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.[1] 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 sometimes try not to use any knowledge of evolutionary processes to improve their guesses about phylogenetic relationships, because the relationship between process and pattern can be tenuous.[2] Those who take this approach argue that invoking a particular evolutionary process 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 [3] used the degree of immunological cross-reactivity between serum proteins as an indication of the evolutionary distance among primates. Zuckerkandl and Pauling [21] proposed that

[1] 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.

[2] This approach is much less common than it used to be. In the "old days" (meaning when I was a young assistant professor), we had vigorous debates about whether or not it was reasonable to incorporate some knowledge of evolutionary processes into the methods we use for inferring evolutionary patterns. Now it's pretty much taken for granted that we should. One way of justifying a strict parsimony approach to cladistics, however, is by arguing (a) that by minimizing character state changes on a tree you're merely trying to find a pattern of character changes as consistent as possible with the data you've gathered and (b) that evolutionary processes should be invoked only to explain that pattern, not to construct it.

© 2001-2023 Kent E. Holsinger

after species diverged, their proteins diverged according to a "molecular clock," suggesting that molecular similarities could be used to reconstruct evolutionary history. In 1966, Harris [5] and Lewontin and Hubby [6, 11] showed that human populations and populations of *Drosophila pseudoobscura* respectively, contained surprising amounts of genetic diversity.

We'll focus first on advances made in understanding the processes of molecular evolution. Once we have a passing understanding of those processes, we'll shift to topics that are generally more interesting to evolutionary biologists, i.e., making inferences about evolutionary patterns from molecular data. Up to this point in the course we've completely ignored evolutionary pattern.[3] 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

If you're interested in the history of molecular evolution, you may be interested in this review of the types of data that population geneticists have used in the last 50 or 60 years to provide insights into evolutionary processes. If you're not interested in the history, feel free to skip this section. I will touch on only a couple of the real high points during lecture. Much of the data being collected now for population genetics is treated as single nucleotide polymorphisms (sometimes with genetic linkage taken into account) or copy-number variation, and it is derived either from low-coverage whole-genome resequencing or from a reduced representation sequencing approach like some version of RADseq or genotyping-by-sequencing. The exceptions are that for some purposes, microsatellites are still the marker of choice and for others, RNAseq can be used to explore differences in gene expression between individuals or under different conditions.

We've already encountered a couple of these (microsatellites and SNPs), but there are a variety of important categories into which we can group data used for molecular evolutionary analyses. Even though studies of molecular evolution in the last 20-25 years have focused mostly on data derived from DNA sequence or copy number variation, modern applications of molecular data evolved from earlier applications. Markers that were used before the advent of (relatively) easy and cheap DNA sequencing had their limitations, but analyses of those data also laid the groundwork for most or all of what's going on in analyses of molecular evolution today. Thus, it's useful to remind everyone what kinds of molecular data have been used to provide insight into evolutionary patterns and processes and to agree on some

[3] Of course, if you really care about making inferences about evollutionary patterns from molecular data, especially patterns above the species level, you'll want to take the course Paul Lewis teaches. He spends pretty much the entire course discussing these problems.

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.

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 (e.g., introns, 5' or 3' untranslated regions), non-transcribed functional regions (e.g., promoters), or regions without apparent function.

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. **Important note**: Don't forget that some loci code for RNA that has an immediate function without being translated to a protein, e.g., ribosomal RNA and various small nuclear RNAs.

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

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.[4]

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

[4]Of course, differences in expression must ultimately be the result either of a DNA sequence difference somewhere, e.g., in a promoter sequence or the locus encoding a promotor or repressor protein, if it is a genetic difference or of an epigenetic modification of the sequence, e.g., by methylation.

