Chapter 3.3 Structure and evolution of genomes 2

The cell 2

The nucleus 2

Genome organization 2

Mitochondria and chloroplasts 3

Figure DNA2Org 4

Chromosomes 5

Cell replication 5

Figure CellCycle 6

Haploid and diploid genomes 7

Size of genomes 8

Genome content 8

The human genome 8

Figure HuGenome1 9

Human genes 10

Repetitive sequence DNA 11

Figure HuGenome2 12

Creation and evolution of genes 13

The hemoglobin gene family 14

Human hemoglobin gene clusters 14

Figure GlobinGenes 15

Pseudogenes 16

Duplicate genes 16

Changes of gene expression during development 16

Figure GlobinKinetics 18

Sickle cell hemoglobin 19

Myoglobin, the immediate ancestor 19

Figure MHb_Aligned 20

Figure GlobinStructure 22

The antibody gene family 23

Antibodies 23

Selection of antibody producing cells 23

Figure IgCells 24

Production of new antibody genes 25

Figure IgGenes 26

The HOX gene family 27

Figure HOXofFly 29

Species 30

Sex maintains species 31

Figure Species 32

Sex generates diversity 33

Figure Sex 34

Recombination 35

Figure recombination 36

Creation of species 37

Kinetics of evolution 37

Figure LifeTree 38

Figure LifeTimeLine 40

Model organisms 41

Figure FindMatch 43

Evolution: paradigm of life 44

Chapter summary____ 46

Chapter 3.3 Structure and evolution of genomes

In this chapter we zoom out from the atomic scale to look at life using a microscope. We focus our microscope on the genome, the collection of structures containing the genetic information transmitted from parents to progeny. In all living organism but a few viruses, the genome is made up of DNA molecules. While in the previous chapter we described the details of the encoding of genetic information in DNA and RNA, the genome is a broader, biological concept since it contains all the information needed to specify one organism. But before we talk about the genome we need to know something about the object the genome is contained in and functions for.

The cell

All free living organisms, are composed of one or more cells. The cell is a lipid membrane bag that contains, and has attached to it, the molecules that enable the organism to live. Cells are typically 2 to 20 microns in diameter. The cell membrane, or wall, is impervious to the water soluble molecules in and outside the cell, which allows the intracellular environment to be defined and controlled by the activities of the cell. The cell wall is however certainly not just a passive barrier. It is covered with pores that allow selected molecules to pass in and out of the cell in a controlled manner. The components in a cell have evolved to function in a fairly narrow range of salt concentrations and pH (acidity). If the sodium, potassium, calcium, etc concentrations are outside the normal narrow range, the enzymes and other proteins can not function. In fact, as we will describe more fully later, small changes in calcium concentration are used to transmit information within the cell. Thus, even if the external environment would seem to be benign, e.g. the fluid inside our body, each cell must still be isolated from that environment. The formation of a cell wall was thus one of the earliest and most important development in the generation of life on earth.

The nucleus

About two hundred million years ago some cells developed an interior membrane enclosed bag, the nucleus, which contains the genome. These organisms are called eukaryotes (eu = true, karyo = kernel), and are represented by the kingdoms of plants, animals, fungi, and protoctista (single cell organisms). The organisms which remain without a nucleus, the prokaryotes, are bacteria.

Like the cell membrane the nuclear membrane has pores which selectively transport molecules in and out. For example, while the DNA genome is inside the nucleus, the ribosomes which translate mRNA into protein are outside. Thus the messenger DNA copied from the DNA genome must pass through pores in the nuclear membrane to get to the ribosomes. The space outside the nucleus is called the cytoplasm.

Genome organization

Viruses many have an RNA genome, e.g. HIV, a single strand or a double strand DNA genome. However, all viruses grow and reproduce only inside the cell of an organism containing a DNA genome, and information supplied by the host DNA genome is required for the virus to replicate. However, there is no obvious reason that a free living organism could not have an RNA genome, and as we will see in a later chapter, it seems likely that very ancient organisms grew and reproduced using only RNA to store genetic information.

Prokaryotes, the bacteria, have a single, circular, double stranded DNA molecule as a genome. Many, in addition, contain a smaller circular DNA molecule, called a plasmid. The plasmids often contain genes that confer drug resistance to the bacteria, and are thus of considerable medical interest. Plasmids can be easily transmitted from one bacterial cell to another, which allows traits like drug resistance to spread far more rapidly in a bacterial population than would occur if drug resistance was only transmitted to the progeny of resistant parents. The transfer of plasmids is an example of horizontal genetic exchange, as contrasted to the vertical (think of the vertical direction representing time) exchange between parents and daughter cells.

In eukaryotes the nuclear genome consists of several long double stranded DNA helixes, each coiled around specific proteins to form condensed structures, the chromosomes.

Mitochondria and chloroplasts

Eukaryotic cells also typically contain organelles with an associated DNA molecule, which is thus also part of the genome of the cell. The mitochondria, found in animal cells, carry out the final stage of oxidation of nutrients, producing most of the energy used by the organism. The small circular DNA molecule in mitochondria contain the genes for most mitochondrial proteins. This genome also codes for t-RNA molecules that, using a slightly different triplet code, make some mitochondrial proteins. However, other proteins that are found in mitochondria are coded by genes in the nucleus. Thus, the mitochondria is a genetic chimera. Plant cells contain chloroplasts, which carry out the initial steps in the conversion of the energy of light into chemical energy. Chloroplasts also contain a small genome. Both mitochondria and chloroplasts are thought to have originated from free living organisms, which became integrated within the host cells in a physical, chemical, and genetic sense to form a symbiosis.

Figure DNA2Org

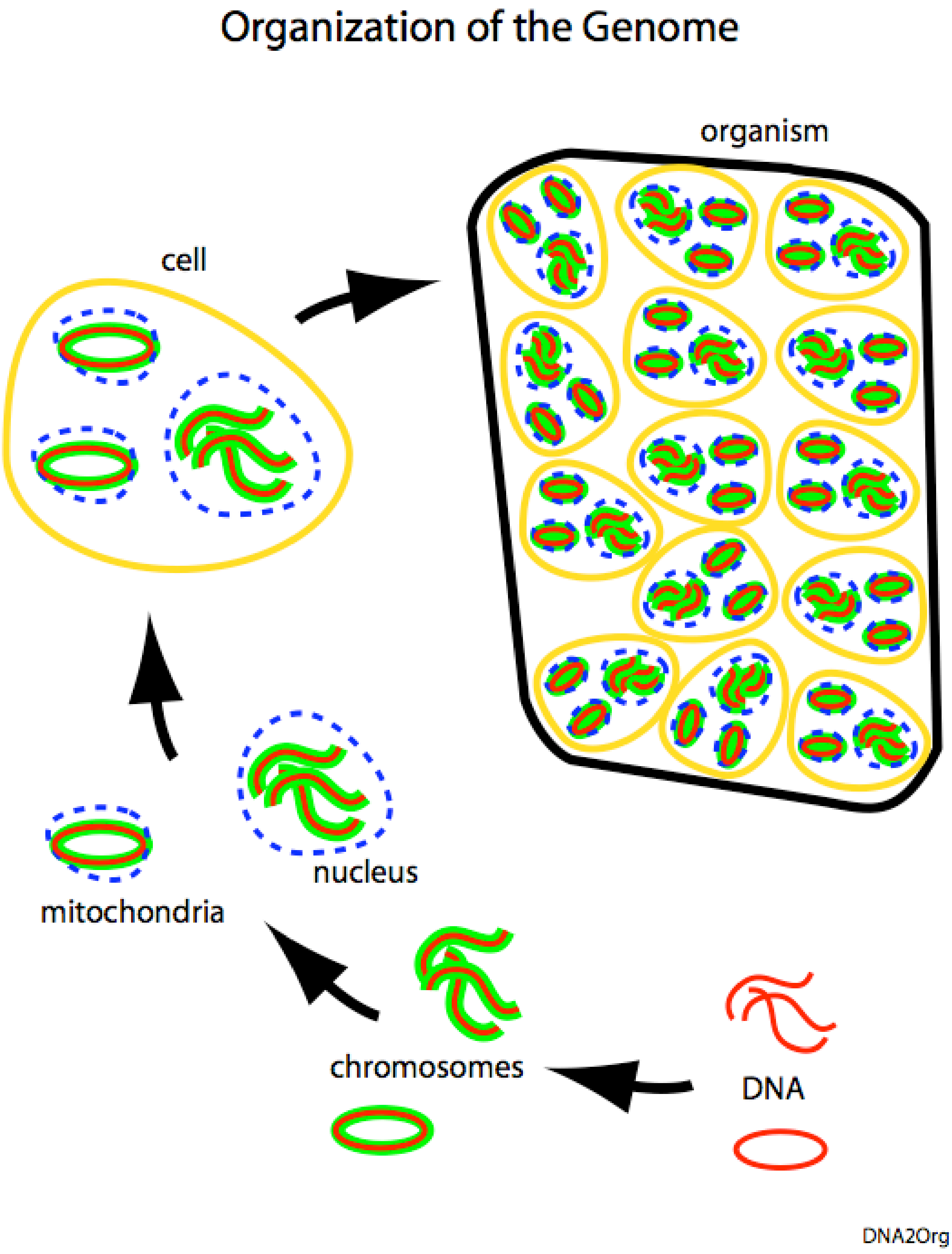


Figure DNA2Org. Individual DNA molecules are part of a hierarchy of larger and more complex structures that comprise a living organism. Each DNA double helix is coiled successively to make a compact linear structure which is covered by a specific class of proteins to make a chromosome. The several chromosomes that represent the major DNA content of the organism are enclosed in a membrane to form a nucleus. The nucleus, along with many other structures, is enclosed in another membrane to form a cell. The cell also contains organelles that contain smaller DNA molecules, e.g. animal cells contain mitochondria. Finally, a collection of cells make up the living organism.

Chromosomes

The multiple linear DNA molecules in eukaryotes are each coiled around a specialized group of proteins, the histones, and this structure is again coiled on itself. Thus, the resulting structure is much thicker and shorter than the extended length of the DNA molecule. The histone-DNA structure is called a chromosome. The density of histones and the resulting degree of compression is not uniform along the chromosome. Regions of DNA that are being actively read, i.e. used as templates to make RNA, are relatively unwound and less covered with histones compared to regions that are inactive.

Cell replication

The production of a copy of each DNA molecule, which was described in the previous chapter, is only the first stage in the process of cell replication. After DNA replication there will be several pairs of DNA molecules, or chromosomes. One copy of each pair must then move to separate regions of the cell and the cell wall must then constrict and pinch off between the two sets of DNA to form two new cells, each with one copy of the genome. When cells are constantly growing and dividing this entire process, the cell cycle, is repeated over and over.

Figure CellCycle

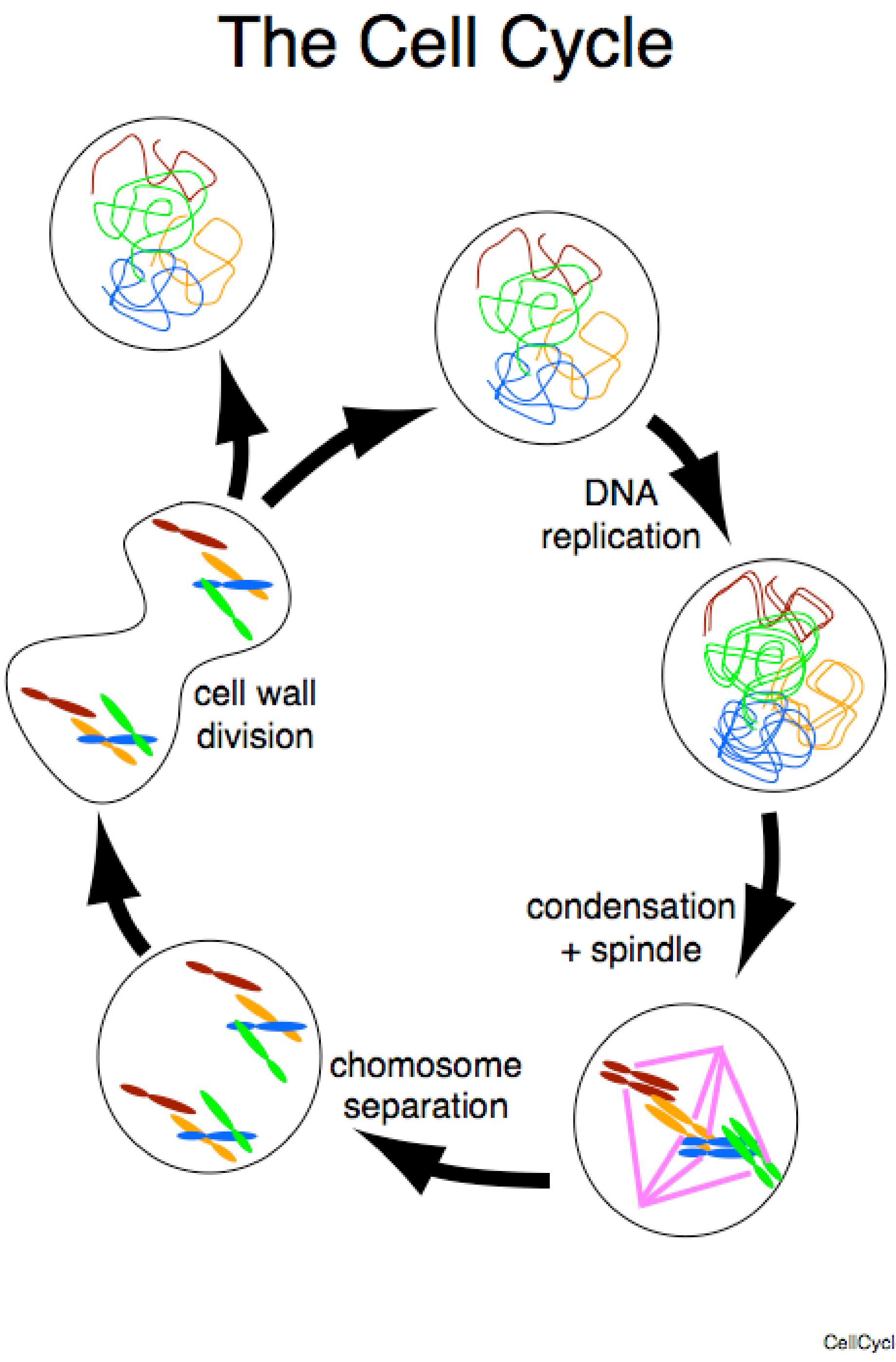


Figure CellCycle. When a cell divides the DNA and the cel must pass through a series of stages. First the DNA is replicated by the DNA polymerase complex. Then the chromosomes containing the DNA are progressively condensed as they are coiled around them selves. At the same time a spindle is made from contractile proteins. The spindle fibers connect the centromere of each chromosome pair with opposing centers and then separate them into the regions that will become daughter cells. After the cell membrane has pinched off to make two cells the DNA returns to its original state of aggregation.

