Molecular Evolution

and Phylogenetic Analysis

David Pollock and Richard Goldstein©

Introduction

All of biology is based on evolution. Evolution is the organizing principle for understanding the shared history of all biological organisms. Evolution describes the similarities between different organisms, as well as explaining how differences emerged. In addition to answering basic questions about the history of life, evolutionary perspectives and information drawn from evolutionary analyses can provide information highly relevent to many biological, biotechnological, and biomedical problems. There is also growing interest in mimicking evolution in the test tube in order to develop RNA, proteins, and organisms with specified properties.

Increasingly, biocomputing has taken advantage of the methods and approaches of evolutionary biology. This is due to the intersection of a number of different developments: the availability of whole genomes from a growing number of organisms, the availability of high-speed computational facilities that allow sophisticated computational and statistical models, and the growing realization of the power of comparative sequence analysis and how such an approach requires understanding the ways that the corresponding organisms evolved and changed. There is, unfortunately, a lack of understanding about how evolution occurs as well as the various tools that are available to analyze evolutionary data. The purpose of this tutorial is to introduce biocomputing professionals to both the approaches and methods.

Evolution will occur when three requirements are met: variation between offspring, difference in survival probabilities due to these variations, and for the variations to be heritable. In this tutorial, we first briefly describe the genotype of an organism, the DNA that (generally) carries the inherited traits and is modified during evolution. We then describe the random changes that can occur in the DNA, from small local changes to more global rearrangements. We discuss how variations in the genome can be described, and what happens to that variation. This leads to a discussion of some of the central debates in evolutionary theory, specifically the role of adaption versus neutralism, the consistancy of the molecular clock, and the role of population dynamics in the evolutionary process. Finally we discuss the techniques that have been used to understand and model the evolutionary process, including phylogenetic analysis.

This material is covered in depth in a number of books such as *Fundamentals of Molecular Evolution* by Wen-Hsiung Li and Dan Graur, *Physical Approaches to Biological Evolution* by Mikhail Volkenstein, *Molecular Evolutionary Genetics* by

Masatoshi Nei, and *Molecular Evolution: A Phylogenetic Approach*, by Roderic Page and Edward Holmes.

Introductory Genetics

With the exception of certain viruses, the basic carrier of genetic information is DNA. DNA is made up of four different bases (Adenine, Cytosine, Guanine, and Thymine, abreviated A, C, G, and T) linked to a deoxyribose sugar and phosphate chain. The two ends of the chain are called the 5' and 3' ends, after the organic chemistry nomenclature

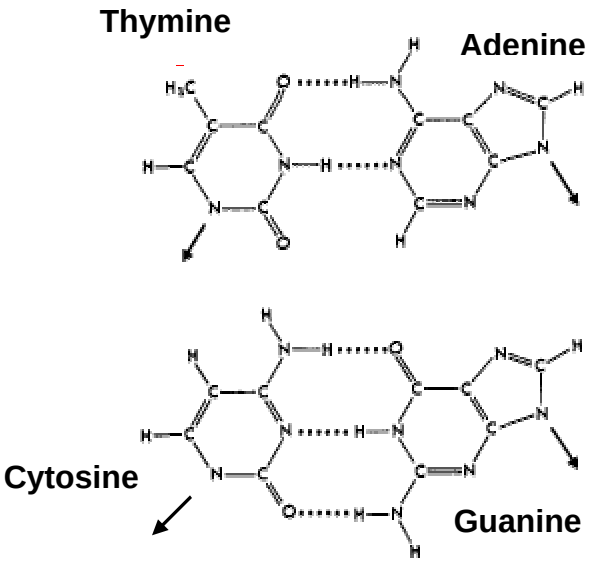


Figure 1: Hydrogen bond patterns for complementary DNA bases.

for the constituent sugars. A and G are called purines and C and T are called pyrimidines, based on the structure of the nitrogenous rings. In double-stranded DNA, T generally forms a hydrogen-bonded pair with A, while C and G form a similar pair. The two strands go in opposite directions, with the 5' end of one strand aligned to the 3' end of the other strand. The form of these four base-pairs and the hydrogen bond are shown in Figure 1. In eukaryotes, these long strands formed into DNA/protein complexes called chromatin, which in turn are packed into structures called chromosomes. Each chromosome has a special assembly called a centromere, which is central to segregation of chromosomes during cell division.

Most organisms are either haploid or diploid. Haploid organisms have one copy of each chromosome. Diploid organisms have two copies of each chromosome (except for the sex chromosomes). These copies may be slightly different, but they are evolutionarily related and are thus called homologous.

In order for the cell to divide, the genetic material must be copied. This can be performed by using each strand of the double-stranded DNA as a template for generating a copy of the other strand. In standard asexual cell division, or mitosis, each chromosome individually duplicates its genetic material, as shown in Figure 2. Cytoskeletal structures, called microtubules, attach to the centromeres and pull the two halves of the duplicated chromosomes apart. For organisms with haploid genomes, mitosis is the only form of reproduction. For diploid organisms, sexual reproduction is also possible. This involves first producing haploid germ cells through the process of meiosis, followed by the sexual association of two haploid germ cells (generally from different organisms) to yield a diploid cell that undergoes further mitotic division to form the adult organism. Meiosis also involves duplication of the genetic material, followed by segregation to yield four germ cells, each with only one copy of each chromosome.

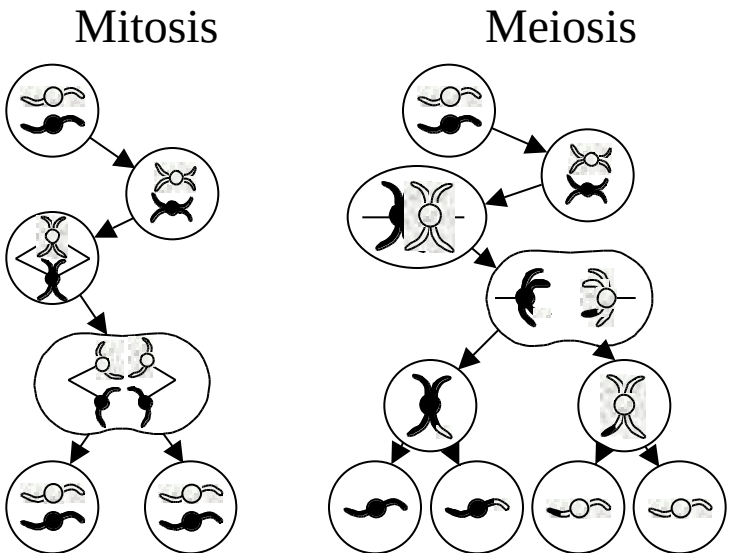


Figure 2: Diagram of mitosis and meiosis

DNA is divided up into units called genes. Genes can be on either strand, and are oriented in the 5' side towards the 3' direction. Thus genes on different strands are oriented in opposite directions. In the past, a gene was considered as the DNA necessary to encode the construction of a single protein. Things are now known to be more complicated, and a single gene can encode an entire protein, a subunit of a protein, an RNA molecule, or fulfill some other functional role. Most eukaryotic DNA consists of elements that are not genes and are often of unknown purpose. They are sometimes refered to as "junk" DNA.

A locus refers to a location of a given gene in the genome. Different variations of the genes are called alleles. If there are multiple alleles in a population, then the locus is said to be polymorphic. As diploid organisms have two copies of each chromosome, it is possible that an individual will have two different varients, that is, two different alleles. If the corresponding genes are different in the two chromosomes, the organism is called a heterozygote, while a homozygote would have two copies of the same allele.

For the protein-coding regions, the DNA is translated into messenger RNA (again through simple base-pairing). In addition to having a sequence complementary to the DNA, RNA also uses uracil, U, in place of thymine, T. The RNA is then translated into the protein sequence in units of three bases, called a codon. Each codon tells the cell machinery which of the twenty amino acids to add next to the growing protein chain. Almost all living organisms use the same translation between DNA and proteins, shown in Table 1. (Mitochondria, for example, use a slightly different code.) For instance, the sequence AUAUGUAUUAAGGCA would be read as AUA/UGU/AUU/AAG/GCA and would code for the sequence Ile (AUA), Cys (UGU), Ile (AUU), Lys (AAG), Ala (GCA). There are three triplets that do not code for any amino acid (represented with '-' in Table 1). These represent "stop" codons and result in the termination of the amino acid chain.

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

Table 1: The genetic code

There are, of course, no markers indicating the boundaries between the codons (indicated above with slashes). Correct translation involves starting at the correct point and accurately translating the RNA sequence three bases at a time.

How the genome changes

The process of evolution works through the generation of random variation among the offspring of each individual. This involves a change in the genetic material that is passed on to the next generation. These changes can involve local changes in the DNA sequence (genetic mutations), larger scale changes in the chromosomes (chromosomal mutations), and loss or gain of chromosomes (genomic mutations):

Genetic mutations

The simplist genetic mutation involves a substitution of one base for another, for example, an adenine becomes a cytosine, causing ATCCGTAGTCCTGAAT to become ATCCGTCGTCCTGAAT. This can be caused by a number of possible events, such as

tautomeric changes where a base is turned around in an abnormal position allowing non-canonical base pairing, chemical modification of the DNA (generally deamination), or radiation damage. Such changes can convert a purine to a purine (A_G) or a pyrimidine to a pyrimidine (C_T), in which case the mutation is called a transition, or a change can convert a purine to a pyrimidine or vice versa (A_T, A_C, G_C, or G_T), which are called transversions. Eight of the twelve possible mutations are transitions. In spite of this, evolutionary change generally involves 50% more transitions than transversions, as these mutations are generally more conservative and less likely to cause less fit offspring.

In regions of the DNA that encode protein sequences, we can also classify mutations according to whether the encoded protein sequence changes. As can be seen from Table 1, the genetic code is degenerate: certain codons (for example, AGA and CGA) code for the same amino acid. A change in the DNA that does not produce a change in the encoded protein (for example, an A_C transition causing AGA_CGA) is called a synonymous or silent mutation, while a similar change that does change the expressed protein (for example, an A_C transition causing AUA_CUA) is called a non-synonymous or missense mutation. A change to one of the stop codons results in a premature termination of the protein chain. These are called nonsense mutations. Approximately 70% of all changes in the third position of a codon are silent, while none of the changes in the second position and only 4% of changes in the first position are silent.