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. In microbial species, only a subset of genes are present in all strains. For example, in *Streptococcus pneumoniae* the "core genome" contains only 73% of the loci present in one fully sequenced reference strain [16]. Similarly, a survey of 20 strains of *Escherichia coli* and one of *E. fergusonii*, *E. coli*'s closest relative, identified only 2000 homologous loci that were present in all strains out of 18,000 orthologous loci identified [19]

Copy number variation Even within diploid genomes, there may be substantial differences in the number of copies of particular genes. In humans, for example, 76 copy-number polymorphisms (CNPs) were identified in a sample of only 20 individuals, and individuals differed from one another by an average of 11 CNPs. [18].

It is worth remembering that in nearly all 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. In some protists, there may be even more, because of secondary or tertiary endosymbiosis. The mitochondrial and chloroplast genomes are typically inherited only through the maternal line, although some instances of biparental inheritance are known.[5] In conifers, chloroplasts are paternally inherited, i.e., through the pollen parent, and mitochondria are maternally inherited, i.e., through the seed parent [15]

Revealing molecular variation

The diversity of laboratory techniques used to reveal molecular variation is even greater than the diversity of underlying physical structures. Various techniques involving direct measurement of aspects of DNA sequence variation are by far the most common today, so I'll mention only a few of the techniques that were most widely used in the past.[6]

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

[5] Recent evidence suggests that mitochondria may occasionaly be inherited biparentally in humans [12].

[6] Note: Several of the techniques in this list are primarily of historical interest. They were widely used in the past, but they are no longer used (or no longer used very much).

DNA-DNA hybridization Once repetitive sequences of DNA have been "subtracted out",[7] 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 once widely used 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 geneticists, 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.[8] 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 positon.

RAPDs, AFLPs, ISSRs With the advent of the polymerase chain reaction in the late 1980s, several related techniques were developed 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.[9]

[7] See below for a description of some of these repetitive seqeuences.

[8] Classical Mendelian genetics, and quantitative genetics too for that matter, depends on genetic variation in traits to identify the presence of a gene.

[9] To be fair, it is possible to distinguish heterozygotes from homozyotes with AFLPs, if you are **very** careful with your PCR technique [7]. That being said, few people are careful enough with their PCR to be

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-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.[10] 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, however, more laborious than identifying AFLPs, RAPDs, or ISSRs.

Nucleotide sequence The advent of automated sequencing[11] has greatly increased the amount of population-level data available on nucleotide sequences. The even more recent arrival of high-throughput DNA sequencing means that sequence information is accumulating even more rapidly. Nucleotide sequence data has an important advantage over most of the types of data discussed so far: allozymes, RFLPs, AFLPs, RAPDs, and ISSRs all hide some amount of nucleotide sequence 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. Of course, as next generation sequencing techniques become less expensive and more widely available, we will see more and more examples of nucleotide sequence variation from many loci within individuals.[12]

Single nucleotide polymorphisms In organisms that are genetically well-characterized, it is possible to identify a large number of single nucleotide positions that harbor polymorphisms. SNPs potentially provide high-resolution insight into patterns of variation within the genome. For example, the HapMap project identified approximately 3.2M SNPs in the human genome, or about one every kb [1]. With the advent of RAD-seq,

---

able to score AFLPs reliably as codominant markers, and I am unaware of anyone who has done so outside of a controlled breeding program.

[10]The rapidly diminishing cost of high-throughput nucleotide sequencing, however, suggests that microsatellites will soon join allozymes, RAPDs, AFLPs, ISSRs, and RFLPs as of interest primarily for historical reasons.

[11]In the old days, sequencing DNA meant running samples on a polyacrylamide gel, transferring them to a membrane, hybridizing with ^{32}P, and exposing X-ray film to the membrane for several days before developing it.

[12]For example, Nora Mitchell's paper on the phylogeny of *Protea* [14] was based on analysis of nucleotide sequence variation at nearly 500 loci.

GBS, and similar approaches, it has become possible to identify large numbers of SNPs even in organisms that are not genetically well-characterized [2, 13].