As shown in Figure CellCycle, DNA molecules are only partially condensed before and during DNA duplication. At this level of condensation the DNA is accessible to the polymerase complex that copies the nucleotide sequence into RNA. However, the individual chromosomes are not resolved in the microscope because they are much thinner than the wavelength of light. However, after duplication DNA becomes more condensed as the DNA histone strands progressively wind around each other to form thicker but shorter structures. Eventually each chromosome is shorter than the diameter of the cell and thick enough that they can be easily seen in the light microscope. In this state the histones of chromosomes can be stained with dyes to reveal a series of colored bands along the chromosome. The bands of different thicknesses and separations constitute a profile of the chromosome, which can be seen using a microscope. The correlation of specific alterations in the chromosome banding patterns with inherited characteristics was one of the first indications that genetic information was carried in the chromosomes. But since chromosomes where known to contain both DNA and protein, this correlation did not, at the time, prove DNA contained the genetic information.

In parallel with DNA condensation , a new structure, the mitotic spindle, is formed by the aggregation of tubulin molecules. The spindle fibers attach to chromosomes at the centromere and extend to a spindle pole so that the two members of each chromosome pair are attached to opposing spindle poles. The tubulin fibers then contract and pull the attached chromosomes to opposing ends of the spindle. The cell wall then pinches off, the spindle dissolves and the chromosomes unwind to return to the pre division state.

Haploid and diploid genomes

Prokaryotes and some eukaryotes have a single copy of the DNA molecules that make up their genomes. This type of genome is called haploid. However, most eukaryote genomes are diploid, they contain pairs of nearly identical DNA molecules, one derived from the mother and one from the father. In a diploid organism there are thus two copies of most genes, and the nucleotide sequence of the two copies may be identical. The two genes may produce twice as much product as a single gene would, or one of the copies may be turned off, i.e. silenced.

If one copy of the gene is defective, in that it doesn't make RNA, or the RNA doesn't make protein, or the protein doesn't do the right thing, the product of the good gene may be sufficient to allow normal function. The mutation, or defect, in the bad gene is then said to be recessive. If the bad gene produces an RNA or protein that interferes with the function of the good gene copy, or if one-half of the normal product is not sufficient for normal function the mutation is said to be dominant. There are many possible ways for two gene copies to interact with each other and determine function.

Size of genomes

As might be expected, the total size of a genome is generally correlated with the complexity of the organism. However, there are wide deviations from this rule, as can be seen in the table below; one salamander species has 20 times the DNA per genome as a human.

ORGANISM	SIZE (kbp)	FORM
MS2 (bacterial virus)	4	single stranded RNA
SV40 (animal virus)	5	circular double stranded DNA
lambda (bacterial virus)	50	linear double stranded DNA
E. coli (bacterium)	4,600	circular double stranded DNA
S. cerevisiae (yeast)	13,000	16 chromosomes
C. elegans (nematode)	97,000	6 chromosomes
D. melangaster (fruit fly)	180,000	4 chromosomes
H. sapiens (human)	3,000,000	23 chromosomes
Amphiuma (salamander)	76,500,000	14 chromosomes

This lack of correlation between organism complexity and genome size, which is most extreme in the more complex organisms, is partially due to the presence of a great deal of DNA which does not code for proteins or RNA. Some non-coding sequences are binding sites for proteins that regulate transcription, but a large fraction of the non-coding DNA has no known immediate function in the organism, and thus has been called "junk" DNA. A few organisms have genomes containing duplications of long regions. The presence of "junk" DNA and large duplications imply that excess DNA is not a large burden for these organisms, and they provide material for the evolution of new genes.

Genome content

The human genome is of special interest to us, and it's as good as any other to illustrate the structure of a complex eukaryotic genome.

The human genome

Humans have 23 pairs of chromosomes, each containing one double stranded DNA chain. One chromosome of each pair has come from the father and one from the mother. The first 22 chromosome pairs are numbered in order of decreasing DNA content, with the accidental inversion of the last two, which anyway have almost identical DNA content. Thus chromosome 1 contains a DNA molecule with about 245 X 10^6 base pairs while chromosome 21 contains a DNA molecule with 47 X 10^6 base pairs. The composition of the 23 rd pair depends on the sex of the human, with a female having two X-chromosomes and a male having one X and a smaller Y chromosome.

Figure HuGenome1

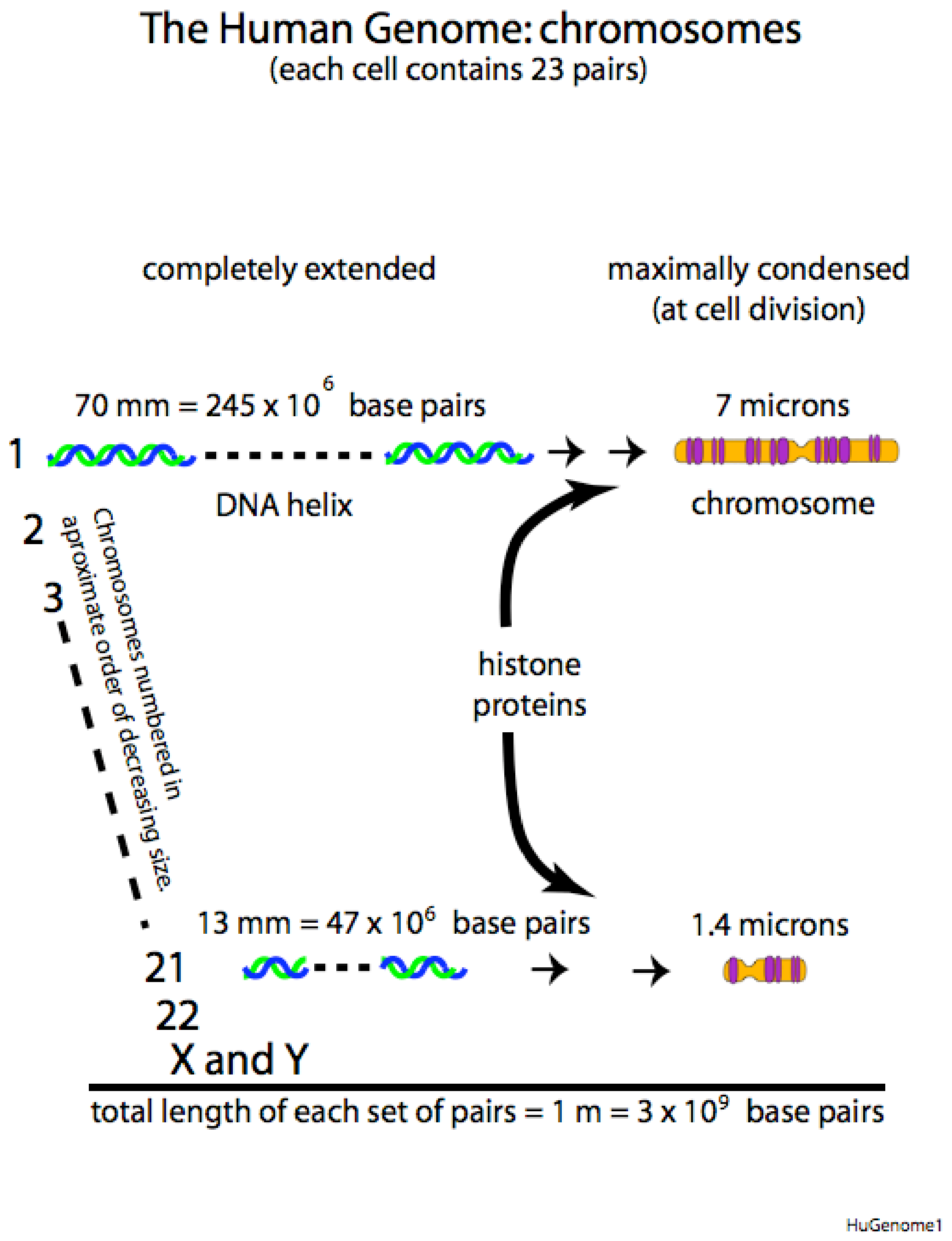

The Human Genome: chromosomes
(each cell contains 23 pairs)
completely extended
maximally condensed
(at cell division)
70 mm = 245 x 10⁶ base pairs
7 microns
1
DNA helix
chromosome
2
3
Chromosomes numbered in aproximate order of decreasing size.
histone
proteins
13 mm = 47 x 10⁶ base pairs
1.4 microns
21
22
X and Y
total length of each set of pairs = 1 m = 3 x 10⁹ base pairs
HuGenome1

Figure HuGenome1. Most human cells contain 23 pairs of chromosomes (sperm and egg cells contain only one copy of the 23 chromosomes and mature red blood cells contain no DNA).

The DNA in each chromosome is tightly coiled around histone proteins so that the length of the chromosome is much less than the extended length of the DNA.

The DNA in chromosome 1 would stretch out to 245 X 10^6 X 0.34 nm or 83 mm, while the chromosome actually has a length of less than 8 μ; a compression of more than 10^4 fold. The factor by which the DNA is compressed by association with histones varies greatly during the cell cycle, with it being the most compact during cell division, when microscopic observations can be most easily made. Compression also varies along the DNA in inverse proportion to the extent to which the region is being transcribed (read) by RNA polymerase. Variation in kind and density of histones along the chromosome creates dramatic visible banding patterns after the chromosomes are stained with dyes. Microscopic characterization of these bands allowed genetic analysis of chromosomal deletions and rearrangements before it was possible to directly determine the nucleotide sequence of DNA. In this way, damage to specific regions of chromosomes was associated with specific diseases, e.g. Down's syndrome, leukemia.

The segment of DNA that codes for one protein (or one RNA molecule if that is the end product), is called a gene. In a human, there are approximately 30,000 genes. However, different cells in our body, for example those that make up different organs, must contain very different collections of proteins. Since all cells in our body contain a complete copy of the DNA genome, the different protein compositions must result from different patterns of gene transcription or translation. Most of the difference is due to variation in transcription.

Only about 24 percent of the human DNA represents genes that are copied to make RNA. After transcription, 95 percent of the RNA is removed and degraded, so that only 5 percent is transported from the nucleus and used as RNA or translated into protein. Thus, ignoring the t-RNA and ribosomal RNA genes, which are a small percent of the transcripts, only 1 percent of the genome appears to be used to code the genetic information of the human.

Human genes

Our knowledge of human genes is extensive, but far from complete. There are many hundreds of human proteins and non-messenger RNA species that have been isolated and purified so that we know at least a partial amino acid or nucleotide sequence and a function for the protein or RNA.

We also know (as of June 2003) 99 percent of the nucleotide sequence of the human genome. Applying rules derived from the genes of well characterized proteins we can predict the location of genes in the entire genome sequence. However, as described previously, the code for determining location of the introns, which are removed from m-RNA before it is translated, is known to only a 50 - 80 percent confidence level. In addition, some segments of DNA are transcribed into more than one m-RNA species, in a process called alternative splicing. The rules for splicing are only partially known.

However, we can apply our imperfect rules to the nucleotide sequence of the entire human genome and obtain the nucleotide and amino acid sequences of predicted gene products. Since most genes are members of a family with related sequences and functions, many predicted gene products can be associated with a family and their function estimated. In most other cases we can recognize at least segments or domains that are similar to sequences from known RNA or protein molecules. These domains can then be associated with functions, such as binding ATP, or becoming embedded in the lipid membrane of the cell wall. Thus having the nucleotide sequence of the entire genome is a little like having a dictionary for a strange language in which only a small fraction of the words have definitions, but most of the other words contain syllables which at least suggest a meaning.

With all these caveats, how many genes do humans have and what do they do? A reasonable estimate for the number is 30,000, which is certainly good to within a factor of 2 (we think). The number of genes in the human chromosome is about that predicted in the genomes of other mammals. Humans may be better at solving differential equations or writing symphonies than other mammals, but these abilities are not obvious from the nucleotide sequences of the genome.

The following table lists the functional class of the most common genes.

	percent	class
	7.5	nucleic acid enzymes: alter nucleic acids
	6.0	transcription factors
	5.0	receptor: bind effectors at cell surface
	4.0	hydrolase: catylize hydrolysis, e.g. proteases
	3.2	regulatory molecules: control rates of reactions
	2.9	protooncogenes: control cell growth
	2.8	kinase: catalize phosporalation by ATP, GTP
	2.8	cytoskeletal: structural framework of cell
	2.1	oxidoreductase: catylize oxidations and reductions
	2.0	transferase: moves a chemical group
	1.9	cell adhesion
	1.7	transporter: moves molecules in or out of cells
	1.4	extracellular matrix: molecules in space between cells
	1.3	ion channel: move ions through cell membranes
	1.2	signaling molecules: part of signaling chain
	1.2	motor: enable movement of cells and components
	1.1	intracellular transporter: moves molecules inside cells

Repetitive sequence DNA

Much of the non-coding DNA in the human genome is a large number of relatively short segments that have very similar nucleotide sequences. One class of these repetitive segments is the SINES, Short Interspersed Elements. SINES, which comprise about 10 percent of the human genome, and are only found in primates, are typically 300 nucleotides long. Most of the SINES contain a nucleotide sequence that can be cut by the Alu restriction nuclease, and thus the SINES are also known as Alu sequences.

Figure HuGenome2

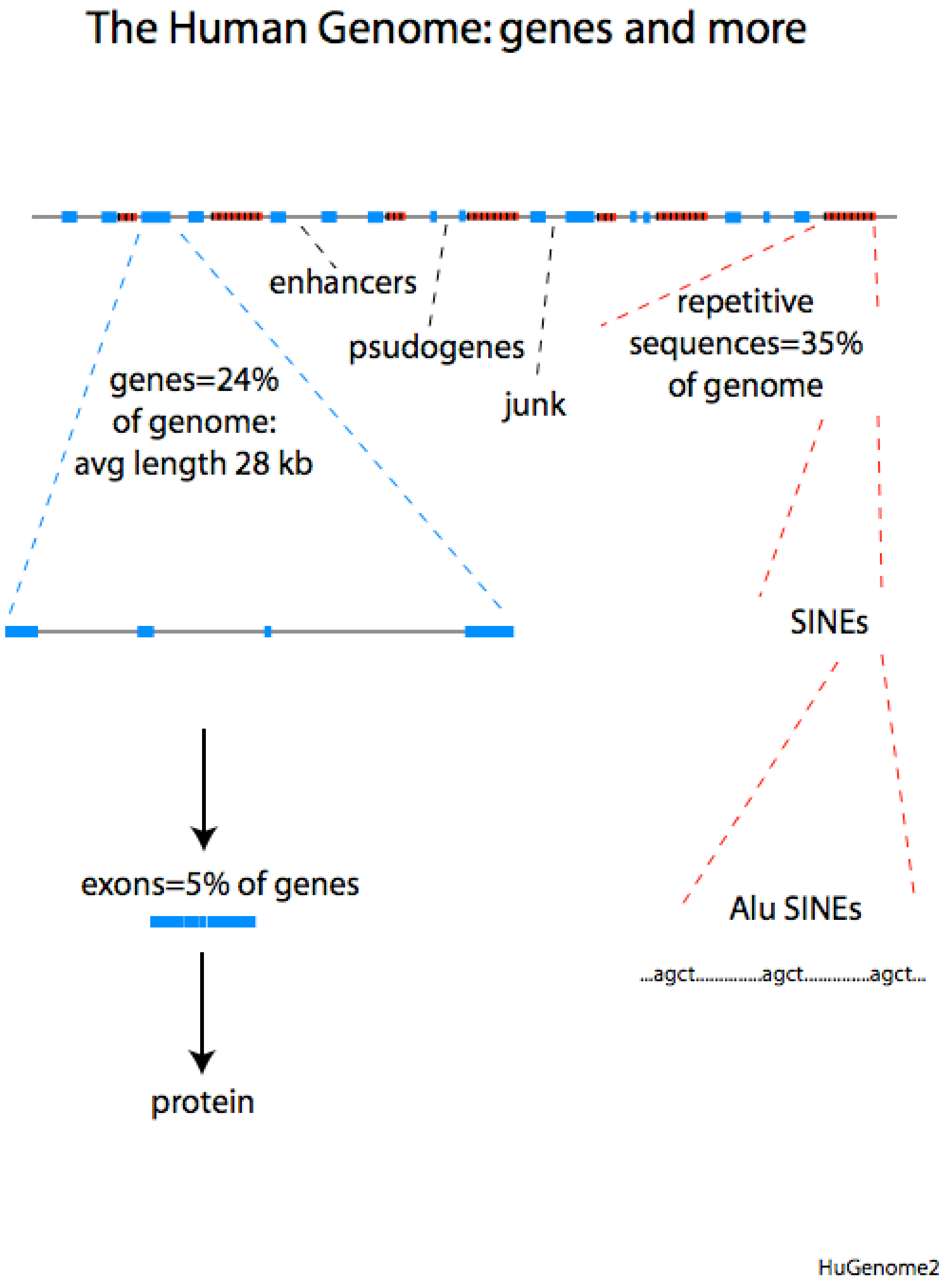

The Human Genome: genes and more
enhancers
psudogenes
junk
repetitive sequences=35% of genome
genes=24% of genome: avg length 28 kb
SINEs
exons=5% of genes
Alu SINEs
...agct.............agct.............agct...
protein
HuGenome2

Figure HuGenome2. In addition to genes, much of the human genome is taken up by repetitive sequences that are believed to originate from viral genomes that were integrated into the genome. Enhancer segments bind to the enzyme complex that makes RNA, and pseudogenes are sequences that were once genes, but no longer are copied into RNA. There is still more DNA that we are clueless about, so it's "junk".