A second form of genetic mutation is the insertion or deletion of segments of DNA. These can be either a few bases, or much larger stretches. If a region of a protein coding region is inserted or deleted that is not a multiple of three, this interrupts the correct translation of the DNA into proteins after this point. This is called a “frame shift” mutation. For instance, in the earlier example AUA/UGU/AUU/AAG/GCA was shown to code for the sequence Ile, Cys, Ile, Lys, Ala. Insertion of a **U** in the middle of the segment resulting in AUA/UGU/**U**AU/UAA/GGC/A would code for the sequence Ile, Cys, Trp, stop. The synthesis of this protein would prematurely terminate at this point so that any other amino acids encoded in the gene downstream of this site would be absent.

Chromosomal mutations

Larger mistakes can be made in duplicating the DNA. It is possible for a region of the DNA to be duplicated. Often the duplication occurs so that the copy is adjacent to the original, but the duplication can be reversed (so that the pattern on one strand becomes copied to the other stand in the inverse order), or it can be duplicated in another part of the chromosome or even on a different chromosome. These various possibilities are indicated in Figure 3.

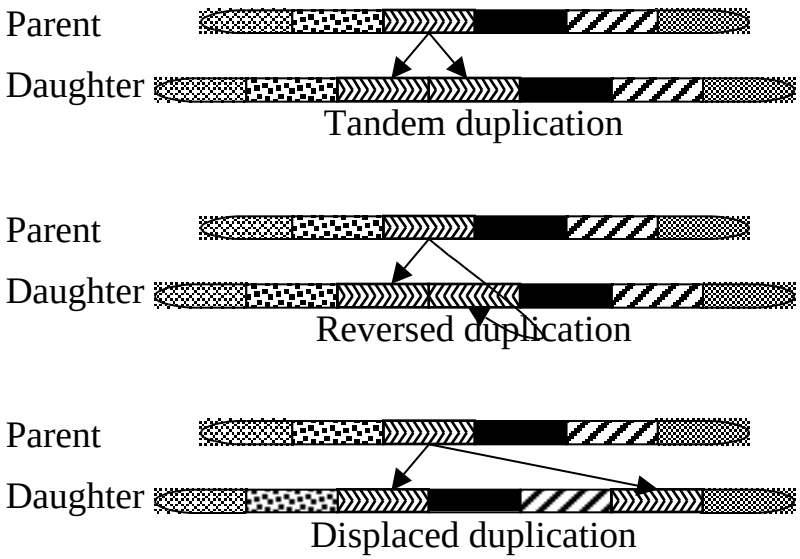


Figure 3: Duplication events

Repeated elements are more likely to be repeated and thus generate larger repeated elements. This can be understood as accidental base-pairing between highly-similar elements. For instance, if expressed as DNA, it would be easy for the above sentence to mutate to "Repeated elements are more likely to be repeated and thus generate larger repeated and thus generate larger repeated elements."

There are other ways that genes can rearrange. A larger section of the genome can simply deleted, inverted to the other DNA strand, or translocated to anther part of the geneome. These processes are illustrated in Figure 4.

The human genome has demonstrated the importance of transposable elements. These are regions of DNA that are capable of either moving themselves to other parts of the genome or directly making multiple copies of themselves. These units seem to exist as independent elements, evolving independently of the host organism with a fitness that only depends upon how many copies of itself it can make, and thus representing the epitome of the "selfish gene" described by Richard Dawkins. These elements can have a major impact on the evolutionary process, in that often adjoining regions of the DNA are moved or copied along with the transposible element. In this way, the existence of these "parasitic" elements may be more synergistic, in that their actions may speed the evolutionary process.

All of the changes described above can occur in haploid as well as diploid organisms, and primarily involve the process of DNA replication whether in mitosis or meiosis. Another common and important form of genomic change involves the exchange of information

between homologous chromosomes, that is, between the two similar chromosomes in a diploid organism.

Consider the diagram of meiosis in Figure 2. This diagram shows the relationship between the two homologous chromosomes. Following replication of the DNA, the homologous chromosomes arrange themselves in the midplane of the cell, and division occurs so that one copy of each (duplicated) chromosome goes into each resulting cell.

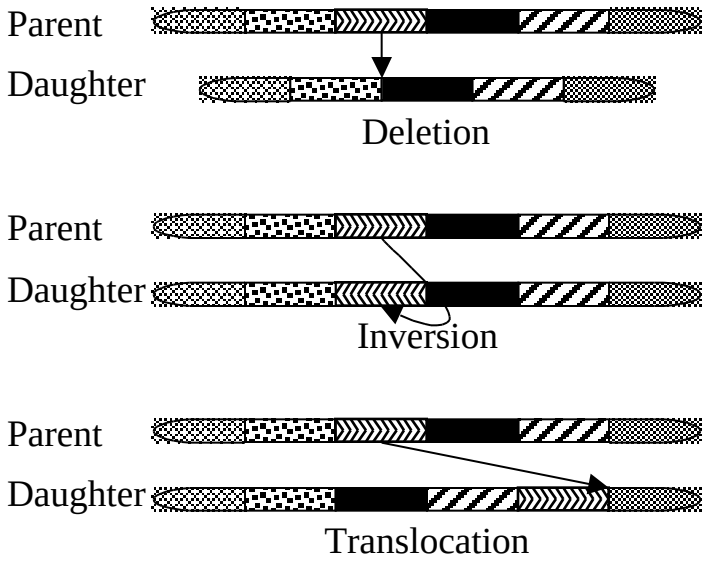


Figure 4: Deletions, inversions, and translocations

The lining up of the chromosomes is important, so that each of the resulting cells receives exactly one copy of each chromosome. In order for this to happen, the two homologous chromosomes have to identify similarities in each other. This process generally involves a “cross-over”, that is, the swapping of some DNA between the two homologous chromosomes, as shown in Figure 2.

This procedure is shown more fully in Figure 5. Two homologous chromosomes are shown, each containing five genes. (In reality, genes are spaced out rather sparcely on the chromosomes.) The two homologous genes are shown in similar shading, but the difference in thickness represents the fact that there may be some differences between these genes. In A), the crossover event occurs between genes 2 and 3, resulting in the offspring getting a chromosome with a set of genes not found in any chromosome of either parent. Occassionally an uneven crossover occurs, resulting in one chromosome with a deleted copy of a gene, and one with an extra copy. The offspring will then have a different number of such genes from either parent. Finally, there is the process of gene conversion, where the gene in one chromosome is replicated to be identical to the gene of its homologous partner. This often occurs with crossovers at this location.

Genomic mutations

It is possible for entire chromosomes to become duplicated, often as a result of errors in meiosis. Such mutations are generally lethal. The most common form in humans is Down's syndrome, where the individual receives an extra copy of Chromosome 21. It is not believed that chromosomal duplication has had a significant impact on genome size. It is also possible for the entire genome to be doubled. This is actually not rare in plants; many hybrid plants are tetraploidal. Following evolutionary divergence of the doubled chromosomes, it is possible for these organisms to evolve to a standard diploid configuration with a genome twice as large. It has been hypothesized that a number of genome doubling events were crucial at various stages of evolution. In general, these hypotheses are not supported by the observed genomes.

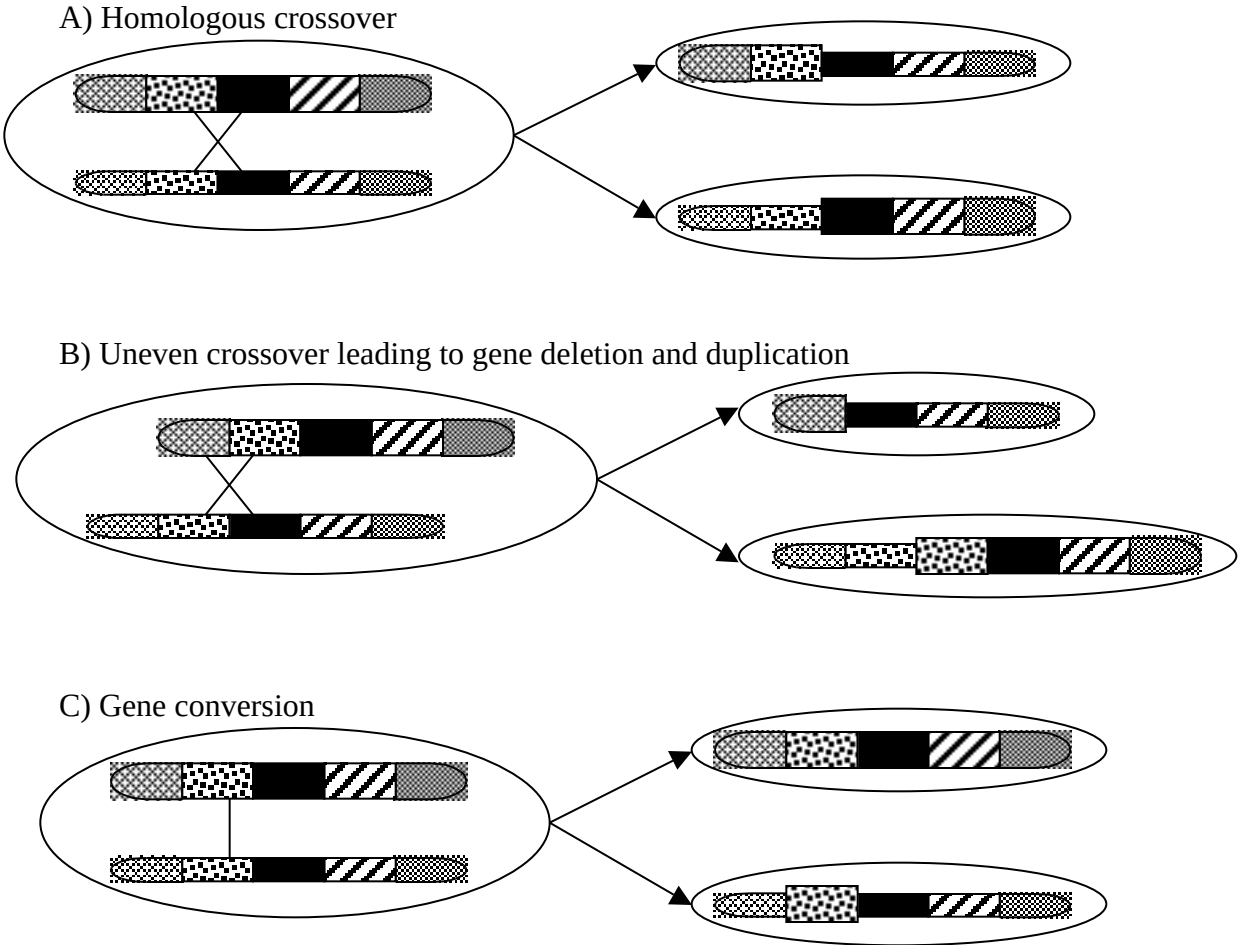


Figure 5: Crossover events

Fate of duplicated genes

A number of the processes described above results in an extra copy of a given gene, so-called gene duplication. There are a number of possible fates for such a gene. It may be that there is an advantage to the organism to having multiple copies, especially if the protein is needed in a large quantity. Multiple identical genes in the genome, for instance,

encode histones. Alternatively, the gene can be able to undergo further mutations until it has lost all function; mutations in the adjacent regulatory regions can cause this protein to be unexpressed. Such genes are called “pseudogenes”. Finally, the gene can evolve to fulfill a completely different function. The flexibility resulting from gene duplication may have had a major impact on evolution.