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 much of what we're going to discuss in this 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.[13]

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 λ be the probability of a substitution[14] occuring at this nucleotide position in either of the two genes during a small time interval, Δt. Then

$$\begin{aligned}
q_{t+\Delta t} &= (1-\lambda\Delta t)^2 q_t + 2\,(1-\lambda\Delta t)\left(\frac{1}{3}\lambda\Delta t\right)(1-q_t) + \mathrm{o}(\Delta t^2) \\
&= (1-2\lambda\Delta t) q_t + \left(\frac{2}{3}\lambda\Delta t\right)(1-q_t) + \mathrm{o}(\Delta t^2) \\
q_{t+\Delta t} - q_t &= \frac{2}{3}\lambda\Delta t - \frac{8}{3}\lambda\Delta t q_t + \mathrm{o}(\Delta t^2) \\
\frac{q_{t+\Delta t} - q_t}{\Delta t} &= \frac{2}{3}\lambda - \frac{8}{3}\lambda q_t + \mathrm{o}(\Delta t) \\
\lim_{\Delta t \to 0} \frac{q_{t+\Delta t} - q_t}{\Delta t} = \frac{dq_t}{dt} &= \frac{2}{3}\lambda - \frac{8}{3}\lambda q_t
\end{aligned}$$

TAMO

[13] By now you should realize that when I write that something is "fairly simple", I mean that it's fairly simple to someone who's comfortable with mathematics.

[14] Notice that I wrote "substitution," not "mutation." We'll come back to this distinction later. It turns out to be really important.

$$q_t \quad = \quad 1 - \frac{3}{4}\left(1 - e^{-8\lambda t/3}\right)$$

The expected number of nucleotide substitutions separating the two sequences at any one position since they diverged is $d = 2\lambda t$.[15] 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 [8]. It assumes

- that substitutions are equally likely at all positions and
- that substitution 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 ($A \iff G$ [purine to purine], $C \iff T$ [pyrimidine to pyrimidine]), occur much more frequently than others, i.e., transversions ($A \iff T$, $A \iff C$, $G \iff C$, $G \iff T$ [purine to pyrimidine or vice versa]). There are a variety of different substitution models corresponding to different assumed patterns of substitution: 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 still requires, however, that the substitution rate be the same in both directions. That's what it means to say that it's time reversible. While it is 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. That is to say that the distance from A to B isn't the same as the distance from B to A.

[15] The factor 2 is there because λt substitutions are expected on each branch. In fact, you will usually see the equation for q_t written as $q_t = 1 - (3/4)\left(1 - e^{-4\alpha t/3}\right)$, where $\alpha = 2\lambda$. α 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.