The SINES are mobile elements; they appear to move around in the genome. Specifically they are retrotransposons, which means that when they are occasionally copied into RNA, the RNA is then copied by the enzyme "reverse transcriptase" to make a DNA fragment which is then inserted back into the genome at a different location.

We can guess that the Alu sequences play a role in the evolution of new genes. Some Alu sequences are found in the exons of genes, and some of these Alu sequences have a sequence that "tricks" the splicing machine that normally removes exons.

Creation and evolution of genes

An active gene can evolve in successive generations by gradual change. The changes from generation to generation must be fairly conservative because after each change the gene must produce a product that works, preferably at least as well as the product of the previous version of the gene.

If a copy of a gene is inserted into a new location on the genome the gene is said to be duplicated (this is very different from just making a duplicate of the entire genome, which of course also makes duplicates of all genes). When a gene is duplicated a more rapid and radical evolution is possible. One gene copy can play the conservative role of carrying out the original function while the other can evolve to perform the function better. In this case the conservative copy may be eventually lost. In other situations the evolving copy may develop a new function. In this case the conservative gene remains, and the number of functional genes in the genome has increased[1]. For many organisms we have evidence that the entire genome was duplicated, sometimes more than once, in the evolution of the present form of the genome.

Since the duplication of a gene generally doubles the rate of generation of gene product, the duplication can be of immediate benefit to an organism that needs more gene product. While there are many mechanisms for increasing the transcription of genes, there is an upper limit that defined by the size of the RNA polymerase and the rate at which it can move along the gene. Above this limit only more gene copies give more RNA. The multiple ribosomal RNA genes are a good example of the use of this method.

If the gene copy does not evolve into a useful new gene, and an increased production of gene product is not needed, the gene will still evolve but there is selection for loss of the ability to be translated into RNA, thus saving the energetic cost of making a useless product. Thus there is an equilibrium between production and deletion of genes as evidenced by the significant number of non-functional genes,

[1] Thus genes aren't made from scratch, they are altered copies of other, earlier genes. This makes the creation of complex genes easier to understand. The creation of genes could be seen as analogous to the (apparently) mysterious appearance of bacteria in meat, wine, and other foods when they are exposed to the air. In a series of investigations starting in the 1860s and Pasteur demonstrated that the bacteria were always there, in the air and everywhere. Both genes and bacteria don't arise by "spontaneous generation".

pseudogenes, that can be seen in the human genome. Of course there is no way to sure way to predict if a pseudogene's fate is to be deleted or is to become a very important new gene. Pseudogenes may be called junk, but they still represent raw material for evolution of new genes. In complex organisms, such as ourselves, there seems to be a great deal of "junk" DNA, and so the extra DNA must be not a heavy burden.

Since gene duplication and evolution is the mechanism for production of new genes, it is not surprising that most genes can be grouped into families. However, after multiple rounds of duplication and subsequent evolution, the differences between the members of the family become progressively greater. Thus the families that are fairly obvious to us probably represent the distal branches in the tree of gene evolution.

The hemoglobin gene family

An important function of our blood is to supply oxygen to all the tissues in the body. Oxygen is obtained from air in the lungs and is then bound by hemoglobin inside red blood cells. As the blood circulates throughout our body the oxygen is released from hemoglobin and diffuses into cells in the surrounding tissue where it reacts, in a series of reactions, with the food we eat to supply energy. The oxygen is actually bound to an iron molecule that is bound to a porphyrin which is linked to the globin protein. It is the iron-porphyrin complex which gives blood its red color.

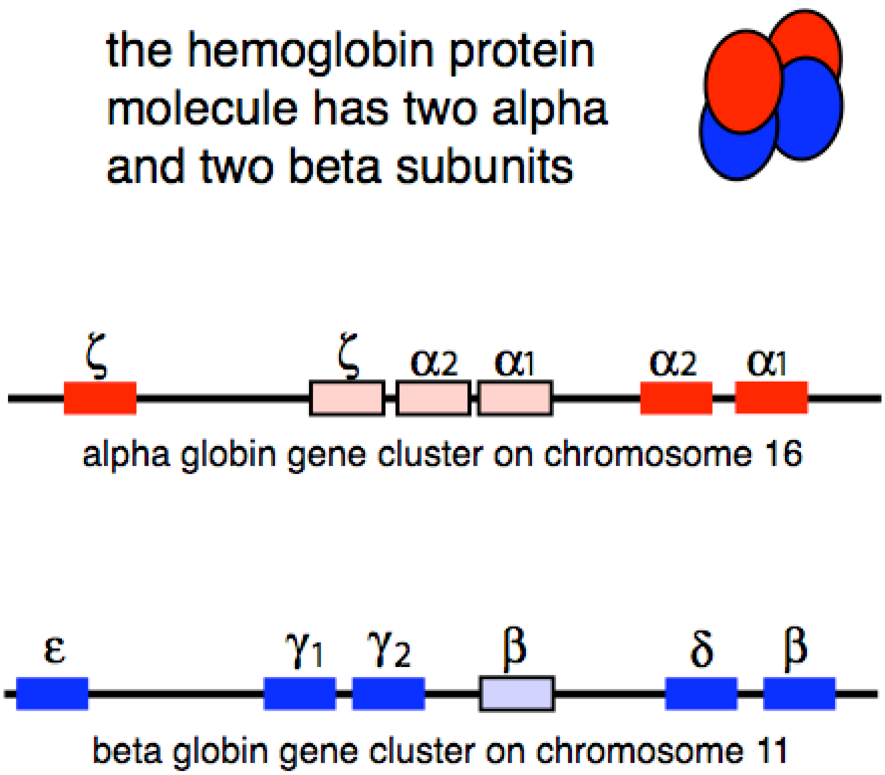
Each hemoglobin protein molecule contains four polypeptide chains, two alpha type and two beta type. The alpha and beta chains are similar in length and amino acid sequence; each chain contains an iron-porphin group and can bind one oxygen molecule. The functional advantage of this hybrid hemoglobin molecule with four chains, in contrast to a single chain hemoglobin, is discussed in the next chapter.

Human hemoglobin gene clusters

A cluster of beta type genes on chromosome 11 and a cluster of alpha type genes on chromosome 16 are diagrammed in Figure GlobinGene, (the Greek notation for the genes is retained here so you can consult the scientific literature if you want to learn more about this system). Now we have a second complication, not only are there multiple chains per hemoglobin molecule, but there are multiple genes for each type of polypeptide chain. However, study of these multiple genes reveals several important general characteristics of the genome.

Figure GlobinGenes

Hemoglobin genes



the hemoglobin protein molecule has two alpha and two beta subunits

alpha globin gene cluster on chromosome 16

beta globin gene cluster on chromosome 11

GlobinGenes

Figure GlobinGenes. Human genes for the globins form clusters on chromosomes 11 and 16. Some of these "genes" are not expressed, and are thus designated pseudogenes, other variants are expressed in embryonic development.

Pseudogenes

The lightly colored genes, one in the beta cluster and three in the alpha cluster, are pseudogenes; genes that produce no messenger RNA and thus of course no protein. These genes arose by duplication, just as all the other genes did. However, after duplication, the promoter region, or some other region essential for production of RNA, was lost or mangled so that the genes no longer function. Pseudogenes are found throughout the genome, not just in the globin gene clusters. The fact that they are common implies that there must be only a modest penalty in fitness for excess DNA, at least in organisms of our complexity. This would be consistent with the fact that only a small fraction of our genome is occupied by functional genes. During the evolution of genomes there is an equilibrium between duplication and deletion events. Duplication provides genes that can be modified without stopping the machine and deletion cleans things up (a process called garbage collection in the computer software world).

Duplicate genes

The functional genes alpha1 ($\alpha 1$) and alpha2 ($\alpha 2$) produce proteins with identical amino acid sequences. Thus these must be genes that have been duplicated recently and haven't had time to mutate. However, the mere duplication of genes can confer an advantage to the organism if a large amount of the gene product, RNA or protein, is required for optimal function. The amount of RNA or protein that one gene can produce is limited by certain kinetic and spatial restrictions, and a common method to bypass these restrictions is use of multiple genes. This technique is used for the transfer and ribosomal RNA species and ribosomal proteins for example. It seems unlikely that duplication of alpha globin genes is necessary for that purpose since one gene is sufficient for production of all the beta globin protein, which is required in equal amount.

The functional genes gamma-1 ($\gamma 1$) and gamma-2 ($\gamma 2$) produce protein chains different in only one amino acid, and this change has no known effect on the function of the protein. Thus this gene pair appears to fall into the same class as the alpha1 and alpha2 genes. However, we do not know for a certainty that these two proteins function identically in all environments and situations. It is always possible that one of the gamma proteins confers an advantage to the organism during, for example, the course of a disease. This proposal may seem unlikely, but there are many examples of mutant proteins that have been shown to confer a resistance to disease, thus, a functional difference in the gamma proteins must be considered a possibility.

Changes of gene expression during development

We've taken care of pseudogenes and duplicates, but what about the other four genes in these clusters that are not alpha and beta? These genes code for hemoglobins that are expressed during the development of the embryo in the uterus. Since red blood cells have life times on the order of weeks, after which they are degraded, the globin composition in red blood cells can be altered by changes in the rates of production of globins by the globin genes. The time course of the relative amounts of globins types in red blood cells in the fetus, embryo, and adult are plotted in Figure GlobinKinetics. The initial source of hemoglobin comes from the epsilon and zeta genes, but they become inactive within a month. The alpha genes are then turned on, and the beta type globin is supplied by the gamma genes. After birth the gamma genes gradually become inactive and the adult beta genes are turned on. The beta type globin produced by the delta gene gradually increases, but is always a minor part of the beta type globin. The unique property of the delta globin, if there is any, is not known.

Figure GlobinKinetics

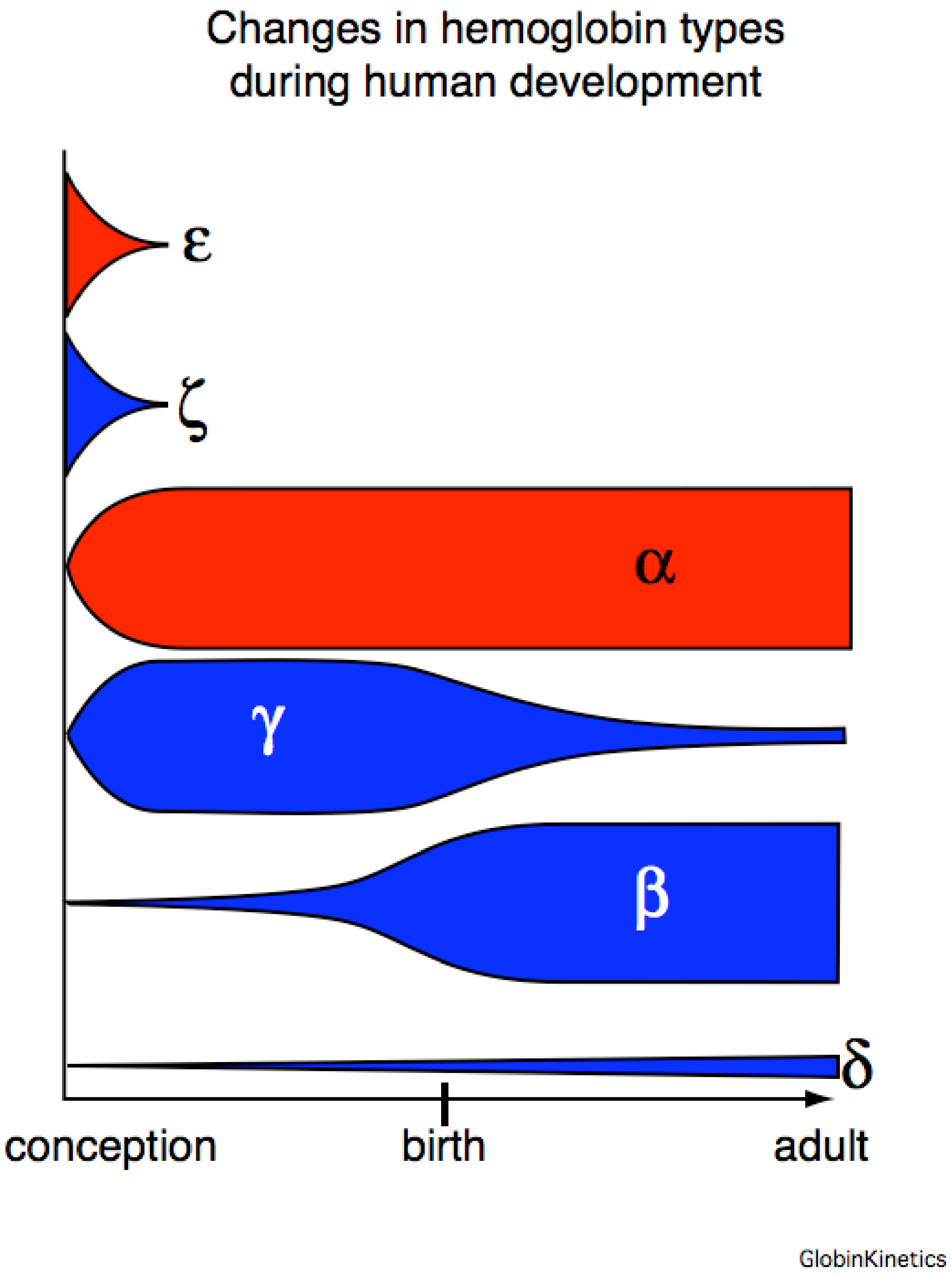


Figure GlobinKinetics. Some globin genes are only expressed in the embryo, others are expressed after birth. and others are expressed throughout the life of the individual.