Classical Theory of Gene Dynamics

Consider a gene where there is a dominant variation or allele. Let us call this allele the “wild type”. A different version of the gene, a new allele, can result from random variation due to any of the previously-described errors in conservation or replication. There are three possible fates of this gene. The individual or the descendents of the individual can fail to reproduce in which case this different allele is removed from the population. Alternatively, the new allele can spread throughout the population so that it now becomes the new "wild-type". This process is called fixation. Prior to removal or fixation, when both versions of the allele are present in the population, the locus is said to be polymorphic. Under certain selective regimes called balanced polymorphism, the new allele neither dies out nor replaces the wild type, but rather is maintained at a particular frequency. Let us see how these alternatives are possible.

Consider a diploid organism, that is, with two copies of each gene, one of which is

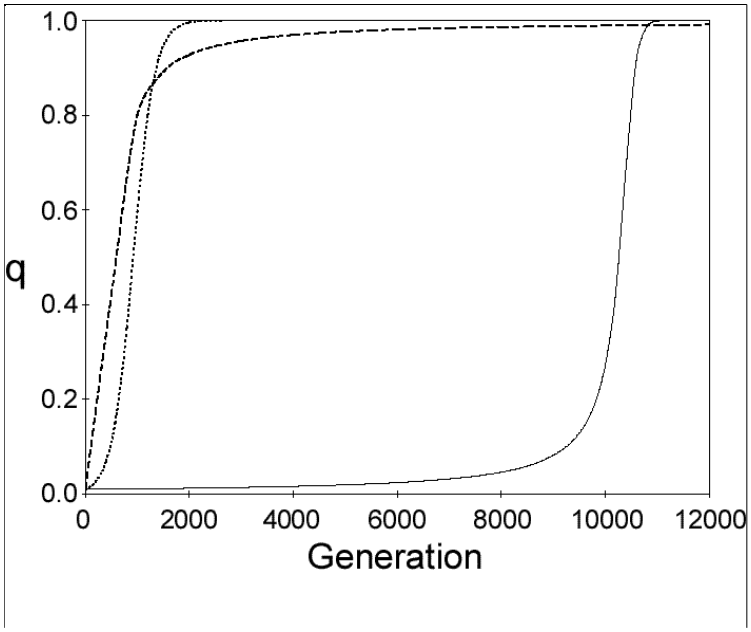


Figure 6: Gradual fixation of a advantageous mutation starting with an initial frequency of 0.01. The relative fitnesses of the various genotypes are s_{BB}=0.01 and either s_{AB}=0.00 (solid line, corresponding to a recessive mutation), s_{AB}=0.005 (dotted line), or s_{AB}=0.01 (dashed line, corresponding to a dominant mutation). The recessive mutation requires longer for fixation due to the slow buildup of BB homozygotes with a competitive advantage. Conversely, it is difficult to achieve total fixation of a dominant advantageous mutation because of the relative fitness of the heterozygotes.

transferred to the next generation. We will call the wild-type allele **A** and the mutant **B**. There are then three diploid genotypes possible, **AA**, **AB**, and **BB**. Remember that an organism is called a homozygote if both alleles are the same (**AA** or **BB** in this example); otherwise it is a heterozygote. If the fraction of the gene **A** in the population is *p* and the fraction of gene **B** in the population is *q* (so that $q = 1 - p$), then at equilibrium where these ratios are maintained and where mating is random the population will have genotype **AA** with frequency $p\_$, genotype **AB** with frequency $2pq$, and genotype **BB** with frequency $q\_$ - the Hardy-Weinberg equilibrium. Since Hardy-Weinberg equilibrium is achieved after one generation in a randomly mixing population, it will approximately hold even when these different genotypes correspond to different genotypes with different fitnesses, where the fitness measures the relative probability of contributing to the next generation.

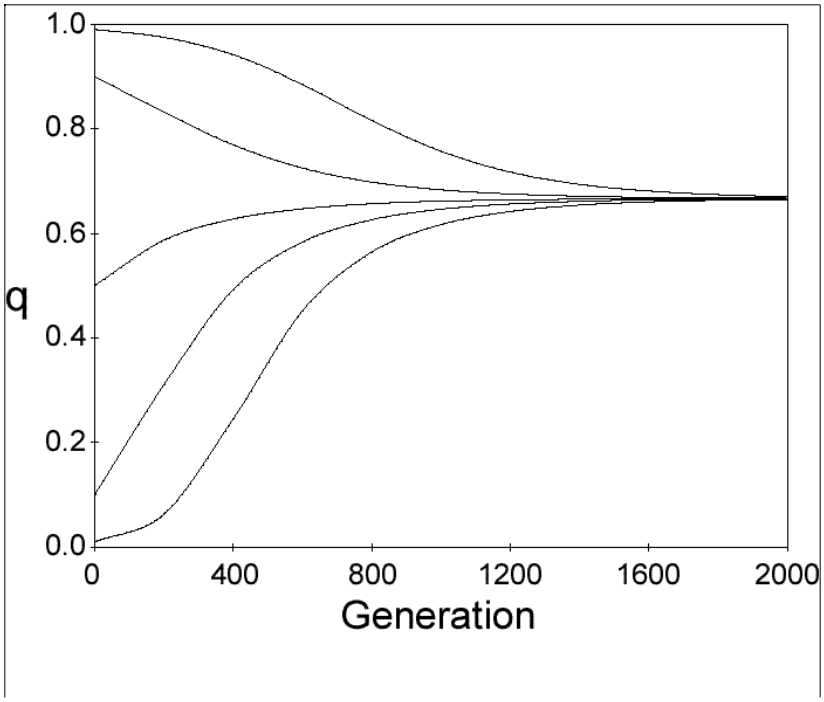


Figure 7: Population dynamics with overdominant selection where the heterozygote has increased fitness relative to either homozygote, for a range of initial gene frequencies. The relative fitnesses of the various genotypes are s_{AB}=0.02 and s_{BB}=0.01.

Let us consider the relative fitnesses of these three genotypes as $\_{AA}$, $\_{AB}$, and $\_{BB}$, respectively. Often we consider fitnesses relative to the wild type **AA**, so $\_{AA} = 1$, $\_{AB} = 1 + s_{AB}$, and $\_{BB} = 1 + s_{BB}$. For an infinite population and the values of *p* and *q* for any given generation we can calculate how much of each allele will be present in the next generation. *q*, the fraction of allele B in the next generation, will given by

$$q = q + \frac{pq(\omega_{AB} - \omega_{AA}) + q(\omega_{BB} - \omega_{BA})}{p^2\omega_{AA} + 2pq\omega_{AB} + q^2\omega_{BB}} \qquad [1]$$

Figures 6 and 7 show two different examples. In Figure 6, we consider the case where the new mutant gene is advantageous with s_{BB} = 0.01. The new allele **B** is fixed in the population with probability 1 with dynamics that depend upon the heterozygote fitness s_{AB}. In Figure 2 we have what is called overdominant selection where the heterozygote **AB** has the highest fitness: in this case s_{AB} = 0.02 and s_{BB} = 0.01. (The classical example of overdominance is the mutation for sickle-cell anemia, where the mutant homozygote (**BB**) is lethal, yet the heterozygote (**AB**) has increased resistance to malaria.) The result is relaxation to a constant steady-state population of some **B** for any initial population.

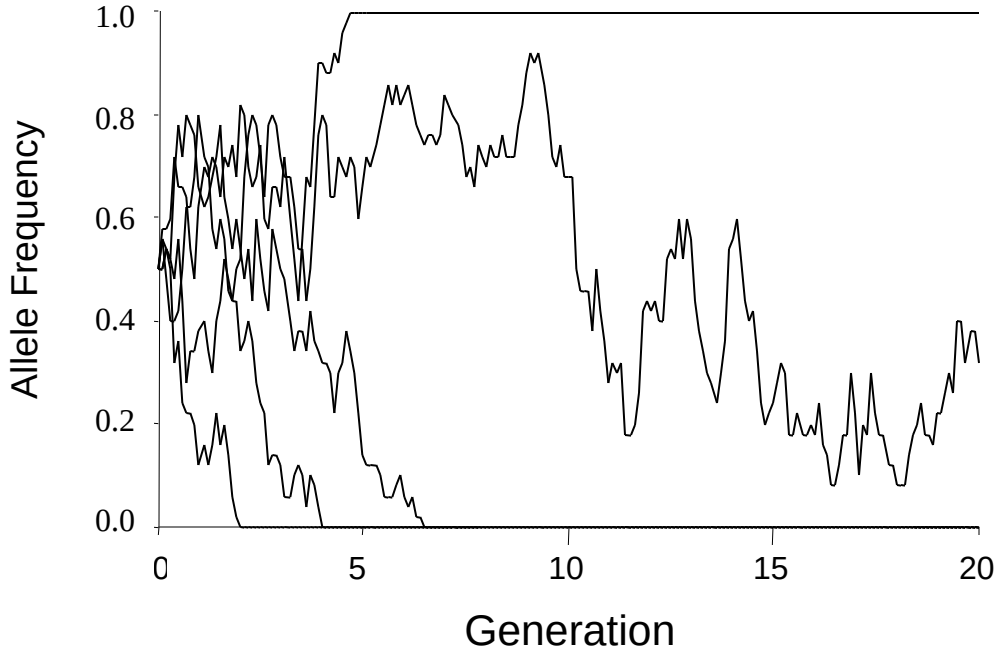


Figure 8: The results of five different runs of two populations of equal fitness, for a constant population of size 50. One allele is generally either fixed or eliminated.