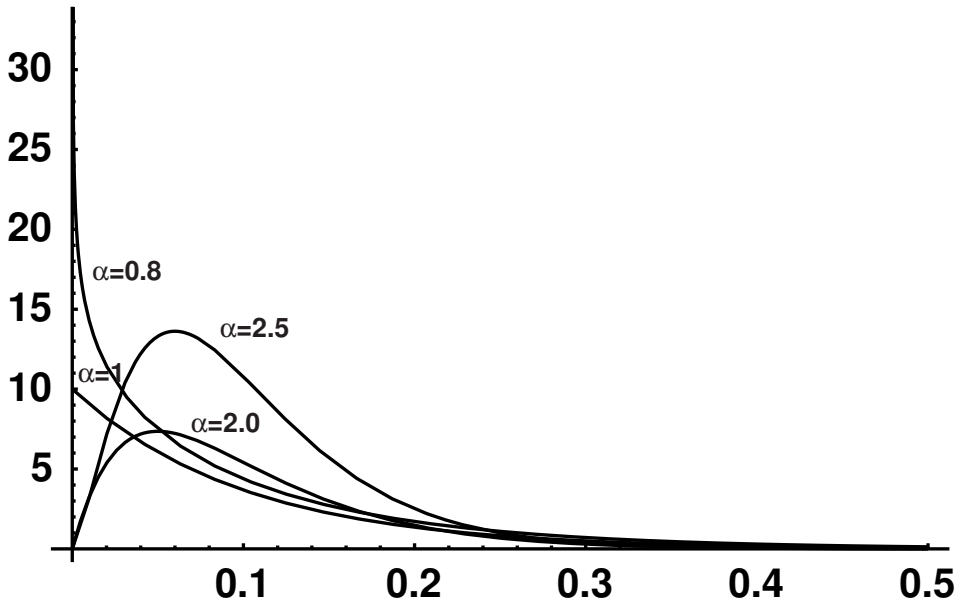

Figure 1: Examples of a gamma distribution.

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 that probability distribution. The gamma distribution is widely to describe the pattern of among-site rate variation, because it can approximate a wide variety of different distributions (Figure 1).[16]

The mean substitution rate in each curve above is 0.1. The curves differ only in the value of a parameter, α, 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.

[16]And, to be honest, because it is mathematically convenient to work with.

The neutral theory of molecular evolution

I didn't make a big deal of it in what we just went over, but in deriving the Jukes-Cantor equation I used the phrase "substitution rate" instead of the phrase "mutation rate."[17] As a preface to what is about to follow, let me explain the difference.

- *Mutation rate* refers to the rate at which changes are incorporated into a nucleotide sequence during the process of replication, i.e., the probability that an allele differs from the copy of that allele in its parent from which it was derived. *Mutation rate* refers to the rate at which mutations arise.
- An allele substitution occurs when a newly arisen allele is incorporated into a population, e.g., when a newly arisen allele becomes fixed in a population. *Substitution rate* refers to the rate at which allele substitutions occur.

Mutation rates and substitution rates are obviously related — substitutions can't happen unless mutations occur, after all —, but it's important to remember that they refer to different processes.

Early empirical observations

By the early 1960s amino acid sequences of hemoglobins and cytochrome c for many mammals had been determined. When the sequences were compared, investigators began to notice that the number of amino acid differences between different pairs of mammals seemed to be roughly proportional to the time since they had diverged from one another, as inferred from the fossil record. Zuckerkandl and Pauling [21] proposed the *molecular clock hypothesis* to explain these results. Specifically, they proposed that there was a constant rate of amino acid substitution over time. Sarich and Wilson [17, 20] used the molecular clock hypothesis to propose that humans and apes diverged approximately 5 million years ago. While that proposal may not seem particularly controversial now, it generated enormous controversy at the time, because at the time many paleoanthropologists interpreted the evidence to indicate humans diverged from apes as much as 30 million years ago.

One year after Zuckerkandl and Pauling's paper, Harris [5] and Hubby and Lewontin [6, 11] showed that protein electrophoresis could be used to reveal surprising amounts of genetic variability within populations. Harris studied 10 loci in human populations, found three of them to be polymorphic, and identified one locus with three alleles. Hubby and Lewontin

[17] In fact, I just mentioned the distinction in passing in two different footnotes.

studied 18 loci in *Drosophila pseudoobscura*, found seven to be polymorphic, and five that had three or more alleles.

Both sets of observations posed real challenges for evolutionary geneticists. It was difficult to imagine an evolutionary mechanism that could produce a constant rate of substitution. It was similarly difficult to imagine that natural selection could maintain so much polymorphism within populations. The "cost of selection," as Haldane [4] called it would simply be too high.

Neutral substitutions and neutral variation

Kimura [9] and King and Jukes [10] proposed a way to solve both empirical problems. If the vast majority of amino acid substitutions are selectively neutral,[18] then substitutions will occur at approximately a constant rate (assuming that substitution rates don't vary over time) and it will be easy to maintain lots of polymorphism within populations because there will be no cost of selection. I'll develop both of those points in a bit more detail in just a moment, but let me first be precise about what the neutral theory of molecular evolution actually proposes. More specifically, let me first be precise about what it does *not* propose. I'll do so specifically in the context of protein evolution for now, although we'll broaden the scope later.