Sickle cell hemoglobin

Many mutant human hemoglobins have been identified, but the most studied is the sickle cell mutation, which is the replacement of glutamic acid by valine as the sixth amino acid in the beta globin chain. This single change produces a protein that will aggregate at low oxygen concentrations and distort the red blood cell into an abnormal sickle shape, which causes the cells to clump in the capillaries, particularly in the joints of the large bones. These cell clumps produce intense pain and the sickled cells are selectively removed and destroyed by the endo-reticular system, causing anemia. A person with both normal and sickle cell hemoglobin, i.e. heterozygous for the sickle cell mutation, has little or no symptoms of disease. However, if all beta hemoglobin contains the sickle cell mutation, i.e. the individual is homozygous for the sickle cell mutation, there are frequent but erratic episodes of disease, which can be incapacitating and sometimes fatal.

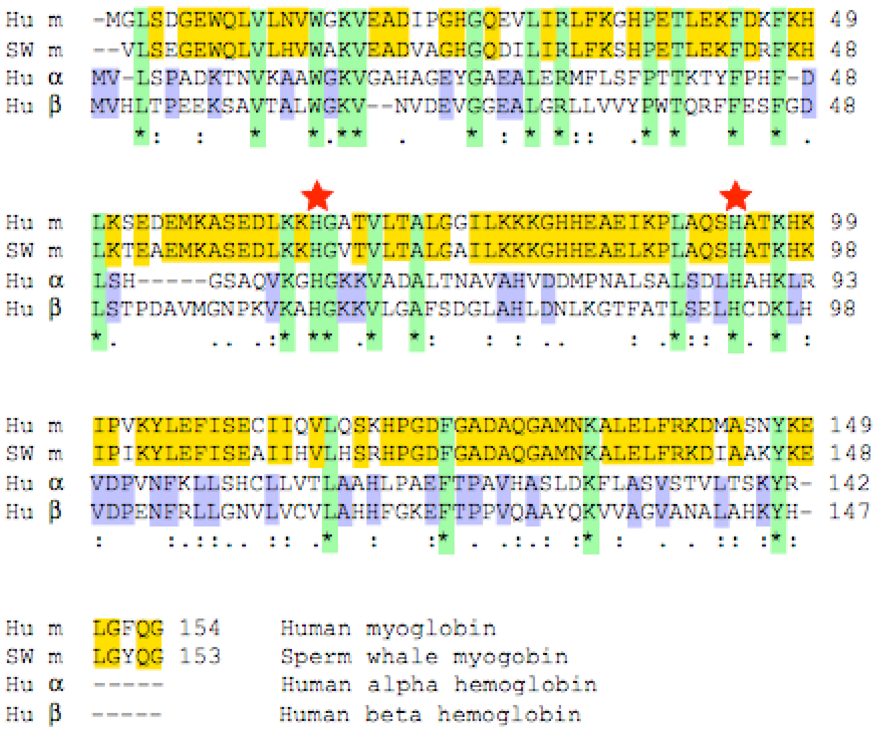
Why is this mutation common in some populations? Why is it not selected against so that it disappears or at least is very rare? The clue is in its uneven geographical distribution. It is most frequently found in people that live in African and Asian locations in which malaria is common; the sickle cell mutation confers resistance to malaria! The presence of the malarial parasite in a red blood cell containing sickle cell hemoglobin increases the probability that the cell will deform and thus be removed from the circulation and destroyed along with the parasite. Thus a mutation that is debilitating when two copies are present also conveys resistance to malaria when only one copy is present. The frequency of the sickle cell mutation in a population is thus an equilibrium between the negative effect on reproduction by sickle cell disease when the mutation is present in both chromosomes and the negative effect on reproduction by malaria when the mutation is absent from both chromosomes.

Myoglobin, the immediate ancestor

The hemoglobin genes have evolved by a process of gene duplication and evolution. The likely ancestor gene coded for another oxygen binding protein that contained only one polypeptide chain, and thus was neither an alpha or beta type. This ancestor gene, or more exactly the progeny of that gene, is still in our genome. It produces myoglobin, which functions to store oxygen in muscle tissue, not to transport it in blood. The amino acid sequence of myoglobin differs more from the hemoglobins than the sequence of the hemoglobins differ from each other. However, the three dimensional shape of myoglobin is quite similar to the hemoglobins.

Figure MHb_Aligned

Sequence alignment of four globin protein chains



```
Hu m  -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH 49
SW m  --VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKH 48
Hu α  MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D 48
Hu β  MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD 48
         *:  :   *   *.**  .    * : *::  .* *   *  * .

Hu m  LKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHK 99
SW m  LKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHK 98
Hu α  LSH-----GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR 93
Hu β  LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH 98
      *.      .. .:* ** .*  *:   : : ..    : .*:: *. * :

Hu m  IPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKE 149
SW m  IPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKE 148
Hu α  VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR- 142
Hu β  VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH- 147
      :    :.::.. :: .*  :   :* . .:.: :* :  . . :: :*:

Hu m  LGFQG 154     Human myoglobin
SW m  LGYQG 153     Sperm whale myogobin
Hu α  -----         Human alpha hemoglobin
Hu β  -----         Human beta hemoglobin
```

MHb_Aligned

Figure MHb_Aligned. The amino acid sequence of human and sperm whale myoglobin (Hu m and SW m), and the human alpha and beta chains of hemoglobin (Hu α and Hu β), show marked similarity.

Yellow zones indicate identity between myoglobins;

Green zones indicate identity between myoglobins and hemoglobins;

Purple zones indicate identity between the two chains of human hemoglobin.

It is thought that the gene duplication that resulted in the creation of the multi-peptide hemoglobins occurred when the ancestors of fish containing bones diverged from the ancestors of the sharks, about 400 million years ago. At that time there were no animals remotely resembling humans on the earth; animals had not yet migrated onto the land. Thus, the globin gene family represents old echoes of animal evolution; some are molecular fossils and some have new functional properties.

Whale myoglobin has been a favorite for study by biochemists for many decades. The first three dimensional protein structure was determined using x-ray diffraction of Sperm whale myoglobin crystals. Whales and humans had a much more recent common ancestor than the divergence of sharks and bony fish. The whale is after all not a fish, but a mammal, as we are. The ancestor of the whales walked the earth and were probably the ancestors of cows. The animals that would evolve into whales then returned to the ocean where the limbs that allowed them to walk on the earth gradually changed into the fins that allow them to swim efficiently. The amino acid sequence of whale myoglobin is quite close to the sequence of human myoglobin, as is seen in Figure MHb_Aligned, and the three dimensional structure of whale myoglobin is similar to the subunits of human hemoglobin as seen in Figure GlobinStructure.

Figure GlobinStructure

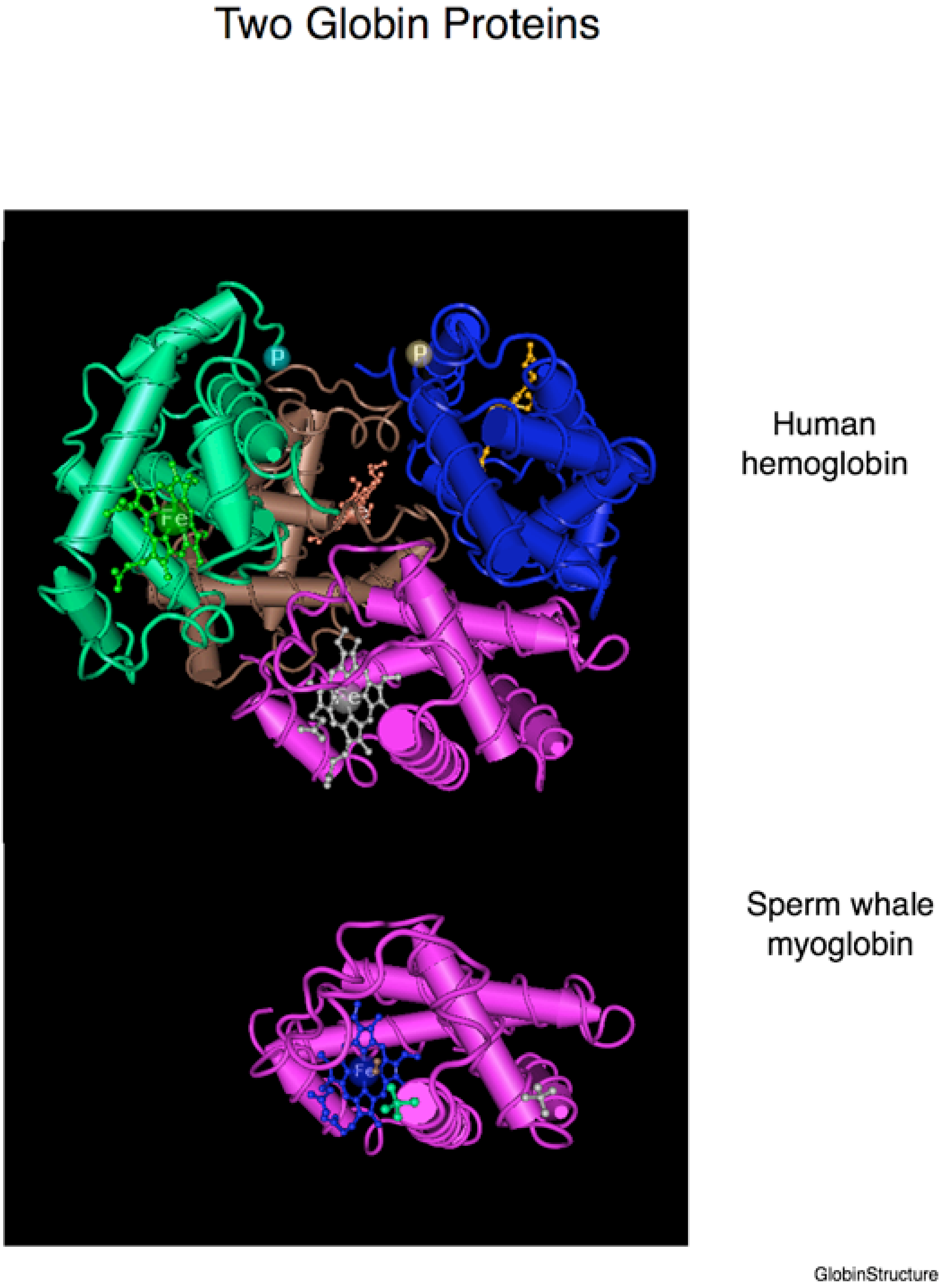


Figure GlobinStructure. In this symbolic three dimensional representation of the hemoglobin and myoglobin proteins the polypeptide chain is represented by a smooth strand. The pink and brown segments of human hemoglobin are the alpha chains, while the blue and green segments are the beta chains. Regions that have the alpha helix secondary structure are indicated by solid cylinders with an arrow head indicating N to C terminal polarity. The porphyrin ring groups, with a central iron (Fe) atom are also indicated.

The antibody gene family

Animals have an immune system which protects them from attack by foreign organisms. The antibody proteins, which bind specifically to foreign molecules, are an important part of this immune system. The production of antibodies is a dramatic and well studied example of evolution that occurs constantly in our bodies over a time span of days; you don't need to dig up fossils to see evolution in this system. The production of antibodies demonstrates that the faithful storage and transmission of information is but one job a living organism must accomplish; it must also create new information to meet new challenges. It makes new information by modifying old information until it does the job.

Antibodies

Antibodies are proteins belonging to several discrete classes, with molecular weights in the range of 100,000 - 500,000 Daltons; they are big proteins. Specific antibodies bind to specific molecules; proteins, DNA, lipids, carbohydrates, etc. A sizable fraction of the proteins in our blood are a diverse collection of antibodies with a very wide spectrum of binding specificities. However, they do have one property in common; they do not bind to any of the molecules that we normally have in our blood or tissues, our “self” molecules.

The function of antibodies is to bind to foreign material, called an antigen, that gains entrance to our body. By binding to the foreign material the antibody facilitates removal or inactivation by other members of the immune system. However, in order to bind tightly to the foreign material, a portion of the antibody must have a shape and charge distribution that is complimentary to a portion of the foreign molecule. At the level of specificity that is required, there needs to be literally millions of different antibodies, each one that binds to a molecular structure that is only defined by the fact that it is not normally present in the body; by definition a structure that the body has never seen before. We know that all proteins are specified by genes that are inherited from parent to child, but there aren't a million genes in the entire human genome, much less a million that could be set aside to code for antibodies. Common sense would thus suggest that antibodies must be formed by some sort of molding process, i.e. the antibodies must acquire their functional shape by folding around the foreign antigen. In fact, in the distant past (decades ago) papers were published that purported to demonstrate just such a molding process.

Selection of antibody producing cells

Just as common sense can trick us into thinking that the earth is flat, it misleads us here. In fact, the millions of different species of antibodies that are needed to bind to foreign antigens are created in cells containing random assemblies of a collection of antibody gene fragments.