Finite Populations and the Neutral Theory

The eventual fate of a mutant gene is deterministic if the population is infinite: there are differential equations that can be solved to give the behavior described above. In reality, populations are quantized and finite. This can have a strong impact on the fate of various alleles. For instance, imagine that neither allele has a selective advantage so that the fitnesses of all three possible genotypes are equal. In this case, the population dynamics of infinite populations would result in the proportion of the two alleles remaining constant in a Hardy-Weinberg equilibrium. In a finite population eventually either the mutant allele would be eliminated or achieve fixation, at least if additional copies of the mutant allele are not produced by further mutations. This is because random fluctuations in the allele fractions would occur; with a finite probability that any allele frequency would decrease to zero from where it cannot recover without further mutations of the

wild type. This process is illustrated (for an extremely small population) in Figure 8. The discreteness of the population is key to this process, as there have to be an integral number of copies of each allele. Consider a single copy of a new allele in a population of size N (so that $q = 1/2N$, with the $2N$ coming from the fact that the individuals are diploid). In the simplest case the heterozygote has the average fitness of the two homozygotes so we can write $w_{AA} = 1$, $w_{AB} = 1 + s$, and $w_{BB} = 1 + 2s$. Kimura derived the probability of eventual fixation of **B**

$$P_{\text{fixation}} = \frac{1 - \exp(-2N_e s / N)}{1 - \exp(-4N_e s)} \qquad [2]$$

where N_e is the effective population size, that is, the population that is actually reproducing at any one time. (For human populations, $N_e \sim N/3$.) Ignoring the difference between N and N_e results in curves of P_{fixation} as a function of s for different population sizes as shown in Figure 9. As would be expected, for a neutral mutation ($s = 0$) the probability of eventual fixation represents the initial fraction of the population, $P_{\text{fixation}} = 1/2N$. In fact, all mutations with values of $|s| < 1/2N$ have approximately probability $1/2N$ of fixation; these mutations are essentially neutral. For larger values of s, the probability of fixation becomes $P_{\text{fixation}} = 2s$. For finite populations there is a chance of a negative mutation becoming fixed in the population, just as there is a chance of a positive mutation being removed.

In classical evolutionary theory, the process of evolutionary change was dominated by chance advantageous mutations that became fixed, what is called adaptive evolution.

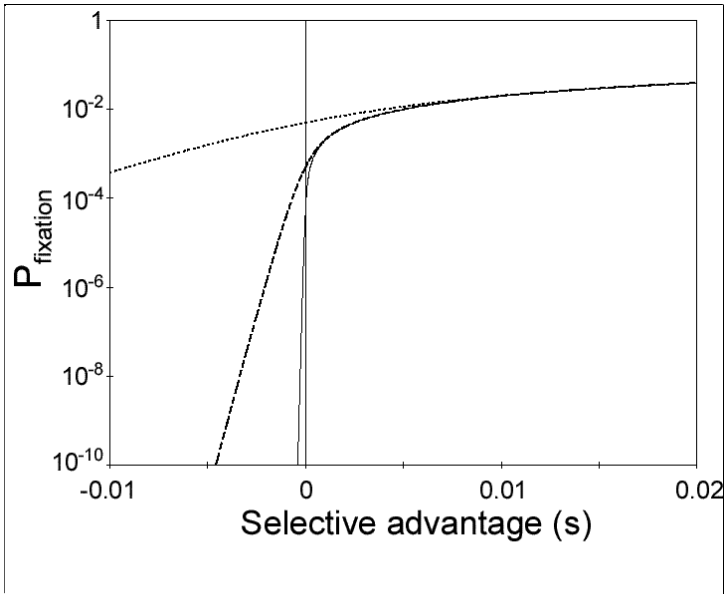


Figure 9: Probability of fixation of a mutation with initial frequency 1/2N as a function of s for various population sizes: N=100 (dotted line), N=1000 (dashed line), and N=10000 (solid line). The difference between population size and effective population size is ignored.

Kimura and Jukes and King proposed the neutral theory, which postulated that the vast majority possible mutations are either deleterious or neutral ($|s| < 1/2N$). As the deleterious mutations will generally be removed from the population by purifying selection, most observed substitutions would be neutral or slightly deleterious. This was used to explain four observations. One observation had to do with the large variation in genotypes in a typical population. It was observed that many genes were polymorphic, that is, multiple alleles exist in the population. In the classical theory, polymorphism could result from overdominant selection (as shown in Figure 7) or from frequency-dependent selection - where there was an advantage to being different from others in the population. In these cases there was selective pressure towards polymorphism. According to the neutral theory, polymorphism was generally a temporary result of a nearly-neutral mutation that had not yet been either eliminated nor fixed. Fixation and elimination times for neutral mutations are quite long, about $4N$ generations. Under these conditions it would be natural to have a large amount of polymorphism in the population. Kimura claimed that the amount of polymorphism in observed populations was too large to be explained by positive adaptation.

A second observation was that any particular gene often tends to evolve at a roughly constant rate in different organisms, that there is a molecular clock. (This is not to say that different genes evolve at the same rate - there are quite large variations in the rate of substitution of different genes.) Figure 10, for instance, shows the relationship between the time of divergence of various species from humans according to the fossil record compared with the dissimilarity in the sequences of alpha-Hemoglobin; the near straight-line behavior across such different organisms (with the exception of chicken) is striking. This molecular clock is a natural result of the neutral theory. Neutral mutations in a population should arise at a rate proportional to the number of genes in the population,

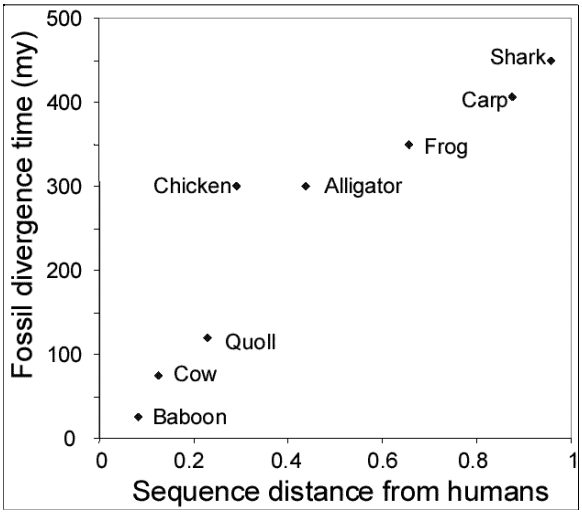


Figure 10: Evidence for a molecular clock. Plot shows the relative time of divergence from man according to the fossil record compared with the sequence divergence. Adapted from Volkenstein 1994.

2_*N*. The probability of fixation is 1/2*N*. Multiplying these two factors together, the total rate of introduction and fixation of neutral mutations should be _, independent of the population size. This would explain the constant rate of genetic evolutionary change. Conversely, for adaptive change the fixation probability is 2*s*, resulting in a total rate of introduction and fixation of 4_*Ns*, proportional to the population size. This would be incompatible with the molecular clock hypothesis.

A third observation was that important genes evolve slower than less-important genes, which evolve slower than regions of the genome that do not seem to serve any purpose. If most mutations were either neutral or disadvantageous, the importance of the gene will correlate with the likelihood that changing the gene will be deleterious. As a result, mutations in less-significant regions of the genome will have a smaller probability of being removed by purifying selection, and thus more chance at fixation.

The last observation has come with the rise of genetic observation and manipulation. Many substitutions at the DNA level do not change the amino acids of expressed proteins - these are called synonymous substitutions. It is likely that the vast majority of these substitutions are essentially neutral. Additionally, is now clear how plastic the resulting amino acid sequences are, that it is not difficult to find many amino acid substitutions that results in proteins with seemingly identical properties. Many of the changes that we can make in the lab seem to be neutral, at least within the accuracy and the context of the experiments.

The neutral theory does not downplay the role of adaptive change. Obviously the characteristics of living creatures show that we are highly adapted to our surroundings. The argument is that adaptive changes, though important, are extremely rare and that most substitutions are either neutral or deleterious. As will be discussed below, one reason for the presence of neutral change at the genetic level is the many-to-one mapping of genotype to phenotype.

One additional concept introduced by Stephen Jay Gould is the idea of the spandrel. According to Gould, certain features of an organism might arise from reasons having nothing to do with selection, either through random neutral drift or as an unavoidable consequence of some other modification. The organism might then be able to adapt this feature for adaptive purpose. Accordingly, it is dangerous to explain the existence of this feature as arising as a positive adaptation towards its eventual purpose, what he called the "Panglossian paradigm" after the character in Voltaire's Candide who sees everything as optimal in this "best of all possible worlds".

Is the neutral hypothesis correct? This is still a topic of much debate. The degree of polymorphism seems to be somewhat between the rate predicted by adaptionists and neutralists. While there seems to be some degree of constancy to the rate of evolution of each gene across different evolutionary lines, the molecular clock seems to run somewhat erratically. Neutralists claim that these irregularities can be explained by accounting for different rates of mutation, different generational times, and some adaptive bursts. Adaptionists can also explain why important regions of the genome evolve slower than

less-important regions: genetic changes in the less-important regions of the genome result in smaller changes in fitness, and smaller changes in fitness are more likely to be advantageous than a larger changes. While it is true that it is not difficult to make genetic changes that have no observable effect on the phenotype or on the organisms chance for survival and reproduction, effects too small to be seen in the lab may still have a large impact on the evolutionary process. Even synonymous substitutions in the DNA code that have no effect on the amino acid sequence of the expressed proteins can affect the probability of survival by affecting the rate of protein transcription.

Eigen's theory of quasi-species

Kimura's neutral theory includes aspects of random change due to the finite size of populations, resulting in stochastic effects that are not included in the infinite-population differential equations. But this model still involves a homogeneous population in which mutations occurs. The polymorphism is a result of the evolutionary dynamics, rather than a critical component. These aspects were to change with the introduction of the theory of quasi-species.

It is easiest to consider the situation by considering a fitness landscape, a term first introduced by Sorvell Wright. The fitness landscape consists of a sequence space, that is, the space of all possible sequences, combined with a fitness value for each point. Each dimension in the fitness landscape corresponds to one allele or base or amino acid. As a result, the sequence space (as well as the fitness landscape) is extremely high-dimensional. It is a strange space, however, in that only relatively few discrete values in each dimension are allowed. For proteins, for example, each dimension consists of twenty discrete points representing the twenty amino acids. A typical simple fitness landscape for a trimer in a two-letter code (A and B) is shown in figure 11. Again, this diagram does not do justice to the high dimensionality of the space. Another useful but misleading representation is where the discrete nature of the sequence alphabet is ignored, resulting in diagrams such as figure 12. The advantage of this representation is that it provides an intuitive idea of fitness peaks, valleys, and ridges.