- *The neutral theory asserts that alternative alleles at variable protein loci are selectively neutral.* This does *not* mean that the locus is unimportant, only that the alternative alleles found at this locus are selectively neutral.

  - Glucose-phosphate isomerase is an esssential enzyme. It catalyzes the first step of glycolysis, the conversion of glucose-6-phosphate into fructose-6-phosphate.
  - Natural populations of many, perhaps most, populations of plants and animals are polymorphic at this locus, i.e., they have two or more alleles with different amino acid sequences.
  - The neutral theory asserts that the alternative alleles are essentially equivalent in fitness, in the sense that genetic drift, rather than natural selection, dominates the dynamics of frequency changes among them.

[18] Notice that I just said that we're going to assume that the vast majority of nucleotide *substitutions* are selectively neutral. This doesn't mean that most nucleotide *mutations* are selectively neutral. Indeed, we'll see that most of them are deleterious.

- By *selectively neutral* we do *not* mean that the alternative alleles have no effect on physiology or fitness. We mean that the selection among different genotypes at this locus is sufficiently weak that the pattern of variation is determined primarily by the interaction of mutation, drift, mating system, and migration. This is roughly equivalent to saying that $N_e s < 1$, where N_e is the effective population size and s is the selection coefficient on alleles at this locus.

  - Experiments in *Colias* butterflies, and other organisms have shown that different electrophoretic variants of GPI have different enzymatic capabilities and different thermal stabilities. In some cases, these differences have been related to differences in individual performance.
  - If populations of *Colias* are large and the differences in fitness associated with differences in genotype are large, i.e., if $N_e s > 1$, then selection plays a predominant role in determining patterns of diversity at this locus, i.e., the neutral theory of molecular evolution would not apply.
  - If populations of *Colias* are small or the differences in fitness associated with differences in genotype are small, or both, then drift plays a predominant role in determining patterns of diversity at this locus, i.e., the neutral theory of molecular evolution applies.

In short, the neutral theory of molecular really asserts only that observed amino acid substitutions and polymorphisms are *effectively* neutral, not that the loci involved are unimportant or that allelic differences at those loci have no effect on fitness.

The rate of molecular evolution

We're now going to calculate the rate of molecular evolution, i.e., the rate of allelic substitution, under the hypothesis that mutations are selectively neutral.[19] To get that rate we need two things: the rate at which new mutations occur and the probability with which new mutations are fixed. In a word equation

$$\begin{aligned} \text{\# of substitutions/generation} &= (\text{\# of mutations/generation}) \times (\text{probability of fixation}) \\ \lambda &= \mu_0 p_0 \quad . \end{aligned}$$

Surprisingly,[20] it's pretty easy to calculate both μ_0 and p_0 from first principles.

[19] Notice that contrary to what I said earlier, here I am assuming that *mutations* are neutral, not just substitutions.

[20] Or perhaps not.

In a diploid population of size N, there are $2N$ gametes. The probability that any one of them mutates is just the mutation rate, μ, so

$$\mu_0 = 2N\mu \quad . \tag{1}$$

To calculate the probability of fixation, we have to say something about the dynamics of alleles in populations. Let's suppose that we're dealing with a single population, to keep things simple. Now, you have to remember a little of what you learned about the properties of genetic drift. If the current frequency of an allele is p_0, what's the probability that is eventually fixed? p_0. When a new mutation occurs there's only one copy of it,[21] so the frequency of a newly arisen mutation is $1/2N$ and

$$p_0 = \frac{1}{2N} \quad . \tag{2}$$

Putting (1) and (2) together we find

$$\begin{aligned} \lambda &= \mu_0 p_0 \\ &= (2N\mu)\left(\frac{1}{2N}\right) \\ &= \mu \quad . \end{aligned}$$

In other words, if mutations are selectively neutral, the substitution rate is equal to the mutation rate. Since mutation rates are (mostly) governed by physical factors that remain relatively constant, mutation rates should remain constant, implying that substitution rates should remain constant if substitutions are selectively neutral. In short, if mutations are selectively neutral, we expect a molecular clock.

