information by plasmids, viruses, transposons under laboratory conditions is misleading. Under natural conditions there are major barriers to such movements. Furthermore, the results of horizontal transfer are often only temporary. Newly acquired genes, especially those on plasmids, transposons, etc., are easily lost. Such genes tend to be acquired in response to selection such as antibiotic resistance and, conversely, they will be lost when the original selective conditions disappear.

f. Experimental problems such as DNA contamination. Bacterial and viral parasites are associated with esentially all higher organisms and completely purifying the eukaryotic DNA is not always easy.

Many of the originally proposed examples of widespread horizontal gene transfer have been severely compromised by the above factors. However, some examples do seem to be valid. One of the most interesting is the recent finding of relatively frequent horizontal gene transfer between the mitochondrial genomes of flowering plants. The genes for certain mitochondrial ribosomal proteins have apparently been transferred from an early monocotyledonous lineage to several different dicotyledonous lineages. Examples include transfer of the *rps2* gene to kiwifruit (*Actinidia*) and the *rps11* gene to bloodroot (*Sanguinaria*).