Figure IgCells

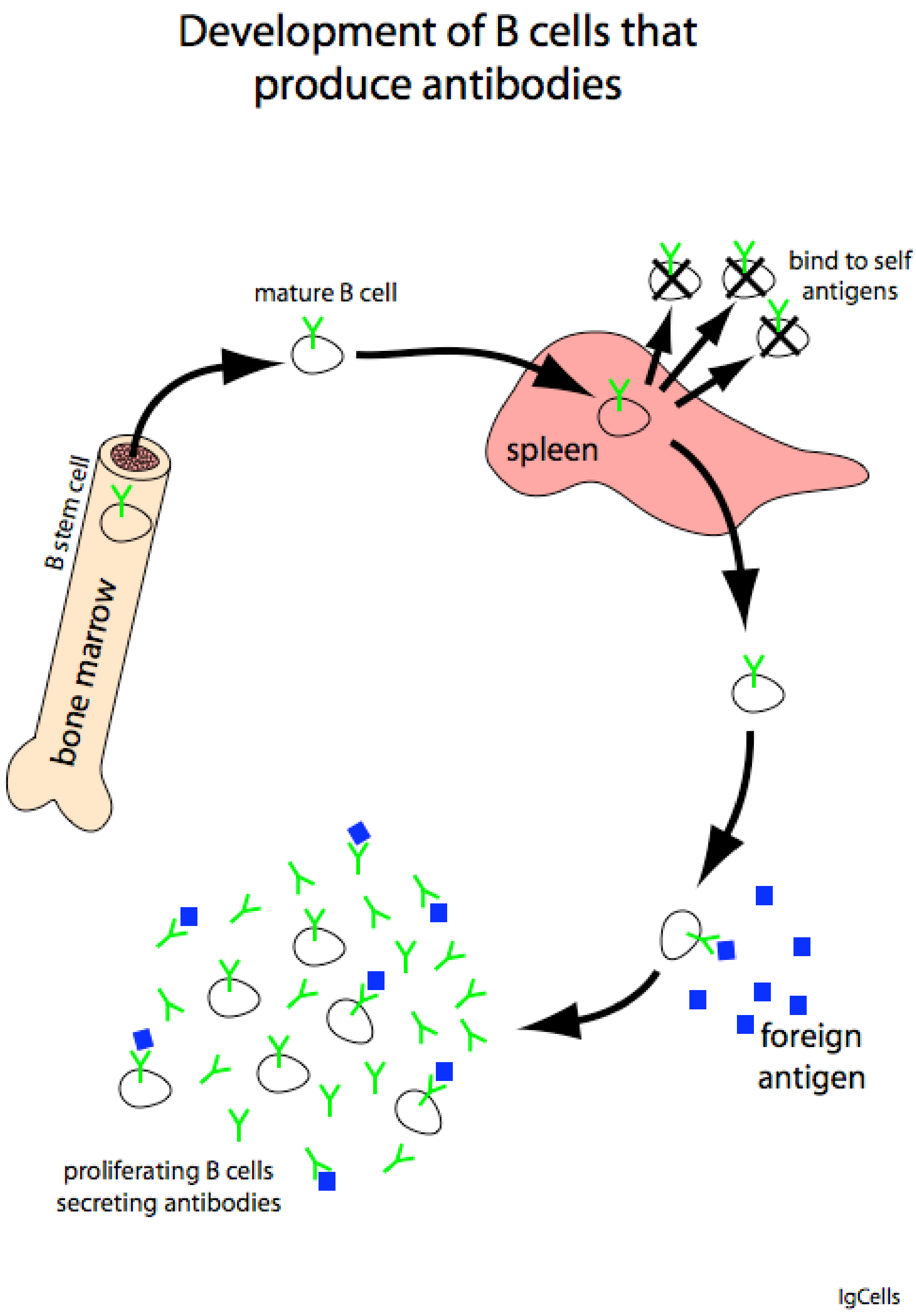

Development of B cells that produce antibodies
mature B cell
bind to self antigens
spleen
B stem cell
bone marrow
foreign antigen
proliferating B cells secreting antibodies
IgCells

Figure IgCells. The cells that produce antibodies (B cells) are produced in the bone marrow and then migrate to the spleen. If the cells produce antibodies that bind to molecules normally present in the body (self-antigens) they are killed by exposure to these antigens. The surviving

B cells then migrate into the body and are free to interact with foreign antigens. Foreign antigens which bind to the surface of a B cell cause it to proliferate and produce more antibodies.

The evolution of an antibody producing cell, a B cell, is seen in Figure IgCells. Stem cells in the bone marrow grow and divide constantly to produce new B cells. Each of these new B cells produces one species of antibody which has a unique amino acid sequence and binds to a small group of molecular shapes. The new B cells are transported to the spleen where they have the opportunity to bind to the proteins and other molecules that are normally in the individual. However, B cells which bind to something here die. This stage of B cell evolution thus eliminates cells that react with "self" antigens. Failure to remove B cells which bind to normal molecules in the body results in auto-immune diseases.

The surviving B cells circulate in blood and tissue where they may come in contact with new, foreign antigen. Now the binding of antigen to antibody on the surface of the B cell stimulates the cell to grow and divide to produce potentially millions of progeny cells, all producing the same antibody. Antigen binding also stimulates the synthesis of an altered antibody molecule which is secreted into the blood and body fluid. The binding of these free antibody molecules to foreign antigen results in a number of processes, most of which involve other components of the immune system, which remove or inactivate the antigen and thus protect the individual.

Production of new antibody genes

We have just described the selection of cells which produce antibodies that bind to foreign antigen, but to have evolution you also need a diverse population for that selection to act upon. You might propose that B cells have a very error prone DNA replication system, perhaps due to an alternative DNA polymerase which made more frequent mistakes as new B cells were produced by the stem cells, or perhaps a defective or even missing error correction system. This would indeed result in a increased number of new antibody molecules, but many of them would not function as an antibody with any antigen. In addition this non-specific mechanism for increasing diversity would also produce altered copies of all the other genes in the B cell. The sum of all these genetic "experiments" would be B cells that were either sick or dead.

Recombination between similar genes is generally more effective than mutation in producing new, but still functional genes. Recombination is most effective if there are multiple genes that can be reassembled. The original genes for antibodies are composed of multiple copies of segments of antibody genes, but during maturation all but one copy of each segment is discarded and a final gene is assembled by recombination between the segments.

Figure IgGenes

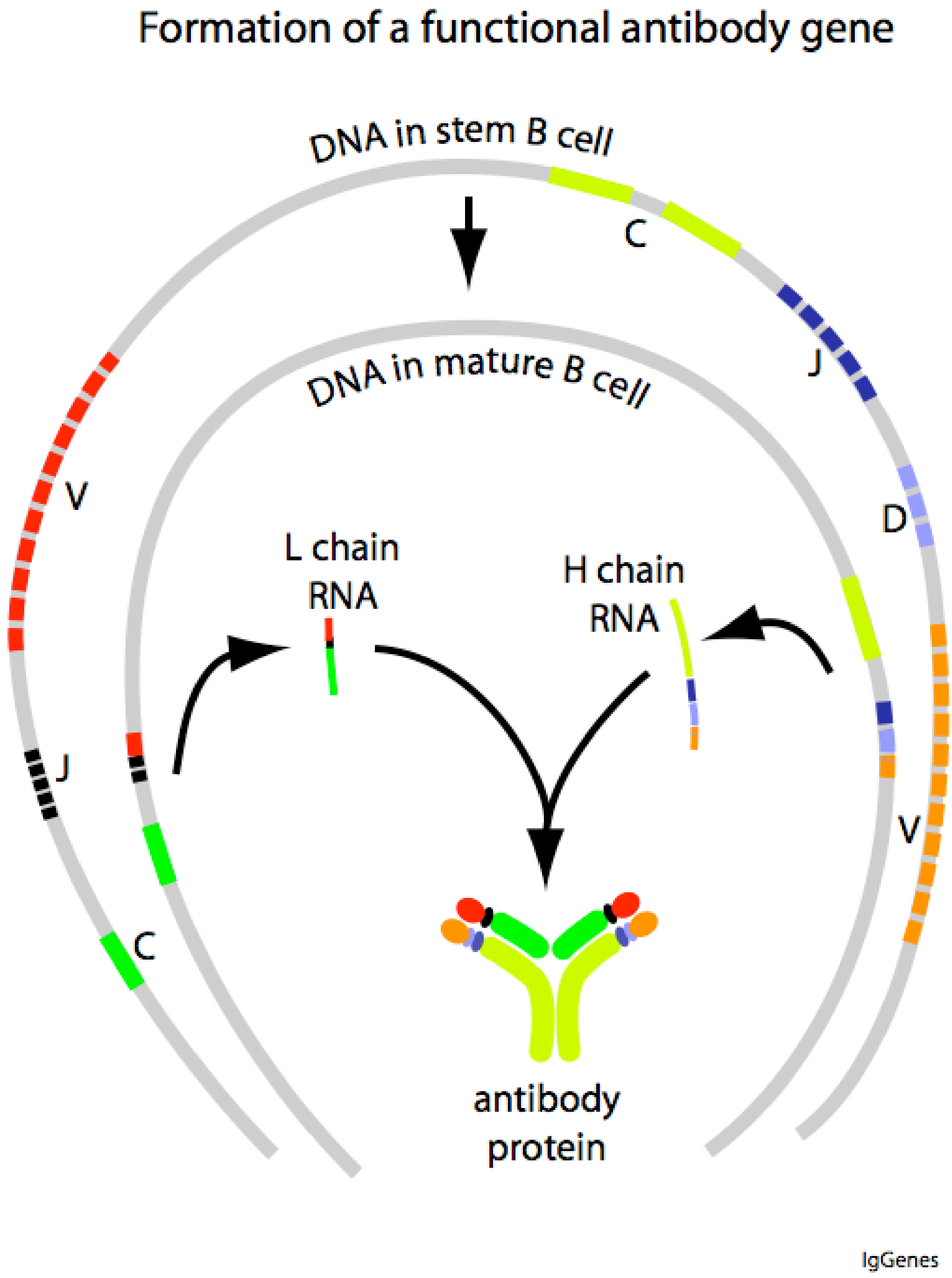


Figure IgGenes. DNA in maturing B cells undergoes extensive recombination in the region of the antibody genes. The multiple copies of domains in the globin genes are selected to generate a single globin sequence. The permutations of the domain copies and 'inaccurate" recombination between them produce a huge number of possible antibody sequences.

As seen in Figure IgGenes, the antibody protein is made of four polypeptide chains, two identical shorter L (Light) chains and two identical longer H (Heavy) chains. The four chains produce a "Y" shaped molecule, with the ends of the arms forming two identical antigen binding regions at the top of the molecule. The lower segments of the chains are known as constant regions because the amino acid sequence there does not vary with the antigen binding specificity of the antibody, although different classes of antibody have different constant segments. It is the amino acid sequence at the top segments of the chains that vary with binding specificity, and they are thus called variable (V) regions. L chains contain a small joining (J) region between the V and C segments, and the H chains contain a small diversity (D) region in addition to the J segment. As seen in the Figure, there are multiple copies of V, J, and D segments in the stem cell genome.

During the formation of a mature B cell recombination occurs within the region containing the multiple copies of the V, J, and D segments. The resulting, shortened genome of the mature B cell generally contains one copy of V, J, and C to specify the L chain and one copy of V, D, J, and C to specify the H chain, although in some cases "extra" copies are removed during the formation of the mRNA chain.

There are a large number of possible combinations between the different V and J segments of the L chain, and an even larger number of possible combinations between the V, D, and J segments of the H chain. There is a special enzyme in B cells that facilitates the high rate of recombination between the antibody segments, but the position of the joint between segments varies over a range of several nucleotides. This sloppy recombination creates more diversity, although at the expense of sometimes producing a joint in which the amino acid triplet code is out of phase, thus producing non-sense protein. Finally, the number of ways L and H chains are assembled to produce the final antibody is the product of the number of different L and H chains.

The HOX gene family

The HOX (homeotic or homeobox) group of genes are found in organisms ranging from flies to humans. These genes are active during embryonic development and code for proteins that control the transcription of other genes. The positional dependent expression of specific HOX genes generate the appendages of insects and the vertebrae and ribs of mammals. These genes all contain a similar 180 nucleotide segment that codes for a 60 amino acid domain that binds specifically to DNA promoter sequences near the genes they control.

In the fly 13 HOX genes are found in one linear cluster. The relative position of each HOX gene along the chromosome is directly correlated with the relative position along the embryo body that the HOX gene controls. Thus, the most 3' HOX gene turns on genes that generate the appendages of the head while the most 5' HOX gene turns on genes that generate the appendages of the tail (abdomen) end of the fly. A dramatic demonstration of the power of a HOX gene is obtained by artificially expressing in the head of a fly the HOX gene that is normally active in the thorax, where it turns on expression of genes that make a leg. True to its native role, the transplanted thorax

HOX gene causes a leg to extend out of the head at the location normally occupied by an antenna.

Figure HOXofFly

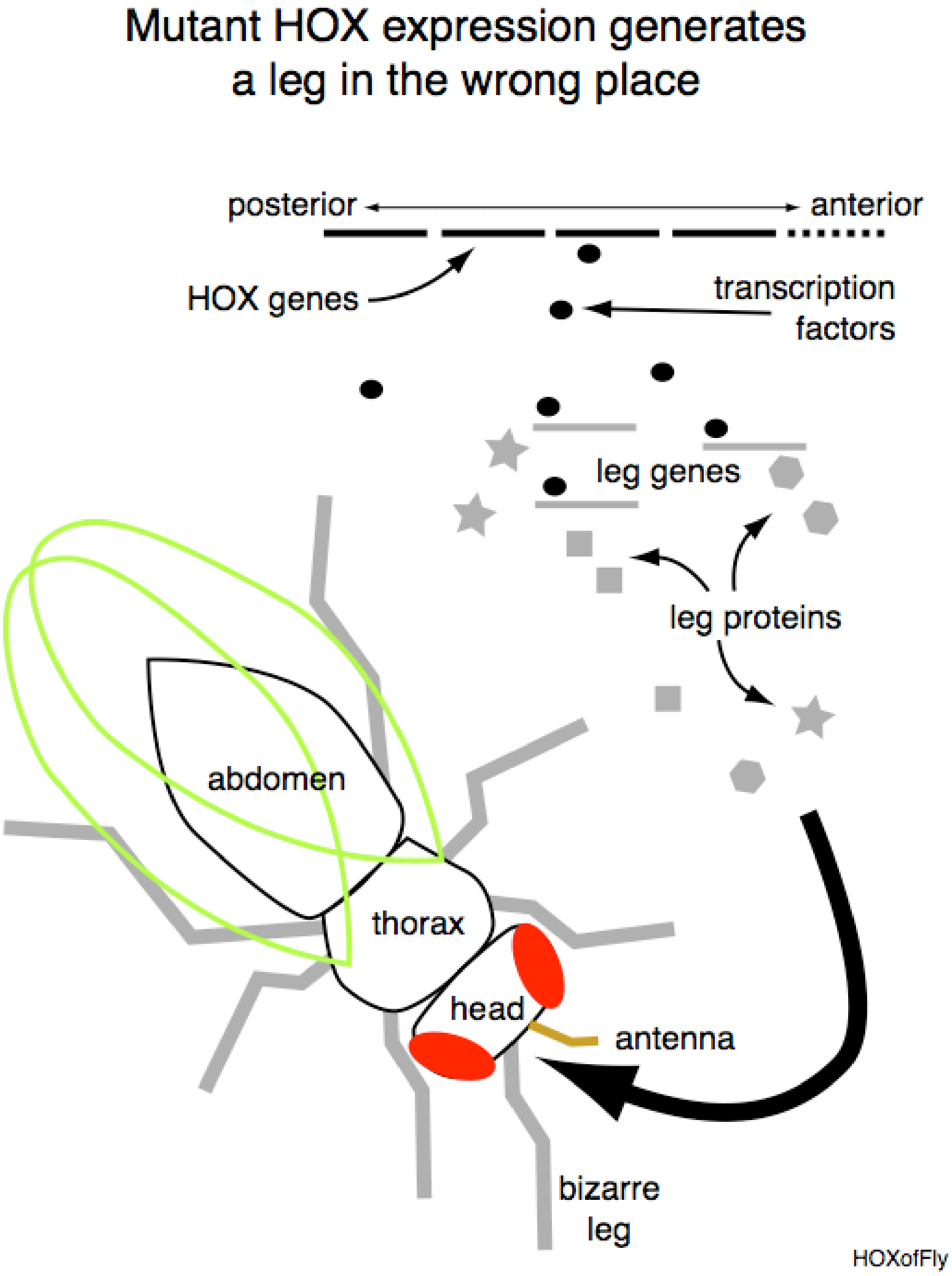


Figure HOXofFly. The embryo of the fly Drosophila develops as a series of segments along its anterior-posterior axis. These segments then partially fuse to form the abdomen, thorax, and head in the adult fly. Each gene in the HOX cluster specifies structures in one body segment by producing a transcription factor that controls the expression of the thousands of genes that code for proteins in the structures of that segment. The position of the HOX gene in the cluster directly correlates with the segment in which that HOX gene is active. If a HOX gene is active in an inappropriate segment, an inappropriate structure will be formed in that segment. In the experiment described by this Figure an activated thorax HOX gene has been inserted in the head, and thus a leg grows in place of the appropriate antenna.

