

# INTRODUCTION TO MOLECULAR POPULATION GENETICS

## Introduction

The study of evolutionary biology is commonly divided into two components: study of the *processes* by which evolutionary change occurs and study of the *patterns* produced by those processes. By “pattern” we mean primarily the pattern of phylogenetic relationships among species or genes.<sup>1</sup> Studies of evolutionary processes often don’t often devote too much attention to evolutionary patterns, except insofar as it is often necessary to take account of evolutionary history in determining whether or not a particular feature is an adaptation. Similarly, studies of evolutionary pattern often try not to use any knowledge of evolutionary processes to improve their guesses about phylogenetic relationships, because the relationship between process and pattern is often tenuous. Invoking a relationship seems often to be a way of making sure that you get the pattern you want to get from the data.

Or at least that’s the way it was in evolutionary biology when evolutionary biologists were concerned primarily with the evolution of morphological, behavioral, and physiological traits and when systematists used primarily anatomical, morphological, and chemical features (but not proteins or DNA) to describe evolutionary patterns. With the advent of molecular biology after the Second World War and its application to an increasing diversity of organisms in the late 1950s and early 1960s, that began to change. Goodman [1] used the degree of immunological cross-reactivity between serum proteins as an indication of the evolutionary distance among primates. Zuckerkandl and Pauling [5] proposed that after species diverged, their proteins diverged according to a “molecular clock,” providing a way that molecular similarities can be used to reconstruct evolutionary history. In 1966, Harris [2] and Lewontin and Hubby [3, 4] showed that human populations and populations of *Drosophila pseudoobscura* respectively, contained surprising amounts of genetic diversity.

In this course, we’ll focus on advances made in understanding the processes of molecular evolution and pay relatively little attention to the ways in which inferences about evolu-

---

<sup>1</sup>In certain cases it may make sense to talk about a phylogeny of populations within species, but in many cases it doesn’t. We’ll discuss this further when we get to phylogeography in a couple of weeks.

tionary patterns can be made from molecular data. Up to this point in the course we've completely ignored evolutionary pattern. As you'll see in what follows, however, any discussion of molecular evolution, even if it focuses on understanding the processes, cannot avoid some careful attention to the pattern.

## Types of data

Before we delve any further into our study of molecular evolution, it's probably useful to back up a bit and talk a bit about the types of data that are available to molecular evolutionists. We've already encountered a several of these (AFLPs, microsatellites, and SNPs), but there are a variety of important categories into which we can group data used for molecular evolutionary analyses, and it's useful to remind everyone what those groups are and to agree on some terminology for the ones we'll say something about. Let's talk first about the physical basis of the underlying data. Then we'll talk about the laboratory methods used to reveal variation.

### The physical basis of molecular variation

With the exception of RNA viruses, the hereditary information in all organisms is carried in DNA. Ultimately, differences in any of the molecular markers we study (and of genetically-based morphological, behavioral, or physiological traits) is associated with some difference in the physical structure of DNA, and molecular evolutionists study a variety of its aspects.

**Nucleotide sequence** A difference in nucleotide sequence is the most obvious way in which two homologous stretches of DNA may differ. The differences may be in translated portions of protein genes (exons), portions of protein genes that are transcribed but not translated (introns, 5' or 3' untranslated regions), non-transcribed functional regions (promoters), or regions without apparent function.

**Sequence organization** Particular genes may differ between organisms because of differences in the position and number of introns. At the whole genome level, there may be differences in the amount and kind of repetitive sequences, in the amount and type of sequences derived from transposable elements, in the relative proportion of G-C relative to A-T, or even in the identity and arrangement of genes that are present.

**Imprinting** At certain loci in some organisms the expression pattern of a particular allele depends on whether that allele was inherited from the individual's father or its mother.

**Expression** Functional differences among individuals may arise because of differences in the patterns of gene expression, even if there are no differences in the primary sequences of the genes that are expressed.

**Protein sequence** Because of redundancy in the genetic code, a difference in nucleotide sequence at a protein-coding locus may or may not result in proteins with a different amino acid sequence.