Eigen modeled evolution by considering a flow reactor containing self-replicating molecules of RNA, with an influx of mononucleotides and an outflow to keep the

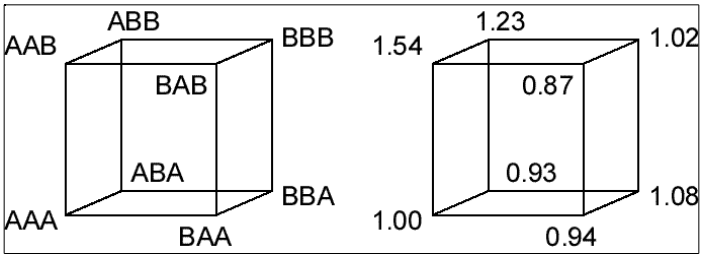


Figure 11: An example of a fitness landscape of a trimer written in a two-letter alphabet. The left of the diagram shows the available sequences and their connectivities, while the right of the diagram shows the corresponding fitnesses.

concentration constant. As with biological RNA, the replication rate is not perfect, but mutations naturally arise. The dynamics can be treated with standard approaches from chemical engineering. The result indicated that rather than having a single "wild-type" genotype, instead there would be a cloud of different genotypes centered in the sequence space around a prototypical sequence. The cloud could represent the ultimate steady-state solution -- not a transient phenomenon eventually resolved by natural selection. It was then possible to talk about the evolutionary process in terms of changes and competition between these various clouds, which took on the role of species in classical evolutionary theory and thus were named quasi-species. Eigen's results demonstrated that in order to consider the fitness of any particular quasi-species, it was necessary to consider the fitness of the prototypical sequence as well as the surrounding sequences. A broader, flatter, but lower fitness peak could out-compete a sharper, narrower albeit higher fitness peak, depending upon the overall mutation rate. Finally, there could be qualitative changes in the evolutionary process brought on by quantitative changes in the mutation rate. Specifically, there was a certain critical mutation rate above which the evolutionary dynamics became random and incoherent, with a loss of the genetic information.

Again things change when we consider finite populations, due to the resulting stochastic nature of the evolution as well as the discrete nature of the individuals. On a flat fitness landscape an infinite population would expand to fill the available space. In reality, the cloud of members retains its cohesiveness. The edge of the cloud is characterized by a dilute population of members. Such members are highly unstable with respect to evolution, in that any fluctuations in population that take their number down to zero results in an extinction of this subpopulation from which it cannot recover. The center of the cloud is more resistant to these fluctuations. The result is that there is a tendency to eliminate the outlying members of the population cloud, so that the cloud remains centered on the prototypical sequence. The resulting cloud can then wander in a stochastic manner in the fitness landscape.

One of the more important results of this approach towards evolution is the dependence of the evolutionary process on the fitness landscape. Certain characteristics of the landscape have been especially emphasized by Schuster, Fontana, and their co-workers. They investigated the fitness landscape for RNA molecules, taking advantage of rapid

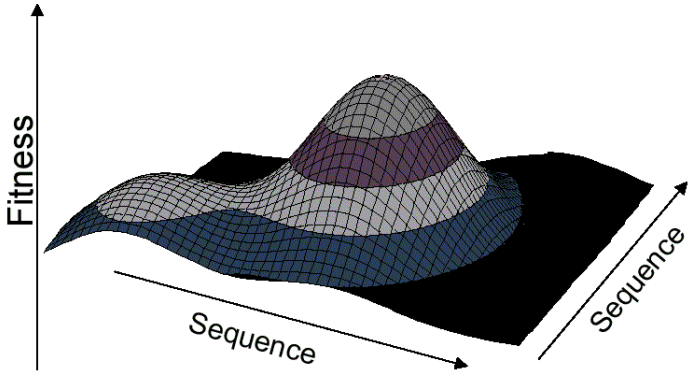


Figure 12: Fitness landscape with sequences represented as continuous variables.

algorithms for computing the ground state conformation. They found that these molecules had a large degree of neutrality, that is, it was possible to make large changes in the sequence while retaining the same structure. Considering the sequence as genotype and the structure as phenotype, there would be many changes in genotype consistent with a single phenotype. The many-to-few sequence to structure mapping results in large "neutral networks", that is, regions in the sequence space with constant fitness.

The dynamics of the population cloud combined with the large neutral networks can have a large impact on the evolutionary dynamics. The sequence cloud is free to randomly sample the neutral network. The individuals on the tail of the cloud allow the population to sample the fitness landscape at some distance from the prototypical sequence in a large number of different directions. If the tail of this distribution overlaps a region of higher fitness, the whole cloud can adapt in this direction with a small change in sequence.. Note that the resulting dynamics (long periods of neutral evolution, punctuated by rare but rapid adaptive change) is exactly what is described by Kimura's neutral theory.

In this model, the properties of the intersection points between various neutral networks become critical. One important property of the fitness landscape is how close the various neutral networks were to each other. For RNA, it is possible to go from one native structure to almost any reasonably common different structure with only a small change of the sequence, a phenomenon known as shape-space covering. As a result RNA can evolve quickly for different structures and possibly different functions. This phenomenon seems not to be true of theoretical models of proteins, in that it is more difficult to go from one structure to another. As a result, evolutionarily-related proteins tend to have the same structure, something called "structural inertia". This may be due to the larger number of amino acids compared with the number of RNA bases, or the relatively small number of sequences that will form a viable protein in any structure.

To summarize the previous section:

- We must include the role of stochastic effects resulting from finite population sizes.
- We should be aware of the presence and effect of neutrality in the fitness landscape. One reason for this neutrality is the many-to-few mapping of genotype to phenotype.
- The properties of the fitness landscape such as the size, distribution, and connectiveness of the neutral networks can have a major effect on evolutionary dynamics.
- Techniques drawn from statistical physics are useful in understanding evolution; it is a natural approach to dealing with the general properties of a large number of individuals behaving stochastically.
- Selective pressure does not necessarily result in the organism selecting the highest peak on the fitness landscape. It is not correct to equate evolution with optimization, or even adaptation.

- Even if a feature fulfills some important function, we cannot conclude that feature evolved in an adaptive way to fulfill this function. These features may represent spandrels.

Modeling evolutionary change

The rate of evolutionary change in DNA depends on the nature of the change as well as its location in the protein. Those regions that encode proteins change more slowly than so-called "junk" DNA. Synonymous changes occur more frequently than non-synonymous changes, and changes occur to important proteins more slowly than to less critical proteins. Non-coding regions change at the rate of 3×10^{-9} substitutions per site per year. Synonymous substitutions in coding regions occur at a similar rate. Non-synonymous substitutions vary greatly, from approximately zero in histones to 3.06×10^{-9} in Relaxin. There are some examples where the rate of non-synonymous substitutions is actually larger than the rate of synonymous substitutions. This is generally seen as evidence of adaptive evolution.

The simples possible model for DNA change is to assume that all substitutions occur at an equal rate, α, the model developed by Jukes and Cantor. For four equally-likely bases, it is straightforward to write down expressions for $P_{xx}(t)$, the probability that a base x at time 0 is still x an evolutionary time t later, as well as $P_{xy}(t)$, the probability that a base x has been replaced by a base y.

$$\begin{aligned} P_{xx(t)} &= \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \\ P_{xy(t)} &= \frac{1}{4}\left(1 - e^{-4\alpha t}\right) \end{aligned} \qquad [3]$$

As described above, transitions and transversions generally occur at different rates. Kimura (1980) developed a slightly more complicated model, where the rate of transitions is α, while transversions occur at a different rate, β.

From \ To	*A*	*C*	*G*	*T*
A	–	β	α	β
C	β	–	β	α
G	α	β	–	β
T	β	α	β	–

[4]

In this case, the substitution probabilities at any time t are given by

$$P_{xx(t)} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$$

$$P_{xy(t)} = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{transitions} \\ \frac{1}{4}\left(1 - e^{-4\beta t}\right) & \text{transversions} \end{cases} \qquad [5]$$

At the same time, Felsenstein developed a model that accounted for varying base frequencies, π_x (where x is A, G, C, or T) by making the substitution rate proportional to the base frequencies. The Kimura and Felsenstein models were soon merged by both Felsenstein and Hasegawa, Kishino, and Yano in what is commonly called the HKY model. In this model, the substitution rate is proportional to both a rate parameter (e.g, β), and the equilibrium frequency of the base that is changed to. This model shares a useful feature with many other models of the evolutionary process, which is that it is reversible and obeys what physicists call "detailed balance", that is, the flux from x to y is equal to the flux from y to x. In this case the composition of the four bases are constant in time, with the extra constraint that their frequencies sum to one. The maximal extension of this kind of model is what is termed the general time-reversible model (GTR), in which there are six free rate parameters, λ_{xy}, controlling the rate of exchange in both directions between each pair of different nucleotides, x and y. If the reversibility criterion is relaxed such that $\lambda_{xy} \pi \lambda_{xy}$, the number of rate parameters doubles to 12, but this makes the calculations more difficult and is not commonly seen as a useful modification.

A more fruitful and necessary development was to consider that the average substitution rate might vary amongst sites. This is often accounted for by assuming that the rates are gamma distributed (Figure 13), a model first successfully implemented by Yang. In these models, the equilibrium frequencies and relationships between the rates within the model remain constant, but the rates of each site relative to other sites are allowed to vary. The gamma distribution, chosen for its flexibility rather than any underlying biological motivation, is modeled by an approximation where average rates are calculated for discrete segments, often only four. Sites can also be divided into pre-specified categories with different average rates, and this is a common first-pass approach for dealing with the different codon positions, which are known to change at different rates according to their probability of changing an amino acid, and the probability that the changed amino acid has very different physicochemical properties.

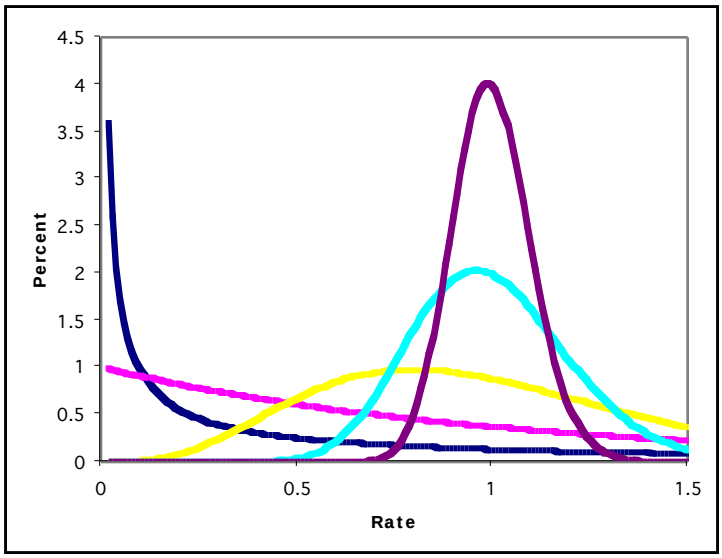


Figure 13. Five gamma distributions at the same scale, with shape parameters equal to 0.2, 1, 5, 25, and 100.