This is already a great story, but it gets better! Since the nucleotide sequence of the HOX genes is similar over a wide range of animals, you might guess that a HOX gene of one organism could turn on the appendage genes of another organism. You are correct; the mouse HOX gene that controls the development of eyes causes eyes to form when it is expressed in the fly. In addition, the squid version of this HOX gene also causes eyes to develop in the fly (fly eyes, not squid eyes)! As you would expect by now, alterations in the human analog of this HOX gene are associated with defective eyes in humans.

Not only are mice, flies, and squids very different animals, the anatomical structure and embryonic origin of the eyes from these animals could hardly be more different from each other. The common action of HOX genes in these organisms thus suggests that the molecular machinery for what was thought to be a rather high level, and thus species specific organization, has instead evolved at a very early stage in the ancestors of all these diverse creatures.

However, as hundreds of laboratories and thousands of scientists explored the details of this system it was shown to be quite complex (no big surprise) and to differ a great deal among organism. The HOX gene cluster seen in the fly has been duplicated twice during the evolution of mammals, and the resulting four clusters are now even on different chromosomes. In addition, some of the individual HOX genes have been duplicated or deleted, so there are a variable number of copies in each cluster. The generation of somites (the repeating body units associated with individual vertebrae) and the activation of the appropriate HOX gene set is a sequential process in mammals, which is driven by a molecular clock. In the fly there is no evidence for such sequential expression.

Species

Now switch from thinking about one animal, plant, or bacterium, to all living things. The diversity of life forms is perhaps as amazing as the complexity of any single organism. However, this diversity has a very non-uniform structure. We see many groups of organisms containing members that are very similar to each other, but few individuals with characteristics intermediate between the groups. The groups of similar organisms are called species.

While the species is an important concept, a complete and unambiguous definition is difficult. Generally, members of one species are defined by the ability to breed and create progeny. They also look alike. However, in some cases groups that are clearly different in appearance and do not normally interbreed may do so if placed in an artificial environment, e.g. a laboratory. In other cases sterile progeny are produced by the interbreeding of species e.g. horses and donkeys. Looking alike may be the only property we can use to group fossils into species, but how alike must members of a group be to constitute one species?

Species can themselves be grouped into an ascending hierarchy based on similarity: Species, Genus, Family, Order, Class, Phylum, and Kingdom. The process of defining the groups of organism and their relation to each other is called taxonomy, and was once the major activity of biologists. It is still an active field, because (believe it or not) new organisms are still being discovered, and new methods of defining relationships are being invented. Many of the new methods use DNA sequence data as the basis for comparison of species and employ algorithms for the comparison that can only be implemented with computers. But why are there discrete species, and how were they created?

Sex maintains species

The existence of species depends on and is maintained by sexual reproduction. A organism created by sexual reproduction contains portions of the genome of each parent, and is thus a hybrid of the parents[2]. Many species, e.g. humans, reproduce exclusively in a sexual mode. However, other organisms most frequently reproduce asexually, only occasionally utilizing a sexual mode. Some organisms do not have separate genders (male or female), but still exchange genetic material during reproduction, e.g. viruses that infect a single cell can exchange genome fragments to produce progeny that are hybrids of the parents.

If a group of organisms never exchange genetic information there is no mechanism to maintain them as a meaningful group. They would evolve as individual progeny organisms, but with no mechanism to link them together to form groups, i.e. species. If the environment was restrictive it would serve to constrict the progeny to have certain properties, but the population would be unlikely to form the discrete groups we see in the real world[3].

The propagation of two species is diagramed in Fig Species. For simplicity, the only characteristic considered here is color, while for real organisms there would be thousands of characteristics that could be followed. As can be seen from the Figure, color must be a complex character, one that probably depends on many genes, and thus is inherited in a complex way that is certainly not decipherable from the Figure. Many characteristics that are of interest to us are complex, however, the information in the scientific literature is mainly concerned with more simple genetic traits that can be easily studied and understood. But of course the scientific literature is mainly reserved for problems that have been solved. This is as it should be, as long as we recognize that the literature is not an unbiased sample of all scientific questions.

[2] A hybrid is conventionally defined as the progeny of two dissimilar strains. Strains are groups of organisms that have some dissimilar characteristics, but are still similar enough to breed. However, essentially all organisms are different from each other; there is most likely at least one nucleotide difference of genomes even among the same strains of the simple viruses. Thus, the progeny resulting from any sexual reproduction is a hybrid is the strictest sense.

[3] Some biologists (but not this one) believe that bacteria can not always be usefully organized into species, presumably because there is not sufficient sexual exchange of genetic information between individuals.

Figure Species

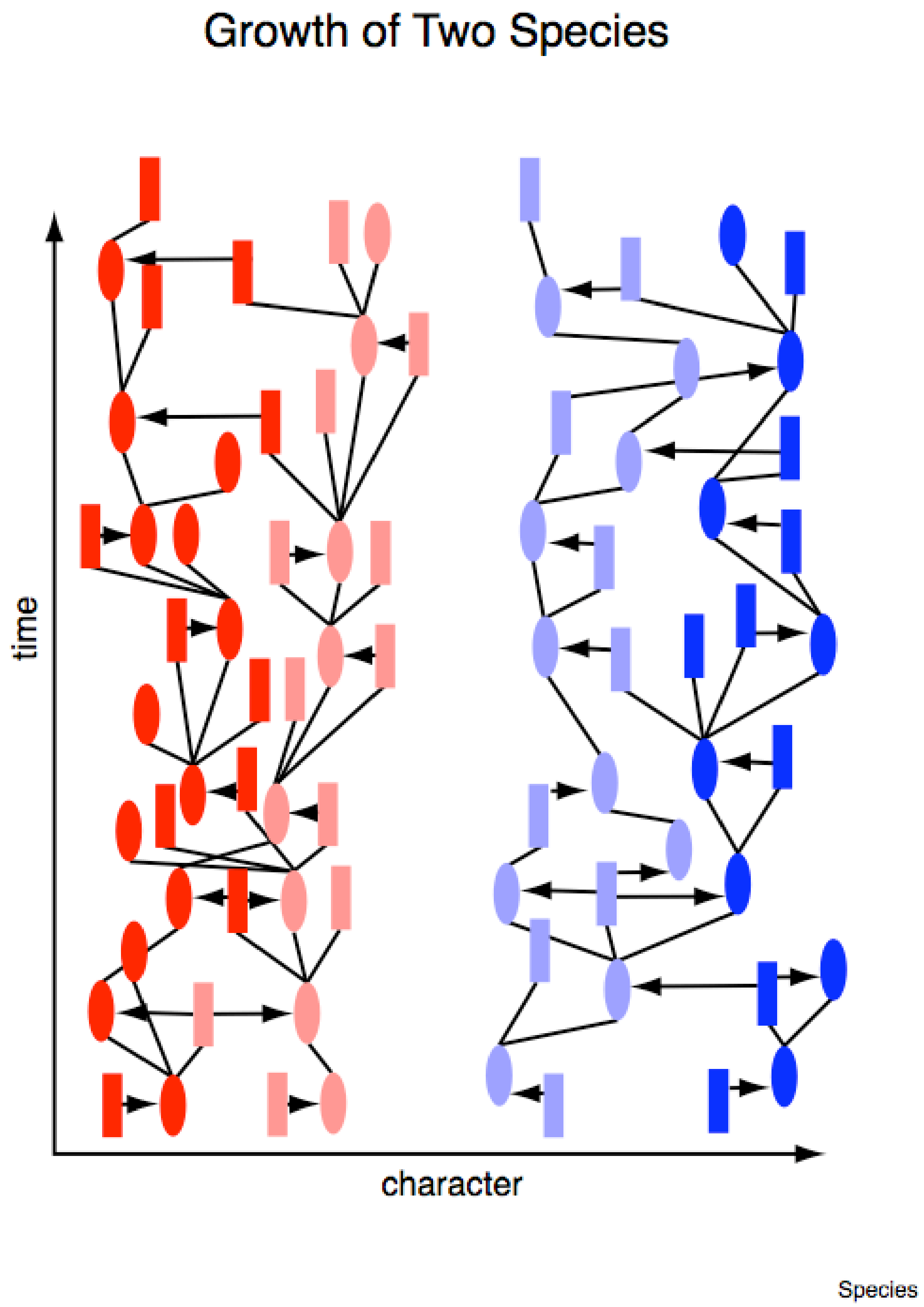


Figure Species. The history of two species, the reds and the blues, is followed in this Figure, as time evolves upward. Reds can only be light or dark red and blues can likewise be only light or dark blue. There are two sexes, rounds and squares. Mating, indicated by horizontal arrows, occurs between one round and one square, with the round subsequently producing one or more progeny, indicated by lines sloping upward. Color intensity is determined by a complex dependency of several genes, and thus no obvious pattern is clear. However, since blues and reds never combine genes, no individuals with intermediate colors are ever produced.

Information exchange occurs only between members of the same species. It is this information exchange that holds the species together as a group. There will be mutations and other genetic changes that occur in the members of the species, but these changes will be transmitted only to other members of the species, and thus the species represents, on average, the sum of all changes that occur and are consistent with survival of the individual (those who die before reproducing do not contribute information to the species). Since individuals of different species can not mate and form individuals with genetic information that is a mix of the two, the species maintain their individual identity.

Humans can only reproduce sexually, however, many other species can reproduce by asexual as well as sexual mechanisms, e.g. bacteria that reproduce by copying their DNA and then dividing into two apparently identical cells. This mode of reproduction is certainly useful if the density of individuals is so low that it is unlikely to encounter another member to mate with.

Almost all regions of the earth contain many different species that are in proximity with each other. Most of these species can not interbreed because of gross physical dissimilarity which prevents mating. But even if two species are similar, and even if they mate, no progeny are produced, or the progeny are not able to reproduce. The failure is the result of the mismatch between the chromosome structure of the two parents that forms the hybrid offspring. In some cases two species may mate and produce a viable offspring, but these hybrids have a defective chromosome composition that prevents them from mating with each other. A well known example among domesticated live stock is the hoarse-donkey pair that can produce healthy mules which however can't breed themselves.

Sex generates diversity

Sex is also a major mechanism for the generation of diversity among progeny. Diversity is required in order for selection to produce an altered population in response to a change in the environment. While the detailed mechanics of sex are varied and complex, the basic process at the cellular level is fairly simple. As diagramed in Figure sex, both male and female parents have a diploid genome, but during a process called segregation, one member of each diploid pair is discarded to produce sperm and eggs that are haploid; they each contain only one set of chromosomes. The union of sperm and egg produces the new individual, which now again contains a diploid chromosome set.

Figure Sex

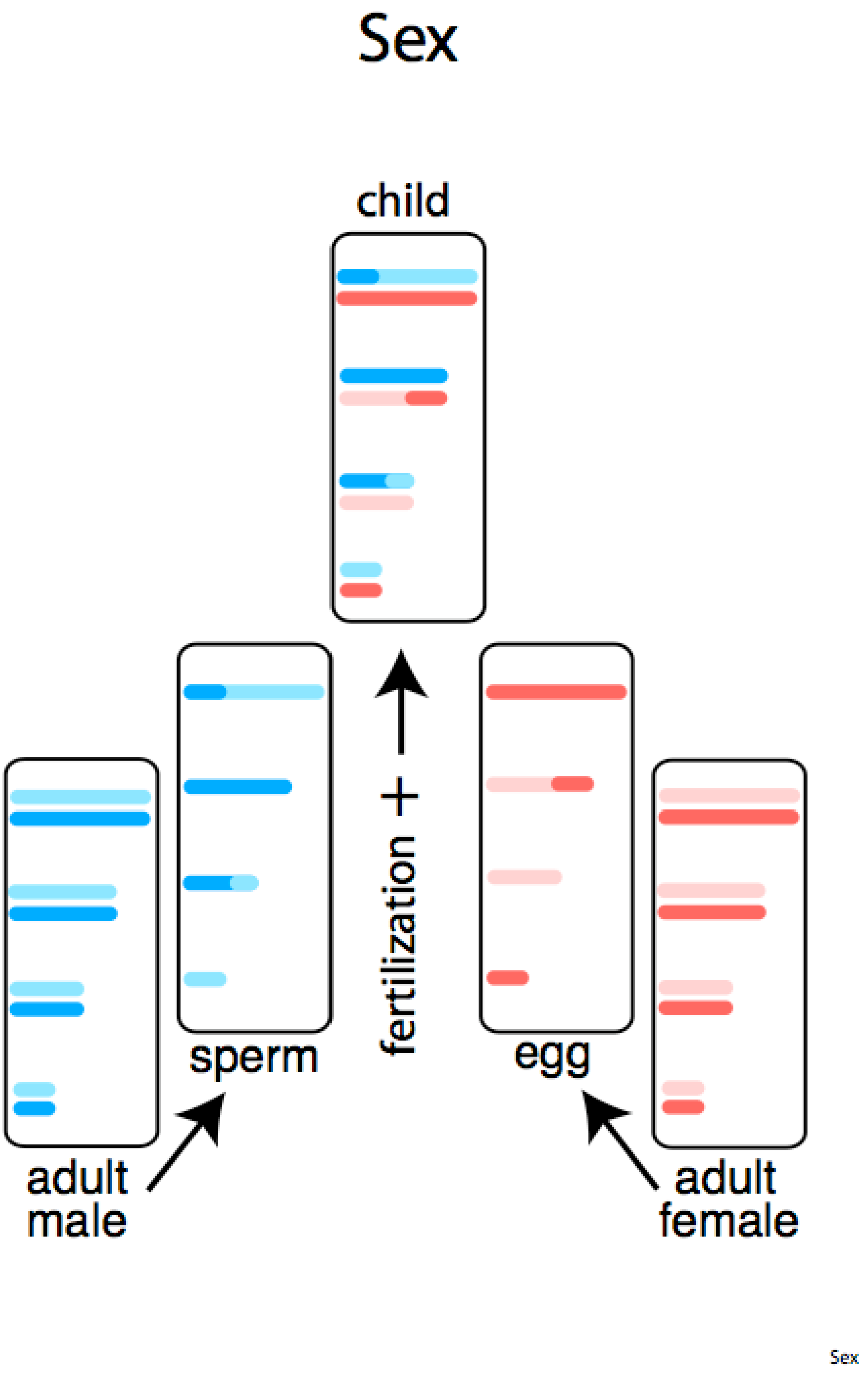


Figure Sex. The adult male and female have two versions of each chromosome. In the formation of eggs and sperm one version of each pair is lost so when the progeny is formed by fusion of egg and sperm it will again have the correct number of chromosomes. Even with only four sets of chromosomes, there are $4^4 = 256$ different combinations of parental chromosomes possible in a progeny. During the generation of sperm and eggs recombination occurs between members of the chromosome pairs, creating chimeras shown here by shades of color. Since recombination can occur at almost any location on the chromosomes, an even larger diversity is created with this process.