**Secondary, tertiary, and quaternary structure** Differences in amino acid sequence may or may not lead to a different distribution of  $\alpha$ -helices and  $\beta$ -sheets, to a different three-dimensional structure, or to different multisubunit combinations.

It is worth remembering that in most eukaryotes there are two different genomes whose characteristics may be analyzed: the nuclear genome and the mitochondrial genome. In plants there is a third: the chloroplast genome. The mitochondrial and chloroplast genomes are typically inherited only through the maternal line, although some instances of biparental inheritance are known.

## Revealing molecular variation

The diversity of laboratory techniques used to reveal molecular variation is even greater than the diversity of underlying physical structures. I'll mention only the most important techniques.

**Immunological distance** Some molecules, notably protein molecules, induce an immune response in common laboratory mammals. The cross-reactivity between an antigen raised to humans and chimps, for example, can be used as a measure of evolutionary distance. The ID between humans and chimps is smaller than it is between humans and orangutans, suggesting that humans and chimps share a more recent common ancestor.

**DNA-DNA hybridization** Once the repetitive sequences have been “subtracted out”, the rate and temperature at which DNA species from two different species anneal reflects the average percent sequence divergence between them. The percent sequence divergence can be used as a measure of evolutionary distance. Immunological distances and DNA-DNA hybridization were used primarily to identify phylogenetic relationships among species. Neither is now widely used in molecular evolution studies.

**Isozymes** Biochemists recognized in the late 1950s that many soluble enzymes occurred in multiple forms within a single individual. Population genetics, notably Hubby and

Lewontin, later recognized that in many cases, these different forms corresponded to different alleles at a single locus, *allozymes*. Allozymes are relatively easy to score in most macroscopic organisms, they are typically co-dominant (the allelic composition of heterozygotes can be inferred), and they allow investigators to identify both variable and non-variable loci.<sup>2</sup> Patterns of variation at allozyme loci may not be representative of genetic variation that does not result from differences in protein structure or that are related to variation in proteins that are insoluble.

**RFLPs** In the 1970s molecular geneticists discovered restriction enzymes, enzymes that cleave DNA at specific 4, 5, or 6 base pair sequences, the *recognition site*. A single nucleotide change in a recognition site is usually enough to eliminate it. Thus, presence or absence of a restriction site at a particular position in a genome provides compelling evidence of an underlying difference in nucleotide sequence at that position.

**RAPDs, AFLPs, ISSRs** With the advent of the polymerase chain reaction in the late 1980s, several related techniques for the rapid assessment of genetic variation in organisms for which little or no prior genetic information was available. These methods differ in details of how the laboratory procedures are performed, but they are similar in that they (a) use PCR to amplify anonymous stretches of DNA, (b) generally produce larger amounts of variation than allozyme analyses of the same taxa, and (c) are bi-allelic, dominant markers. They have the advantage, relative to allozymes, that they sample more or less randomly through the genome. They have the disadvantage that heterozygotes cannot be distinguished from dominant homozygotes, meaning that it is difficult to use them to obtain information about levels of within population inbreeding.

**Microsatellites** Satellite DNA, highly repetitive DNA associated with heterochromatin, had been known since biochemists first began to characterize the large-scale structure of genomes in DNA hybridization studies. In the mid-late 1980s several investigators identified smaller repetitive units dispersed throughout many genomes. Microsatellites, which consist of short (2-6) nucleotide sequences repeated many times, have proven particularly useful for analyses of variation within populations since the mid-1990s. Because of high mutation rates at each locus, they commonly have many alleles. Moreover, they are typically co-dominant, making them more generally useful than dominant markers. Identifying variable microsatellite loci is more laborious than identifying AFLPs, RAPDs, or ISSRs.

---

<sup>2</sup>Classical Mendelian genetics, and quantitative genetics too for that matter, depend on genetic variation in traits to identify the presence of a gene.

**Nucleotide sequence** The advent of automated sequencing has greatly increased the amount of population-level data available on nucleotide sequences. Nucleotide sequence data has an important advantage over most of the types of data discussed so far: allozymes, RFLPs, AFLPs, RAPDs, and ISSRs may all hide variation. Nucleotide sequence differences need not be reflected in any of those markers. On the other hand, each of those markers provides information on variation at several or many, independently inherited loci. Nucleotide sequence information reveals differences at a location that rarely extends more than 2-3kb.