For both the single rate and the gamma models, the most complex model for which analytical transition equations (such as Equations 3 and 4) exist is the Tamura-Nei model, which is a slight extension of the HKY model to allow different transition rates between purines and between pyrimidines. These equations are often inverted to estimate the number of substitutions between sequences, molecular distances, from the observed base frequencies and base substitutions, P_{xy}. With Kimura's model, for example, if the proportion of observed transitions (P_{xy}, where x and y are in the same nucleotide class) is P, and the proportion of observed transversions (P_{xy}, where x is a purine and y is a pyrimidine, or vice versa) is Q, then the molecular distance estimate for the number of substitutions ($2\alpha t + 4\beta t$) is

$$\hat{D} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right) \qquad [6]$$

The accuracy of these distances can be assessed using analytical estimates of the variance of these estimators, which can also be calculated directly. An interesting phenomenon with distances from models more complicated than Jukes and Cantor is that when accuracy is measured by mean square error (MSE=variance plus square of the bias) or by utility in tree reconstruction, the appropriate analytical distance measures are often less accurate than distances based on simpler models. In the realistic case where the transition rate is considerably higher than the transversion rate, there is a large range of times where the transition parameter in Equation 5 cannot be estimated accurately, and its variance is so large that it dominates the variance of the entire estimator. The solution is to estimate the transition and transversion components separately and get an average estimate of their ratio from many sequences. This ratio can then be used to recombine the components with generalized least squares, resulting in a distance that is both more accurate and less biased than distances based on simpler models. If computational time is not critical,

accurate distances can also be estimated using maximum likelihood (see below), but it is still essential to obtain relative rate ratios from multiple sequence comparisons. The simple distance equations should not be used.

As the substitution models become more complicated, the computational time required for maximum likelihood (ML) calculations increases roughly as a multiple of the number of parameters. This is because, by definition, ML involves finding the maximum of the likelihood function, which means finding the set of *all* the parameter values that has the highest probability of having produced the sequence data. Furthermore, with smaller data sets there is a definite risk of over-parameterizing the model. Since many of the models described above are nested, it is useful to compare the difference in the log likelihoods of each model when the parameters are maximized with respect to the likelihood function. Twice the cumulative density function (cdf) of this statistic can be approximated by the chi-square distribution χ^2_η, where η is the number of degrees of freedom separating the two nested models (that is, the difference in the number of parameters between them). If the chi-square assumption is in question, the cdf can be determined by parametric bootstrapping, which involves numerical simulation of the simpler model using the maximum likelihood estimators (MLEs) to see how large the log likelihood differences between the models would be due to chance alone.

In order to accommodate amino acid substitution, the substitution model must expand to 20 states at a minimum. In a general reversible model, there would be 219 free parameters, which would lead to dramatic over-parameterization in most data sets if they were allowed to freely optimize. Instead, amino acid substitution models are often generated by counting substitutions in a large number of sequence comparisons, sometimes including all proteins in the available database, with the counts adjusted for the total number of counts observed in each individual protein in the database, which accounts for varying mutability between proteins. The original such mutation data matrix (MDM) was Dayhoff's point adjusted mutation (PAM) matrix, which was calculated from simple trees and parsimonious reconstruction of ancestral states. A simpler pairwise approach was utilized by Jones, Taylor, and Thornton to automatically create a matrix from the by then dramatically larger sequence databases, and Koshi and Goldstein showed that these matrices could be more accurately reconstructed using maximum likelihood techniques that integrate over all ancestral states, albeit with a considerable computational burden. MDMs have at this time been created from many, if not most, conceivable subsets of the protein databases, with the most important modifications being segregation by rate of substitution and by secondary and tertiary protein structure. The BLOSUM matrices are created using only more divergent proteins, and are preferred for use in detecting remote homology relationships when scanning databases. Structural categories seen to have important differences in their substitution matrices include alpha helices, beta sheets, and loose coils, internal and external sites, and transmembrane regions.

It is also possible to model evolution of codons using a 64 x 64 transition matrix. A general reversible model of such a matrix would have 2016 transition parameters in addition to 63 free equilibrium frequencies, which is at this time far too many, and such a model has not been utilized. Rather, codon matrices are generally constructed using a

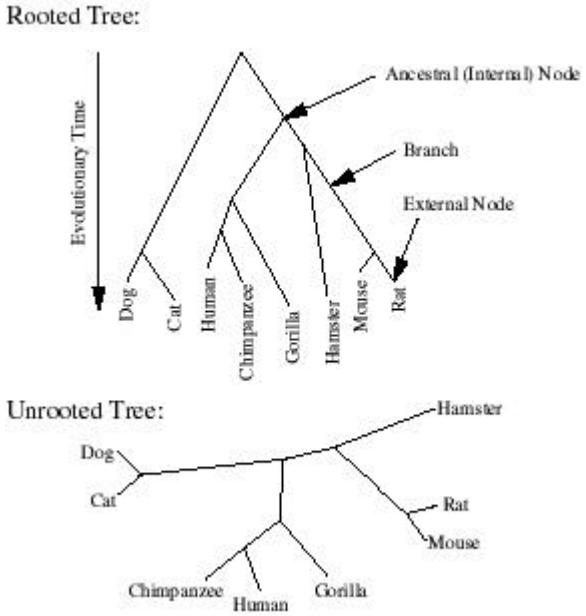
single nucleotide model for synonymous changes, and overlaying a matrix of amino acid substitution probabilities. Such models assume that there is no effect of the codon on these two models, that is, that amino acid substitution probabilities are not affected by the codons as long as the substituted codon is directly accessible, and that synonymous nucleotide substitution probabilities are not affected by the codon they are in. The former assumption is intuitively reasonable, but the latter assumption is known to be demonstrably wrong: many cases of "codon bias" have been documented, particularly in highly expressed proteins. Thus, this assumption is likely to be dropped in future models for large data sets as computational power increases. Another feature of the way these matrices are constructed from compound amino acid and nucleotide substitution matrices is that only single substitutions between codons are considered. This assumption also appears reasonable, except that in population processes, double substitutions can occur at rates that are faster than the simple multiple of their individual probabilities, particularly when selection is involved. Deleterious but recessive alleles can rise to reasonably high numbers in large populations, and so stand a reasonable chance of mutating to a second allele that is no longer deleterious, or even advantageous. This sort of evolutionary tunneling has been shown to occur in paired sites in RNA helical structures, and it is reasonable to infer that it may also occur in proteins.

Rather than use a pre-specified MDM, it is also possible to create an MDM based on some function of the physicochemical properties of the amino acids. Such functions have been created for use in codon models by Yang, for site-specific amino acid frequency models by Bruno and Halpern, and for the k-class models used by Goldstein and colleagues (see below). These functions are sometimes more or less loosely based on principles of fitness and the probability of substituting less fit amino acids, but they are also used similarly to the gamma model, as convenient functions that can be tested and optimized to see what is the best way to match the data without adding too many extra parameters.

The k-class models are probably the most notable recent addition to protein evolutionary models, in that they are the first to allow for multiple classes of distinct rate patterns at different sites. They also allow the data at each site to determine which rate class is most appropriate by integrating over all rate categories at each site, in a similar fashion to the way gamma models are implemented. This type of model has not yet been implemented for DNA evolution, but promises to be useful enough that implementation for DNA is certain to be forthcoming in some standard programs in the near future. A very recent innovation is the use of empirical "energy potentials" derived from three-dimensional structures. These have been used in the context of a sequence threaded onto a known structure to create substitution probabilities that are dependent on the local residues in existence at the time the substitution occurs.

Probability Calculations on Phylogenetic Trees

Since sequences are not independent realizations of a particular protein, but rather are related by the gene duplication and serial speciation events that have occurred in their history, any calculations beyond a first pass approximation must take these relationships into account. A statement of these relationships is a phylogenetic tree, which includes information on the relative order in which the gene duplication and speciation events occurred, and also the amount of time (or probable number of substitutions) separating different events. In a rooted tree (Figure 14), the root or top of the tree is the deepest point, and is considered the ancestor of all the other sequences. Evolutionary time proceeds with movement along the tree in a downward direction from the root, until the tips of the tree are reached, which represent the descendent sequences in the sample. For purposes of calculation, and because gene duplication is a binary splitting process, phylogenetic trees are usually represented as purely bifurcating trees, without any higher multifurcations. Since two genes never become one, there are also no networks, although



attempts to represent recombination can make these graphs considerably more complicated. When the process is viewed moving backwards in time, the lineages can be said to coalesce, although that terminology is more commonly used within populations. Oftentimes, determination of the root of a tree is one of the more problematic possible inferences, and for reversible models the placement of the root does not affect probability calculations. Therefore, it is common to represent the tree as an unrooted tree (Figure 14), where no statement is made as to which point is ancestral to all other points.

The lines between gene duplication are called branches (equivalent to edges in graph theory), while the points where gene duplication occurs are called nodes (vertices). The tips where the sequences reside are sometimes called external nodes, and are connected to the rest of the tree by external branches, whereas the remaining branches are called internal branches, and the remaining nodes are internal nodes. For any unrooted tree with N sequences, there are 2N-3 branches, N of which are external, and N-3 of which are internal. There are N-2 internal nodes. The order in which the branches occur,

irrespective of the branch lengths, is called the topology of the tree, and the number of possible unrooted topologies relating a set of sequences rises extremely quickly as

$$\prod_{i=3}^{N} 2i - 5 \qquad [7]$$

The topology and branch lengths of a phylogenetic tree may be considered parameters of a model that includes the phylogenetic tree. Various values for these parameters need to be considered in order to find the most plausible tree or set of trees given the data, but for a particular tree with specified topology and branch lengths, and the particulars of the model of substitution, such as the equilibrium frequencies, substitution probabilities, and any other parameters germane to the model, it is possible to calculate probability directly.

Substitution probabilities for any time t (or branch length, b) are generally calculated by integrating over all possible paths that might have led to the substitution of one nucleotide, amino acid, or codon, for another. Since general equations for these probabilities do not exist except for the simpler DNA models, these probabilities are determined by taking the exponential of the instantaneous transition probability matrix, e^{Qt}, which is defined as $Se^{-\Lambda t}S^{-1}$, where S, Λ, and S^{-1} are the matrix of eigenvectors, the corresponding diagonal matrix of eigenvalues, and the inverse of S, respectively.