Diversity is created in three ways by this process. First, the child has a one copy of each gene from each parent (the sex chromosome are more complicated), thus it is a new individual. However, some of the genes from the two parents will be identical, and other genes may produce a product with indistinguishable function. Secondly, the selection of the one chromosome from each pair of parental chromosomes, which ends up in the child, is random. Thus, a huge number of combinations are possible even if the number of chromosomes is low, and thus all children from one pair of parents are likely to be different. Finally, during the process of selecting one member of each parental chromosomal pair to produce the egg or sperm, genetic recombination occurs between the pair, which generates a chromosome that is not present in the parent. The recombinant chromosome is thus a hybrid between the parents of that parent.

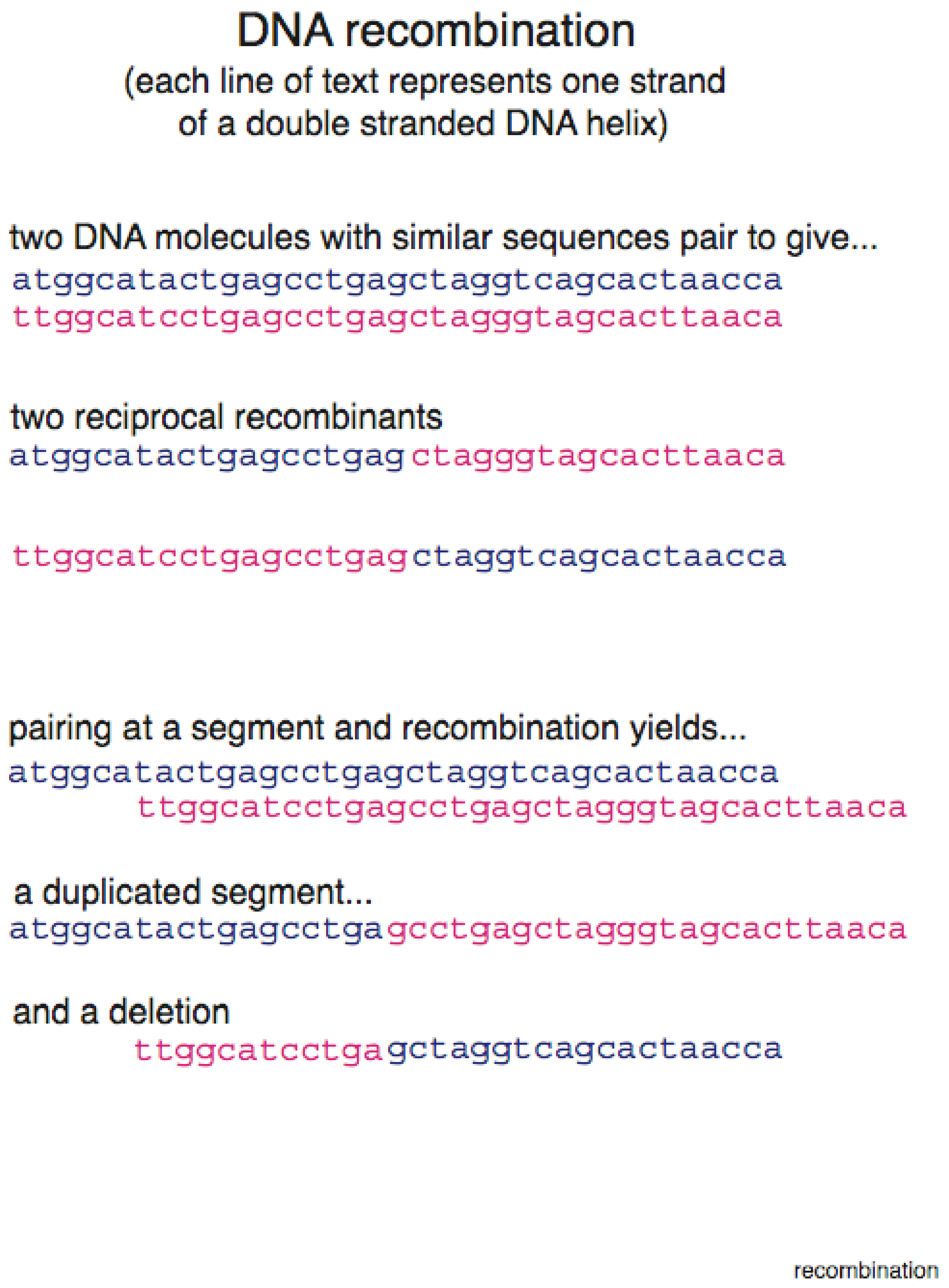
Recombination

Generation of diversity is necessary for evolution to occur. Some errors in DNA replication produce a more fit individual which thus constitutes evolution. Such "productive errors" thus balance the more common negative effects of mistakes in information transfer. An alternative process to create diversity is rearrangement of regions of DNA. Segments of two molecules are exchanged to create hybrids in a process called recombination. The switch from one molecule to the other usually occurs in regions of very similar nucleotide sequences for both molecules. It can be argued that recombination creates more useful diversity than mutations, although mutation is needed to created the variability that is rearranged by recombination.

In a typical cell that has a DNA molecule from each parent, recombination can thus occur between these very similar molecules. When germ cells are formed, they will contain only one member of the pair of DNA molecules present in other cells. If this DNA molecule is a recombinant, the progeny will contain some genes from one parent and some genes from the other. Since recombination can occur in the middle of a gene, the progeny may contain a chimeric protein.

However, in a large DNA molecule there will be several segments with a similar or even identical nucleotide sequence, perhaps by chance. When recombination occurs between non-homologous segments the products will contain deletions and duplications.

Figure recombination



DNA recombination

(each line of text represents one strand of a double stranded DNA helix)

two DNA molecules with similar sequences pair to give...

```
atggcatactgagcctgagctaggtcagcactaacca
ttggcatcctgagcctgagctagggtagcacttaaca
```

two reciprocal recombinants

```
atggcatactgagcctgag ctagggtagcacttaaca

ttggcatcctgagcctgagctaggtcagcactaacca
```

pairing at a segment and recombination yields...

```
atggcatactgagcctgagctaggtcagcactaacca
      ttggcatcctgagcctgagctagggtagcacttaaca
```

a duplicated segment...

```
atggcatactgagcctgagcctgagctagggtagcacttaaca
```

and a deletion

```
      ttggcatcctgagctaggtcagcactaacca
```

Figure recombination. In the cell two double stranded DNA molecules pair along similar sequences during cell division and during the process of producing DNA for the egg or sperm (this process is distinct from the pairing of complementary single strands of DNA). An enzyme system can then break and rejoin the two DNA helixes to form recombinants, chimeras that contain segments of each of the original two molecules. However, pairing can occur along small segments that are not really homologous, and recombination at such sites generates a molecule that has a repeated segment and a molecule that has a deletion.

Creation of species

So that's why species are discrete and have some stability, but how are they created? Of course a new species is not created out of thin air, it must be the descendant of another species. In the most common scenario, a group of individuals from the parent species becomes isolated from it, and changes under the selective pressure of a different environment. The isolation prevents the entire group from evolving together, rather, two species are formed. After a certain degree of divergence, the two groups of organisms sufficiently different that they can not interbreed even if they should become neighbors again.

The above scenario for creation of two species from one is essentially a continuous process, even though in the end it may produce to groups of individuals that every biologist would agree were different species. Thus, there would a time in the evolution of the two species when some individuals from the two groups could interbreed but others could not. Or, separation into two species might be detected by a gradual decrease in the probability that offspring from mixed parents would survive as compared to offspring from parents in the same group. In addition, it could be difficult to determine the exact extent of the divergence into two species (as defined by sexual reproduction of mixed parents) by looking at the differences in morphology of the two groups. Thus, while generally animals, plants, and bacteria can be grouped into discrete species, there are examples where populations form an almost continuous spread.

Kinetics of evolution

It might seem intuitive that species would evolve gradually. Darwin assumed gradual evolution when he formulated his theory. In a few cases there is a fossil record that documents a gradual change in a species. However, generally fossils suggest sudden changes followed by long periods in which morphology does not change much. Proponents of gradual change in evolution have suggested that lack of a chronological series of fossils with incremental changes in morphology is only the result of a incomplete fossil record. While it is certainly true that the fossil record is unfortunately the product of the erratic probability of the organism being preserved and the probability of it being found, most paleontologists now believe in punctuated equilibrium, popularized by Eldredge and Gould. This proposal is that the fossil record is mostly correct; species arise suddenly (suddenly for a paleontologist means within a million years or so) and then remain fairly constant over tens or hundreds of millions of years.

Figure LifeTree

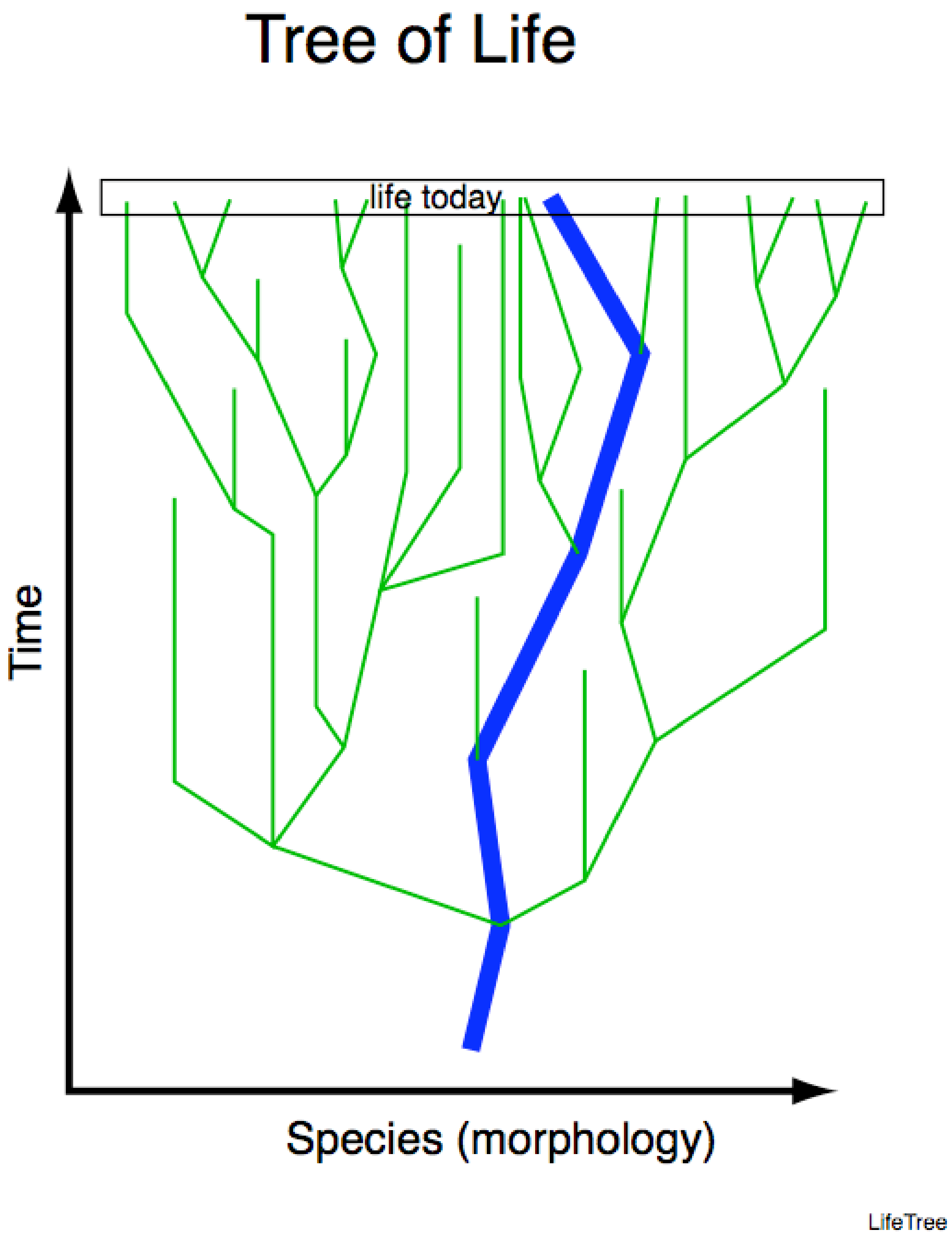


Figure LifeTree. The creation, evolution, and elimination of species can be visualized as a tree, where time increases along the vertical axis and morphology (and the sum of all other characteristics) is indicated on the horizontal axis. The branches that extend to the top of the

Figure are the species that are alive today, while the ones that end before reaching the top are extinct. Nearly vertical lines represent species that have changed little over time while diagonal lines represent periods of rapid change. The evolution of one species, say humans, is indicated by the thick blue line. Note that no species alive today has, strictly speaking, evolved from another that is alive today. However, two species alive today may be closely related, and thus appear to be similar, because they recently evolved from a common ancestor.

Most of our direct knowledge of ancient organisms is derived from fossils. Animals and plants that are large and have some resemblance to those now alive are the most dramatic and thus were the first to be described. Fossils of bacteria have actually been found and identified, but it's difficult to obtain much information from them. However, from a biochemical (which I would say is a fundamental) viewpoint, most of the interesting evolution had already taken place when bacterial first appeared on the earth. All this is to say that the paleontology record is heavily biased toward the late phases of evolution. However, to give an approximate idea of the time scale of evolution we can look at past periods that have been given names by geologists and paleontologists.