**Single nucleotide polymorphisms** In organisms that are genetically well-characterized it may be possible to identify single nucleotide positions that harbor polymorphisms. These SNPs potentially provide high-resolution insight into patterns of variation within the genome. For example, the HapMap project has identified approximately 3.2M SNPs in the human genome, or about one every kb.

As you can see from these brief descriptions, each of the markers reveals different aspects of underlying hereditary differences among individuals, populations, or species. There is no single “best” marker for evolutionary analyses. Which is best depends on the question you are asking. In many cases in molecular evolution, the interest is intrinsically in the evolution of the molecule itself, so the choice is based not on what those molecules reveal about the organism that contains them but on what questions about which molecules are the most interesting.

## Divergence of nucleotide sequences

Underlying everything else we’re going to discuss in this last part of the course is the idea that we should be able to describe the degree of difference between nucleotide sequences, proteins, or anything else as a result of some underlying evolutionary processes. To illustrate the principle, let’s start with nucleotide sequences and develop a fairly simple model that describes how they become different over time.

Let  $q_t$  be the probability that two homologous nucleotides are identical after having been evolving for  $t$  generations independently since the gene in which they were found was replicated in their common ancestor. Let  $\lambda$  be the probability of a substitution occurring at this nucleotide position in either of the two genes during a small time interval,  $\Delta t$ . Then

$$\begin{aligned}q_{t+\Delta t} &= (1 - \lambda\Delta t)^2 q_t + 2(1 - \lambda\Delta t) \left(\frac{1}{3}\lambda\Delta t\right) (1 - q_t) + o(\Delta t^2) \\ &= (1 - 2\lambda\Delta t)q_t + \left(\frac{2}{3}\lambda\Delta t\right) (1 - q_t) + o(\Delta t^2)\end{aligned}$$

$$\begin{aligned}
q_{t+\Delta t} - q_t &= \frac{2}{3}\lambda\Delta t - \frac{8}{3}\lambda\Delta tq_t + o(\Delta t^2) \\
\frac{q_{t+\Delta t} - q_t}{\Delta t} &= \frac{2}{3}\lambda - \frac{8}{3}\lambda q_t + o(\Delta t) \\
\lim_{\Delta t \rightarrow 0} \frac{q_{t+\Delta t} - q_t}{\Delta t} = \frac{dq_t}{dt} &= \frac{2}{3}\lambda - \frac{8}{3}\lambda q_t \\
q_t &= 1 - \frac{3}{4}\left(1 - e^{-8\lambda t/3}\right)
\end{aligned}$$

The expected number of nucleotide substitutions separating the two sequences at any one position since they diverged is  $d = 2\lambda t$ .<sup>3</sup> Thus,

$$\begin{aligned}
q_t &= 1 - \frac{3}{4}\left(1 - e^{-4d/3}\right) \\
d &= -\frac{3}{4}\ln\left[1 - \frac{4}{3}(1 - q_t)\right]
\end{aligned}$$

This is the simplest model of nucleotide substitution possible—the Jukes-Cantor model. It assumes

- that mutations are equally likely at all positions and
- that mutation among all nucleotides is equally likely.

Let's examine the second of those assumptions first. Observed differences between nucleotide sequences shows that some types of substitutions, i.e., transitions ( $A \iff G$ ,  $T \iff C$ ), occur much more frequently than others, i.e., transversions ( $A \iff G$ ,  $A \iff C$ ,  $T \iff A$ ,  $T \iff G$ ). There are a variety of different substitution models corresponding to different assumed patterns of mutation: Kimura 2 parameter (K2P), Felsenstein 1984 (F84), Hasegawa-Kishino-Yano 1985 (HKY85), Tamura and Nei (TrN), and generalized time-reversible (GTR). The GTR is, as its name suggests, the most general *time-reversible* model. It allows substitution rates to differ between each pair of nucleotides. That's why it's general. It requires, however, that the substitution rate be the same in both

---

<sup>3</sup>The factor 2 is there because  $\lambda t$  substitutions are expected on each branch. In fact you will usually see the equation for  $q_t$  written as  $q_t = 1 - (3/4)(1 - e^{-4\alpha t/3})$ , where  $\alpha = 2\lambda$ .  $\alpha$  is also referred to as the substitution rate, but it refers to the rate of substitution between the two sequences, not to the rate of substitution between each sequence and their common ancestor. If mutations are neutral  $\lambda$  equals the mutation rate, while  $\alpha$  equals twice the mutation rate.