For a reversible model, calculation of the likelihood or probability of a phylogenetic tree given the data may begin at any point on the tree, which will be considered the starting or root node, R, for calculations. Assuming independence between sites, the likelihood for the entire data set is the multiplicative sum of the likelihoods at each site, and the likelihood at each site is the sum of the likelihood of each of the possible states multiplied by the equilibrium frequency of those states, or

$$L = \prod_{l=i}^{L} \sum_{k=1}^{K} \pi_k P(D_i \mid R_i = k), \qquad [8]$$

where L is the length of the sequence alignment, K is the number of states, and i and k are the site- and state-specific subscripts, respectively. D_i is the sequence data at site i in the alignment. The likelihood of each state, k, is the multiple of the probability of producing the data on each branch leading away from the root node, and these branch-specific probabilities, or fractional likelihoods, are simply the probability of the data below the descendant node times the probability of changing from state k to state j along a branch of length b, or

$$P(D_i \mid R_i = k) = \prod_{b}^{B_R} \sum_{j}^{K} P_{kj}(b) P(D_{i,N} \mid N_i = j), \qquad [9]$$

where B_R is the number of branches leading away from the root, N is the descendant node on the other end of the branch, and $D_{i,N}$ is the subset of data at site i that includes only sequences ancestral to the descendant, N. Similar fractional likelihoods can be calculated for each descendant node, N, considering only the branches leading away from the root,

and so on iteratively until a tip is reached, at which point the probability of the observed state is one and any other state is zero. Computationally, these calculations are made most efficiently by beginning at the tips and proceeding iteratively upward through the tree to the root.

Another criterion for comparing phylogenetic trees is the method of maximum parsimony, in which the minimum number of changes possible is calculated for each topology. This method does not take branch lengths into account, nor is it based on an explicit model of sequence evolution, although different changes can be weighted differently in the cost function. Changes are determined by making optimal character state assignments (the ancestral states) at each internal node. Parsimony is computationally much faster than likelihood calculations, and was important in the recent past when computers were slower. It is less efficient at obtaining the correct answer than likelihood, however, and is subject to a variety of biases, including “long-branch attraction”, that make it less useful than likelihood. In addition, it lacks the power of maximum likelihood and distance techniques to utilize and evaluate complicated evolutionary models, and is less efficient at using additional sequence information to increase the accuracy of topology reconstruction. For these reasons, parsimony is used less and less as a primary means of evaluating topologies, although it remains popular for analysis of its mathematical properties due to the simplicity of its calculations.

Searching Topology Space

Topology space, the set of all possible topological arrangements of phylogenetic trees, becomes very, very large for even moderate number of taxa. For likelihood, parsimony, and other criterion-based phylogenetic methods, it is therefore not possible to exhaustively enumerate all topologies for all but the smallest datasets. For slightly bigger datasets, the method of “branch and bounds” can be used, in which a pretty good tree is found as quickly as possible. Topologies are then constructed by iterative hierarchical addition of taxa, but whenever a topological route becomes worse than the best current topology, it is abandoned, since the final topology (with a complete set of taxa) is guaranteed to be worse still. For moderate or large numbers of taxa, heuristic methods to move through topology space have been developed by trial and error, with the hope being that topology space is smooth enough with respect to the search algorithm that all reasonably probable topologies are visited, and in particular that no optimal topologies are missed. The first step is to construct a reasonably good tree, and this is often done by “stepwise addition” of taxa, where at each step a new taxon is added at the best spot on the tree given the topology for the previous subset of taxa. This continues until all the taxa are added. An alternative is “star decomposition” (Figure 15), in which the taxa are all clustered together at a single point in the starting star phylogeny, and then branches are added to separate clusters into smaller clusters, until the tree is entirely bifurcating.

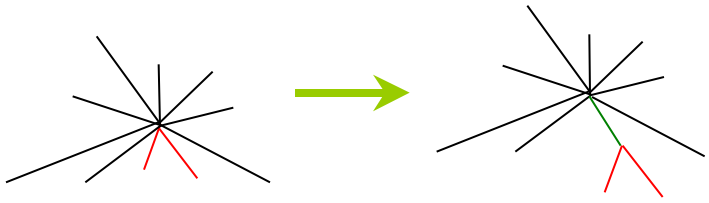


Figure 15. Star decomposition

Once an initial tree is constructed, topology space is generally traversed by some form of branch swapping. The simplest procedure, which happens to work quite well, is “nearest neighbor interchange”, where an internal branch is collapsed to a single node with four branches emanating from it. A new branch is then added to separate these branches again in any one of the three possible arrangements. Other procedures include “pruning and regrafting”, where a branch leading to a subtree is cut off the main tree and this branch then randomly reattached to some other branch in the tree, and “tree bisection and reconnection”, where the tree is split into two subtrees that are then reattached by creating a branch between randomly chosen branches in each tree. Despite the difficulty of the problem, and the fact that “valleys” of suboptimal trees are sometimes encountered that separate peaks in topology space, these methods often work surprisingly well.

When molecular distances are used, it is possible to develop a criterion for tree comparison such as “minimum evolution”, where the tree is most preferred that can minimize the difference between distances between species calculated on the tree and the starting distances. It is more common (and much faster) to use a heuristic tree reconstruction method to get a single tree estimate. The oldest and most obvious way to do this is the unweighted pair group method using arithmetic averages (UPGMA), whereby the taxa with the closest pair of distances are merged first, followed by the next pair, and continuing until all groups are joined into a single tree. This method. UPGMA does not take rate variation among branches into accounts, and can therefore be deterministically wrong when such rate variation exists. For this reason, it is generally preferred to use neighbor joining (NJ) or its derivatives. NJ also tries to join nearby taxa, but adds the criterion that the pair should be farthest from all the other taxa. NJ has some biases of its own, including the fact that results can deteriorate as distant taxa are added. Recent advances have corrected some of these problems by using the variance of distance estimates as a weight for both placing the node when taxa are joined (BIONJ) and choosing which nodes to cluster (Weighbor).

An additional heuristic clustering class of methods are the quartet-puzzling methods. It has already been mentioned that maximum likelihood can be used to obtain distance estimates between each pair of species, and these distances can then be used to create a tree using e.g., NJ. Similarly, in quartet puzzling the optimal topology and set of branch lengths is found for each set of four species, or quartet. An algorithm is then used to sort through or “puzzle” these quartets to find the topology that contradicts as few of these quartets as possible. This method is slower than distance methods and faster than a complete maximum likelihood solution, but is also less accurate than a maximum likelihood solution, and it is uncertain if it is more accurate than likelihood-based distance methods. Some recent studies also indicate that it may be vastly overconfident in its assessment of whether it has gotten the right answer.

Confidence Limits

As with any statistical approach, both parametric and non-parametric methods can be applied to assess confidence in estimates of any underlying parameters in phylogenetic analysis, including topology, branch lengths, and substitution models. The non-parametric bootstrap is probably the most common means of assessing confidence in details of the topology. In this procedure, new alignments are created by randomly sampling sites in the alignment with replacement. The same procedure is then repeatedly applied to the new alignments (often 100, but sometimes 1000 if computational time is free and there is nothing better to do) to produce a set of "bootstrapped" topologies. Each data partition (branches that split the taxa into the same subgroups) is then tallied up to obtain the fraction of times that each branch is observed. While ideally these fractions would translate directly into confidence estimates for individual branches, simulations have revealed that there is a downward bias, such that a bootstrap value of 70% or more is often considered significant support for a branch, and values of 90-95% are often considered highly significant. The actual relationship between bootstrap values and confidence limits probably varies with the amount and type of data, and with the structure of the true underlying topology.

Bootstrap values are usually displayed attached to the optimal tree obtained with the original data set. Since there are usually many partitions in the bootstrapped trees that are not present in the original tree, not all partitions are displayed. Those that are not in the original tree are unlikely to have high bootstrap values. Occasionally, if a particular node is not supported by a certain number of bootstraps, that node will be collapsed, resulting in a multifurcations in what is called a "consensus tree". Although this procedure is conservative ("I'm not going to show a partition unless I'm sure about it"), it throws out a great deal of information about the structure of the tree and is probably overly pessimistic. Penny and colleagues have experimented with alternatives, in particular nearest neighbor bootstraps, which give a bootstrap percentage for a particular partition plus the two other partitions that can be reached by nearest neighbor interchange.

In any model-based analysis, a more powerful means of obtaining confidence limits is through parametric bootstrapping. Standard non-parametric bootstrapping is very general and easy to implement, and works perfectly well in the limit of infinite data. Athough it does not specify model parameters, it requires estimation of the frequency of each pattern, which means there may be insufficient data to obtain good estimates of these frequencies. Non-parametric bootstrapping takes the patterns in the data and resamples from them, without making a statement (i.e, without building a model with parameters) about how those patterns were generated. Parametric bootstrapping, in contrast, is based on a model, in particular the model and set of parameters that has the greatest probability of producing the data. Once the parameters are determined, data is re-simulated based on those parameters, and relevant statistics are re-calculated. Through repeated simulations, one can obtain distributions and confidence limits for both the statistics and the underlying parameters.

In phylogenetic analyses, complete parametric bootstrapping is not generally feasible for moderately large datasets, simply because the original maximum likelihood procedures

are themselves computationally burdensome. For this reason, it is more common to run parametric bootstraps without re-estimating the topology every time. It is also possible to test specific topological questions, such as the probability that a particular group is monophyletic.

Posterior Probability

Recently, posterior probability, or Bayesian, analysis has been making a large impact on phylogenetic analysis. The core of the posterior probability approach is to consider the entire distribution of probability space, rather than simply the maximum, as with ML. Bayesian proponents tend to claim that the center of Bayesian analysis is the "priors", the probability of the model parameters before looking at the data. Likelihood proponents argue that prior support is perfectly acceptable so long as it is obtained from general data and is not "made up". Both arguments have philosophical bases and are hotly contested, and it is not worth going into great detail here. Rather, we will focus on the practical aspects of posterior probability analysis in relation to phylogenetics.

The central calculation in posterior probability analysis is the same as with ML, which is calculating the probability of the observed sequence data, *D*, (with the alignment usually treated as though it were observed) given a specific set of model parameters, θ. The posterior probability of any specific set of parameter values can be determined by calculating

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{\sum_{\theta=-\infty}^{\infty} P(D \mid \theta) P(\theta)}. \qquad [8]$$

In the continuous case, the summation changes to an integral, and the probability of any range of parameter values can be determined by summing probability over that range. Ignoring the issue of the prior, the major computational difference is then that in ML one must perform some sort of hill-climbing algorithm to identify the maximum, whereas in posterior analysis one must have a means of efficiently moving through the entire posterior probability space (or at least the portion of that space with high probability density).