Figure LifeTimeLine

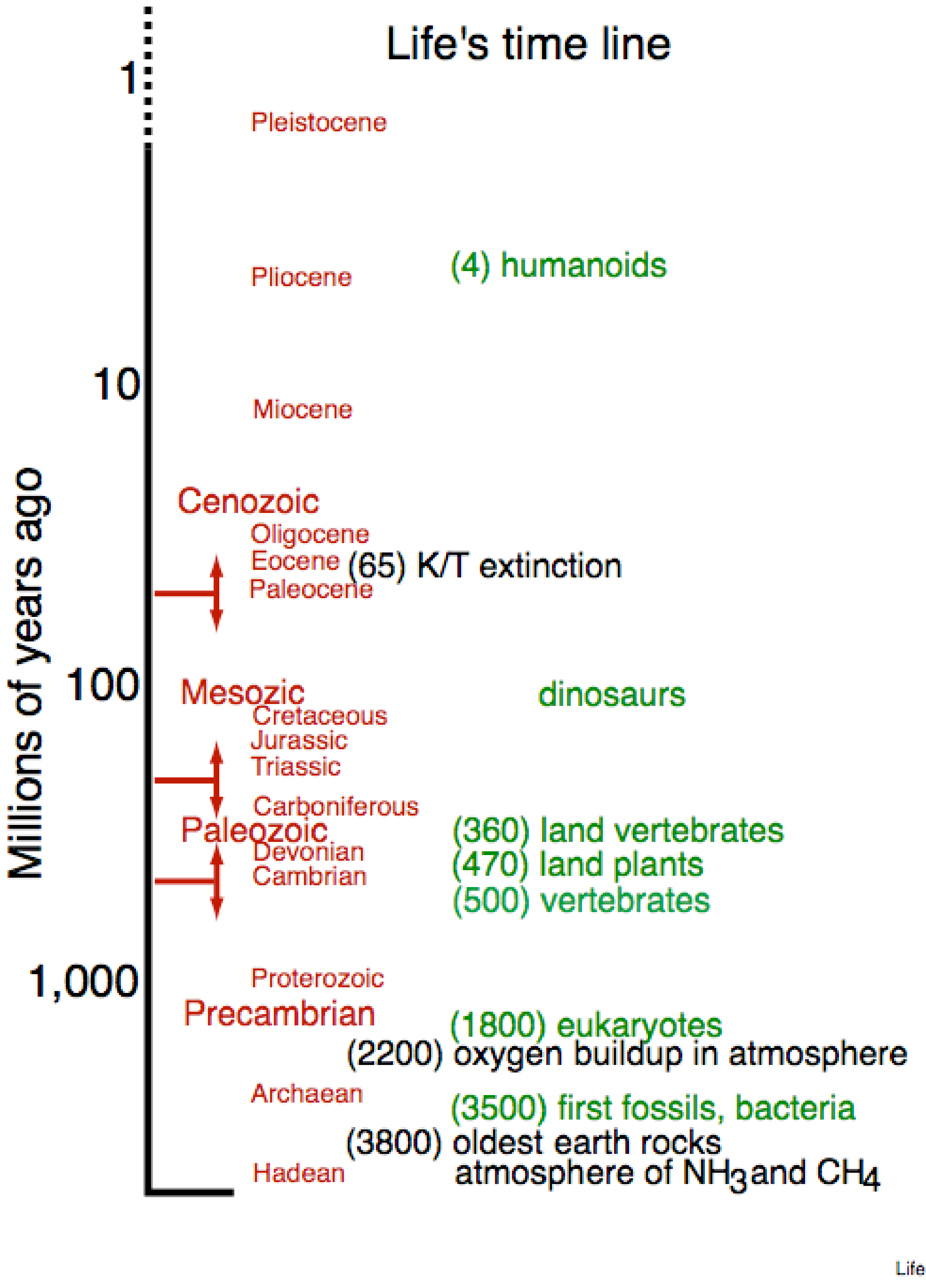


Figure LifeTimeLine. Note that the vertical time line is on a logarithmic scale. This scale can be justified by the fact that we know less and less about life the further back in time we go, and the logarithmic scale compresses as it extends to older times. The oldest fossils appear close (on this scale) to the time when the surface of the earth become solid and rocks formed (the earth formed about 4.5 billion years ago). Note that the atmosphere of the early earth contained methane and ammonia, but no oxygen, and the first organisms evolved in this environment. The appearance of oxygen, generated by life itself, caused dramatic changes in the biochemistry of all organisms. As an example, the early hemoglobins probably functioned to protect the organism from chemical damage from oxygen, and only much later served to deliver atmospheric oxygen to the new enzymes that could now use it. There have been several sudden and extreme changes in earth's environment which have resulted in the extinction of a significant fraction of the species alive at the time. One of these, the K/T extinction, thought to be due to the impact of a large meteor, occurred about 65 million years ago. Among other effects, it appears to have caused the extinction of the dinosaurs.

Model organisms

The medical and biological research communities use model organisms[4]. In medical research the term "model organism" seems natural, since the defined goal is to understand human, and thus all other life forms are seen as models for humans. However, even in more fundamental research that is not focused on the human there is a use for models. Models are species selected because they have some set of useful properties for experimental work. Each model species has objective advantages and disadvantages, but the choice of species also has a subjective and historical component linked to the sociology of the scientists that make that group decision. Some species are popular for several decades, but then as the center of gravity of research shifts other species become popular. For several decades, 1940 to 1970, the bacterium *E. coli* and several viruses that infect this bacterium, e.g. T4 and lambda, were the subject of many papers. In later decades, because the fundamental principals of molecular biology had been worked out using these relatively simple organisms, yeast, round worms, the fruit flies, Zebra fish, and mice became more popular. The ebb and flow of interest in these organisms was actually more complicated than a simple shift from simple to complex; some of the most fundamental studies of genetics were performed using fruit flies during the 1930's and mice have always been used for medical and toxicology work.

The US National Institutes of Health has awarded research grants worth millions of dollars to determine the complete nucleotide sequences of the genomes of these model organism and the functions and behavior of the gene products. The motive was not based on any intrinsic interest in the organisms themselves. As we have suggested, the strategy of this research is based on the belief, or rather the understanding, that there is a great deal of similarity, even down to the level of nucleotide sequence of the genes, between these organisms and humans. The extent that understanding gene sequence and function in the single cell organism that is used to make bread, yeast, helps unravel some of the problems of human biochemistry, surprises even the scientists that work in the field.

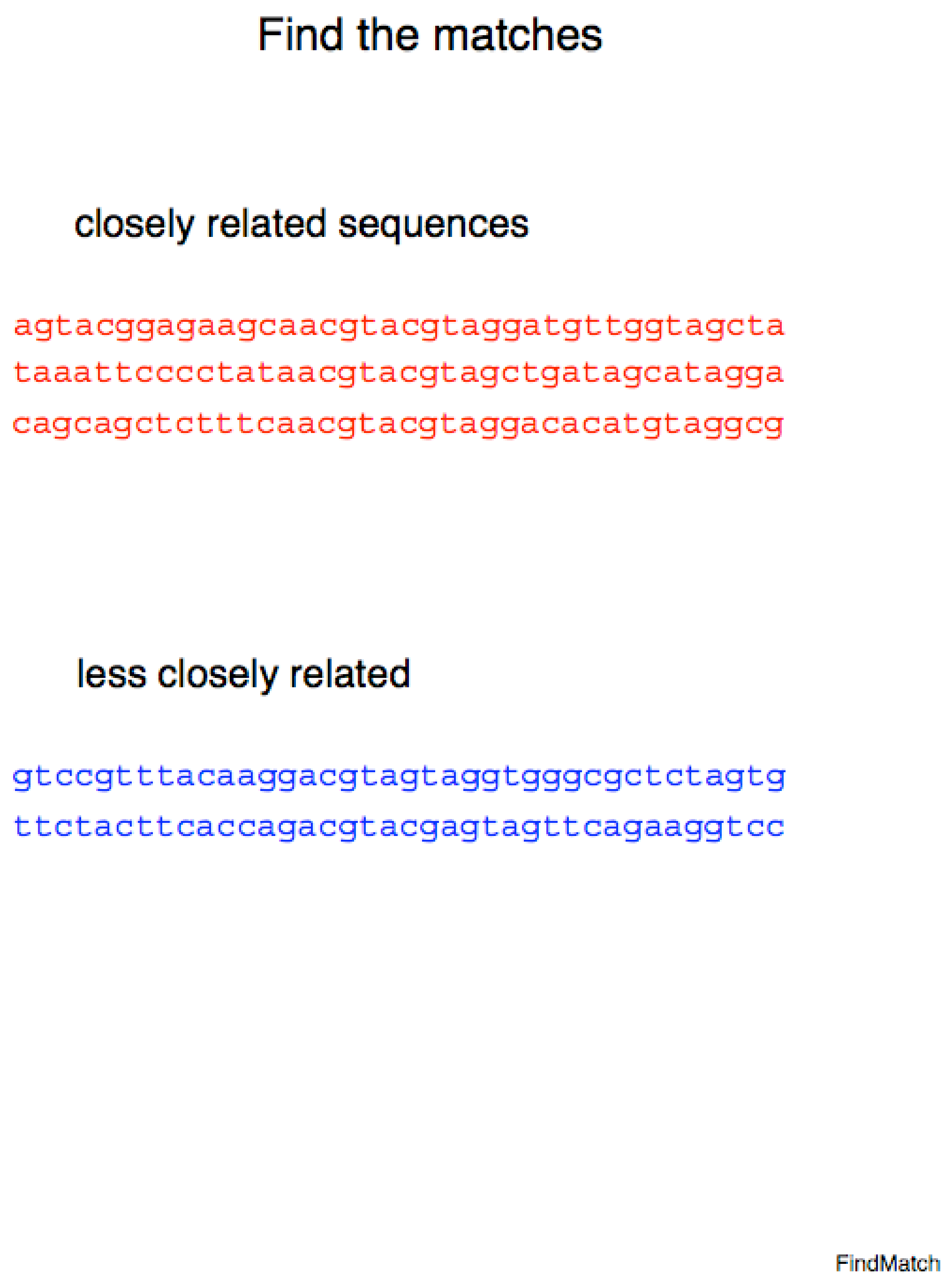
The basic paradigm is that we are all brothers and sisters, particularly at the molecular level, even if our common parents existed many millions of years ago. If we

[4] I distinguish medical and biological research because it has very different goals. Medical research attempts to find diagnostic tests and cures for disease states, while biological research attempts to understand life. Of course the two often blur together, and one person may be involved in both, or a hybrid of the two. However, they still have separate goals.

were not all related there would be no explanation for the relation between our genes, and there would be little justification for spending large amounts of energy and money to determine the nucleotide sequences of model organisms.

A specific example of the use of comparison between multiple species is comparison of genomic nucleotide sequences between species of yeast that were closely related and those that are more distant from each other. The goal was to find transcription control sequences. These are much more difficult to identify than the genes themselves because there are fewer rules that define them. The strategy was based on the logic that much of the genome was not actually code for information and thus there would be a great deal of variation between species yeast. However, control sequences are important, and thus would be similar between species. Many conserved sequences were identified in this way, and their function confirmed. The principal is illustrated in Figure FindMatch.

Figure FindMatch



Find the matches

closely related sequences

```
agtacggagaagcaacgtacgtaggatgttggtagcta
taaattcccctataacgtacgtagctgatagcatagga
cagcagctctttcaacgtacgtaggacacatgtaggcg
```

less closely related

```
gtccgtttacaaggacgtagtaggtgggcgctctagtg
ttctacttcaccagacgtacgagtagttcagaaggtcc
```

FindMatch

Figure FindMatch. In this example the task is to find the mostly common sub-sequence in the five larger sequences. The assumption is that this sub-sequence is mostly conserved because it performs a function, while the flanking nucleotides are free to change by mutation. I have made comparison simple by artificially aligning the sequences so the common regions are above each other. Even so, this very simple task should convince you of the value, or rather necessity, of using a computer to find common sequences.

Evolution: paradigm of life

The process of evolution has been mentioned many times in this chapter on the genome. That is because evolution, implicit or explicit, is the basic concept that makes the living world understandable, at both the macroscopic and molecular scale.

It obvious that the genome must contain the information needed to make the machinery that keeps the organism alive. However, there are two other very fundamental and related processes that must occur, and the genome contains the information for accomplishing them also.

First, unless the organism will be only a transient object, it must produce progeny. Another option for an organism might seem to be immortality. However, immortality is not a real option. However well constructed, there will always be situations in which any real organism will be destroyed. Falling trees, intense storms, earthquakes, meteor collisions, lucky predators; let your imagination go wild here. No, the only real immortality is the virtual one; you must produce progeny that contain your genetic information.

Secondly, in the long run it is necessary for organisms to change, because the environment will eventually change and if you can't change also you won't be able to survive and compete with other organisms. Since there is no mechanism for an organism to change in a fundamental way, the only way to effect change is to produce progeny that are variations on the parents. While parents may not be able to predict which changes will be beneficial, if they produce progeny with some diversity, at least some of the progeny may able to survive.

A final, less obvious, but very important requirement for an organism is that it must have been created. Of course this is just the same process as producing progeny, but seen going backward in time. However, going backward in time is a little different when seen in its entirety, since eventually history must reveal progressively greater simplicity, until we finally see the transition of non-living to living objects[5]. The farther back in time we go, the more difficult it is for us to deduce the nature of the biologic world and the relationships between the animals, plants, and bacteria, i.e. who is the ancestor of whom. The beginnings of life are the most difficult to imagine and reconstruct. The transition from non-living objects to living organisms seems most likely to have been a diffuse ill-defined event.

[5] I am not saying here that as time and evolution go forward organisms always and necessarily become more complicated and “advanced” (the inverse of claiming that as we look at older and older ancestors they become more simple), an idea that was favored by the Victorians. As time goes forward a species may evolve into a simpler organism, perhaps by becoming a parasite which comes to depend on the biosynthetic machinery of the host and thus looses some of its genes. Thus, as we go backwards in time the chain of ancestors may become more complicated, but only for a time. Eventually the organisms must become more simple.

This process of reproduction and change is called evolution. The basic idea is simple enough: while progeny are very similar to their parents, they are slightly different from them and different from each other. The environment the progeny live in exerts a selective effect, and thus some progeny produce many other progeny while others produce few or no progeny or even die. The environment thus selects some individuals for survival, and thus produces a change in successive generations. Evolution by variation in progeny and selection of the more fit is the model for the creation and organization of all living organisms on the earth. Evolution is thus the paradigm of life.

It is only at the level of the genome that we are forced to think about the entire organism. The collection of molecules that make up one organism must work as a unit and this unit either succeeds or fails to create progeny. While the different parts, e.g. different genes, may contribute unequally to success, the entire collection, the individual organism, is the object that produces the progeny, and thus is the smallest unit that can evolve in the Darwinian sense[6]. However, the interactions between individuals, the behavior of a group, can certainly influence the survival of individuals in that group, and thus evolution also acts at the level of groups.

The genetic information in an organism exists because it has enabled the parents of the organism to survive to produce progeny. No such relationship exists between the information carried by the Internet and the survival of the Internet[7].

[6] Of course genes in the genome of that organism do evolve as the organism evolves, but are the genes the unit of evolution. On this issue I side with Gould.

[7] Here I exclude information in the headers of packets and the information passed between routers, since the function of this information is also to enable the Internet to function and survive.

Chapter summary

The genome of a living organism is the collection of molecules (usually DNA) that contains the total genetic (inherited) information for that organism.

Organisms large enough to be seen by eye contain many cells, with each cell containing the same genome. The major part of the genome is in a nucleus where the DNA molecules are wound around proteins to produce a very condensed structure. A minor part of the genome is contained in other structures in each cell, e.g. mitochondria. Humans and many other organism have a diploid genome: there are two almost identical copies of each DNA molecule, one from the mother and one from the father.

A gene is the segment of a DNA molecule that contains the information to produce a stable RNA or protein molecule. Genes make up only a small fraction of the human genome. Some of the “extra” DNA represents control sequences or old genes that are no longer used. Much of the “extra” DNA is virus-like segments that can move around in the genome.

Current genes evolved from older genes, often by a process of duplication and mutation. This process creates gene families. Ultimately all genes are related to each other just as all organisms are related.

The genes for antibodies evolve at a rate many orders of magnitude faster than other genes. This requires a unique fragmented structure for the antibody genes plus the activity of modified replication enzymes that facilitate rapid evolution.

Sex is a mechanism for two organisms to exchange and share genetic information to create hybrid progeny. This exchange of information is the mechanism for maintaining a species. The division of organisms on the earth into discrete species is a result of exchange of genetic information within species.

As with genes, organisms are not created from scratch, rather they evolve from other organisms. Since all organisms are related, it is often useful to compare function and structure of different organisms in order to deduce the function of genes and gene products. This is why “model” organisms are useful in scientific and medical research.

The unifying principal for understanding life is evolution.