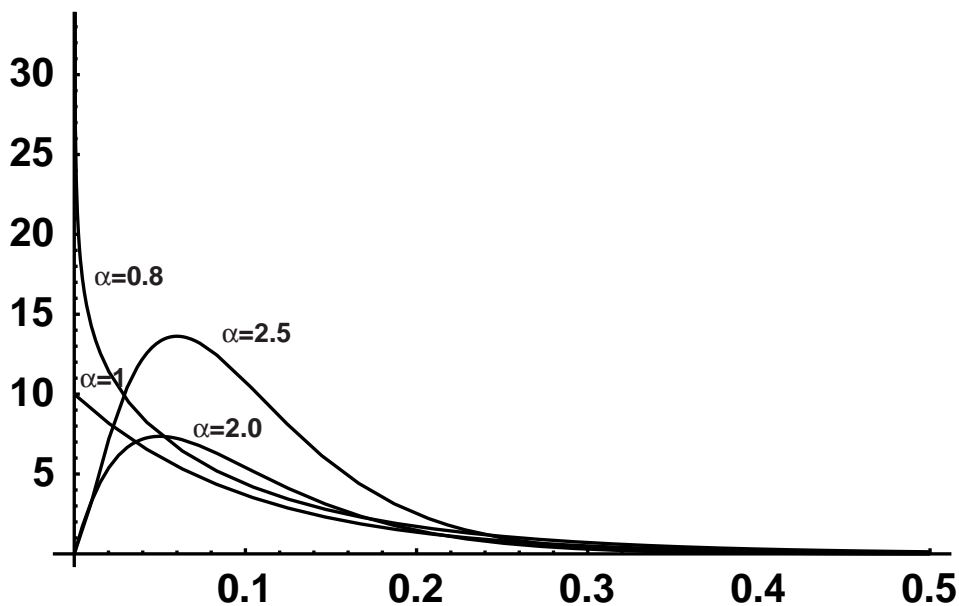


Figure 1: Examples of a gamma distribution.

directions. That's why it's time reversible. While it would be possible to construct a model in which the substitution rate differs depending on the direction of substitution, it leads to something of a paradox: with non-reversible substitution models the distance between two sequences  $A$  and  $B$  depends on whether we measure the distance from  $A$  to  $B$  or from  $B$  to  $A$ .

There are two ways in which the rate of nucleotide substitution can be allowed to vary from position to position—the phenomenon of among-site rate variation. First, we expect the rate of substitution to depend on codon position in protein-coding genes. The sequence can be divided into first, second, and third codon positions and rates calculated separately for each of those positions. Second, we can assume *a priori* that there is a distribution of different rates possible and that this distribution is described by one of the standard distributions from probability theory. We then imagine that the substitution rate at any given site is determined by a random draw from the given probability distribution. The gamma distribution is widely used to describe the pattern of among-site rate variation, because it can approximate a wide variety of different distributions (Figure 1).<sup>4</sup>

The mean substitution rate in each curve above is 0.1. The curves differ only in the

---

<sup>4</sup>And, to be honest, because it is mathematically convenient to work with.

value of a parameter,  $\alpha$ , called the “shape parameter.” The shape parameter gives a nice numerical description of how much rate variation there is, except that it’s backwards. The larger the parameter, the less among-site rate variation there is.

## References

- [1] M. Goodman. Immunocytochemistry of the primates and primate evolution. *Annals of the New York Academy of Sciences*, 102:219–234, 1962.
- [2] H. Harris. Enzyme polymorphisms in man. *Proceedings of the Royal Society of London, Series B*, 164:298–310, 1966.
- [3] J. L. Hubby and R. C. Lewontin. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54:577–594, 1966.
- [4] R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, 54:595–609, 1966.
- [5] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York, NY, 1965.

## Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.