Posterior probability space is generally explored using the Metropolis-Hastings algorithm or derivatives thereof, although some use has been made of importance sampling algorithms, and Gibbs sampling has been used for alignments. The basics of Metropolis-Hastings exploration is a Markov chain consisting of a proposal mechanism for moving from point x to point y, and an acceptance mechanism, by which it is decided to move to point y or stay at point x. Every point must be reachable by the proposal distribution, q, and the proposals must be symmetric, such that q(y|x)=q(x|y). The acceptance mechanism is to move to y if it has greater posterior probability, or if not, to move to y with probability P(y)/P(x). With importance sampling, a means of generating independent points is used that is hopefully as close to the posterior probability distrubution as possible. Each point generated is then weighted by the ratio of the posterior probability to the proposal probability. With both methods, the set of points sampled is a sample from the posterior probability distribution, and therefore the probability of any parameter, or

range of parameters, can be found by simply counting up the proportion of points that fall within the appropriate range.

Credible intervals, the Bayesian equivalent of confidence limits, are extremely intuitive as the regions of parameter space containing, for example, 95% of the probability space surrounding the mean parameter values. The points from an importance sampler are independent, but for a Markov chain they are autocorrelated, meaning that the variance of estimates cannot be calculated from the number of points sampled. For this reason, chains are often subsampled every 100 or 1000 points or so, but still, an accurate estimate of the variance can only be obtained by running multiple chains. Another important factor is that the chains must be at equilibrium before sampling begins. It is often difficult and somewhat of an art to determine when equilibrium is reached, the prior points representing "burn-in" that must be thrown out.

Philosophical considerations aside, there are three pragmatic reasons that posterior probability analysis is becoming more popular in phylogenetic research. First, ML requires optimization of each parameter, meaning that computational time is multiplicative with the complexity of (the number of parameters in) the model. With Metropolis and related approaches to posterior analysis, multiple parameters may be varied simultaneously, meaning that computational costs are not necessarily multiplicative. Since phylogenetic models are getting quite complex, and topology space alone has far too many parameters to deal with exhaustively, this is an important consideration. Second, since all parameters must be optimized simultaneously in ML, it is easy to have over-parameterization of the model and lose power for analyzing a few parameters that are of greatest interest. In posterior analysis, it is possible to integrate out parameters that are not of interest, thereby increasing the power for a given data set (although the effective reduction in parameters is uncertain). Third, the credible intervals are not only intuitive, but they can be calculated directly from the sample of the posterior probability space. This means that there is no need to do parametric or non-parametric bootstrapping, which saves a huge amount of time.

Adaptation, Coevolution, Functional Divergence, and Ancestral Reconstruction

Some of the most interesting uses of phylogenetic analysis have to do with understanding deviations from the simple models. With neutral evolution, one expects that evolution will proceed in a straightforward manner, but with selection and adaptation, there may be bursts of change, coevolutionary interactions, and dramatic changes in the evolutionary process. In addition, one may want to make inferences about the ancestral condition in order to infer the direction of evolution and correlate evolutionary changes to environmental change. The pursuit of such questions has a long history, including methods that ignore phylogenetic relationships and methods that do not have an explicit model. Recently, many of these questions have been addressed in a phylogenetic likelihood context, and the accuracy and power of is often dramatically improved.

Adaptation has been successfully detected by comparing rates of substitution between ostensibly neutral sites (usually synonymous changes at 3rd codon positions) and amino

acid substitutions. This was first done for entire proteins or regions of proteins, which detects what might be called “diversifying selection”, in which selective forces are constantly in favor of more change at functionally important sites. This sort of selection is observed in molecular “cat and mouse” situations where interacting molecules are chasing each other in the evolutionary process. Examples are pathogen proteins that interact with host defense systems, and molecules involved in interactions between the sexes. Brief bouts of selection have also been detected on particular branches or lineages, which can be considered “adaptive bursts”, in response to some (often unknown) external or internal environmental change.

Coevolution, sometimes called covariation or correlated mutation (correlated substitution) is the correlation of evolutionary change between sites, whether within or between proteins. This can be generally defined as a situation where changes at one site change the probability of substitution at another site. Coevolutionary analysis has been shown to be particularly susceptible to detecting spurious correlations that arise from phylogenetic relationships rather than coevolutionary relationships when phylogeny is not accounted for. Another problem with coevolution is that in evaluating evidence for interactions between sites, information is limited to those sites under consideration, meaning that large phylogenies are needed in order to have enough data to rise above random fluctuations. For the same reason, it can also be difficult to justify complex models, which may be vastly overparameterized compared to the amount of data at a pair of sites. Approaches to coevolutionary analysis have included use of physicochemical properties to orient amino acid residues, frequency-based analyses such as use of information measures, and analysis of correlation in timing of substitutions by reconstructing where changes occurred on the phylogenetic tree. The relevant likelihood approach (used in Pollock’s LnLCorr program) is to compare the likelihood of a model of independent evolution between sites with the likelihood of a model with coevolutionary interactions between sites. Because the amount of information at any one site is low, it is necessary to perform a parametric bootstrap rather than relying on χ^2 approximations in order to get confidence intervals. Pairwise coevolutionary interactions tend to be weak, but have been detected most clearly between residues that are adjacent in alpha helices. Other interesting general interactions involve maintenance of charge distribution on the surface of proteins, and weak positive correlations in size spread over broad regions of the protein interior.

Finally, it is very plausible that when protein function changes, which may happen after gene duplication, so too should the evolutionary process at some sites in a protein. Gu has built likelihood models for detecting such “functional divergence” by evaluating whether the rate of evolution at individual sites changes dramatically between homologous but functionally divergent protein pairs. Some of the rate divergence detected with this method may be reasonably explained by neutral change, or by divergence of structural features unrelated to functional shifts. Nevertheless, this is an extremely promising approach to understanding functional change and innovation in proteins.

Available Software

Phylip: An old standby from the founding father of phylogenetic analysis using likelihood. Set up as a package of stand-alone programs, Phylip includes distance calculation and basic reconstruction methods using distances, a large variety of parsimony methods, and of course, maximum likelihood. Also includes programs for bootstrapping (another Felsenstein thing) and building consensus trees, and for displaying trees. A new version, 3.6, has just come out. For a long time Phylip was only available as a UNIX program, and it retains that UNIX command-line feel. Relies on Adachi's MOLPHY program for likelihood-based analysis of proteins.

*PAUP**: Perhaps the most user-friendly and popular of the academic programs, although the current version has been in beta testing for almost a decade now and the manual is incomplete. PAUP* is sold for a nominal fee. Swofford first wrote PAUP as a program devoted entirely to parsimony when parsimony was king, and he has been moving steadily to include more and more statistical methods, i.e., distances and maximum likelihood. PAUP* is flexible, fast, and essential for phylogenetic analysis, but probably would have done better to abandon the parsimony methods to a separate program, since it can be difficult to track how parameters and flags are set, and which of them apply to which analyses. Still, PAUP* is a highly usable program and very good for beginners to learn the possibilities of phylogenetic inference. Also, despite its user-friendly interface, PAUP* is available on UNIX and can be automated for batch processing, making it highly versatile for advanced analyses. New developments are often contained in hidden commands that are only slowly released to the general user, so it is worth asking if a new type of analysis is available. PAUP* has almost nothing designed for proteins.

PAML: Ziheng Yang has been the most productive researcher in recent years coming up with new, more complicated, and biologically realistic models of DNA and protein evolution. In particular, this includes gamma models of rate evolution, codon-based models, ancestral reconstruction, and detection of adaptation. Every new analysis is quickly incorporated into PAML, so it is usually on the cutting edge, and has techniques that are not available in other programs, particularly for analysis of protein evolution. PAML's strength is in its models, and it is not built for speedy phylogenetic reconstruction to the degree that PAUP* is, and as a UNIX-based program, its user interface can be difficult for those who are not computer-savvy.

MrBayes and *BAMBE*: These two programs mark the onset of posterior probability, or Bayesian, analysis into phylogenetics. BAMBE came first, but MrBayes is perhaps easier to use. Although driven by a Unix-style command-line interface, MrBayes is based on the same programming model as PAUP*, so many of the commands are directly transferable and will be an easy jump for users of PAUP*. Includes all the usual Bayesian-style analyses, and a wide variety of models for both DNA and protein evolution, and has included more and more options at a rapid pace. Watch for updates; MrBayes will rise in importance in the near future.

Other software: Many of the analyses mentioned here are available in other packages, and many of the analyses are only available in specialty programs, or research programs

that are extremely difficult to use. A list of many programs is available from Felsenstein's group at zoology.uwashington.edu. GCG is a commercial program that contains most of the simpler analyses, and MEGA contains basic analytical distances, although not maximum likelihood or least squares distances, which limits its utility. HiPhy is an interesting likelihood package for programming. The Oxford Zoology group has produced a variety of useful programs for Macintosh computers, and TreeView is a useful viewing program, although it tends to be sensitive and buggy (our departmental computer manager blames it for destroying the file system on his NT machines). BIONJ, WEIGHBOR, LnLCorr, and k-class analysis are available as specialty programs from their authors. There are also numerous packages focused on population genetics analyses.

References

Li, W. H and Graur, D. Fundamentals of Molecular Evolution. (Sinauer, Sunderland) 1991. This book contains an overview of the manner in which molecules evolve.

Volkenstein, M. V. Physical approaches to biological evolution. (Springer-Verlag, Berlin) 1994. This book describes some of the more theoretical aspects of evolution, especially Eigen's pseudospecies theory.

Page, R. D. M and Holmes, E. C. Molecular Evolution: A Phylogenetic Approach. (Blackwell Science, Oxford) 1998. This book also gives an overview of the forms of molecular evolution, what genes are and how they change, how phylogenies are reconstructed, as well as how these approaches can be used to answer important biological questions.

Hillis, D. M, Moritz, C, and Mable, B. Molecular Systematics, 2nd edition. (Sinauer Associates) 1996. Chapter 11, "Phylogenetic Inference" is particularly helpful, although the excess details on parsimony can easily be skipped over, and Chapter 12, "Applications of Molecular Systematics" also has useful information. Many of the other chapters are focused on research techniques.

Durbin, R., Eddy, S., Kroch, A., and Mitchison, G., Biological Sequence Analysis. (Cambridge University Press) 1998. This book has a good description of some of the basics of probability theory and hidden Markov models. Chapter 7 and 8 are on phylogenetic trees, but is somewhat confusingly written; Hillis is an easier read. Most of the chapters go into more detail than the average reader will be interested in, but similarly to a Scientific American article, the chapters are generally more accessible at the beginning. Readers should understand the basics of alignment that are presented in this book, although perhaps not all the details.