Diversity in populations

Protein-coding genes consist of hundreds or thousands of nucleotides, each of which could mutate to one of three other nucleotides.[22] That's not an infinite number of possibilities, but it's pretty large.[23] It suggests that we could treat every mutation that occurs as if it

[21] By definition. It's new.

[22] Why three when there are four nucleotides? Because if the nucleotide at a certain position is an A, for example, it can only *change* to a C, G, or T.

[23] If a protein consists of 400 amino acids, that's 1200 nucleotides. There are $4^{1200} \approx 10^{720}$ different sequences that are 1200 nucleotides long. For context, there are only about 3.28×10^{80} elementary particles in the universe (`https://www.popularmechanics.com/space/a27259/how-many-particles-are-in-the-entire-universe/`).

were completely new, a mutation that has never been seen before and will never be seen again. Does that description ring any bells? Does the infinite alleles model sound familiar? It should, because it exactly fits the situation I've just described.

Having remembered that this situation is well described by the infinite alleles model, I'm sure you'll also remember that we can calculate the equilibrium inbreeding coefficient for the infinite alleles model, i.e.,

$$f = \frac{1}{4N_e\mu + 1} \quad .$$

What's important about this for our purposes, is that to the extent that the infinite alleles model is appropriate for molecular data, then f is the frequency of homozygotes we should see in populations and $1 - f$ is the frequency of heterozygotes. So in large populations we should find more diversity than in small ones, which is roughly what we do find. Notice, however, that here we're talking about heterozygosity at individual nucleotide positions,[24] not heterozygosity of halpotypes.

Conclusions

In broad outline then, the neutral theory does a pretty good job of dealing with at least some types of molecular data. I'm sure that some of you are already thinking, "But what about third codon positions *versus* first and second?" or "What about the observation that histone loci evolve much more slowly than interferons or MHC loci?" Those are good questions, and those are where we're going next. As we'll see, molecular evolutionists have elaborated the framework extensively[25] in the last sixty years, but these basic principles underlie every investigation that's conducted. That's why I wanted to spend a fair amount of time going over the logic and consequences. Besides, it's a rare case in population genetics where the fundamental mathematics that lies behind some important predictions are easy to understand.[26]

References

[1] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[24] Since the mutation rate we're talking about applies to individual nucleotide positions.

[25] That mean's they've made it more complicated.

[26] It's the concepts that get tricky, not the algebra, or at least that's what I think.

[2] Robert J Elshire, Jeffrey C Glaubitz, Qi Sun, Jesse A Poland, Ken Kawamoto, Edward Buckler, and Sharon E Mitchell. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, 6(5):e19379, May 2011.

[3] M Goodman. Immunocytochemistry of the primates and primate evolution. *Annals of the New York Academy of Sciences*, 102:219–234, 1962.

[4] J. B. S. Haldane. The cost of natural selection. *Journal of Genetics*, 55:511–524, 1957.

[5] H Harris. Enzyme polymorphisms in man. *Proceedings of the Royal Society of London, Series B*, 164:298–310, 1966.

[6] 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.

[7] R C Jansen, H Geerlings, A J VanOeveren, and R C VanSchaik. A comment on codominant scoring of AFLP markers. *Genetics*, 158(2):925–926, 2001.

[8] T. H. Jukes and C. R. Cantor. *Evolution of Protein Molecules*. Academic Press, New York, 1969.

[9] M Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[10] J L King and T L Jukes. Non-Darwinian evolution. *Science*, 164:788–798, 1969.

[11] 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.

[12] Shiyu Luo, C. Alexander Valencia, Jinglan Zhang, Ni-Chung Lee, Jesse Slone, Baoheng Gui, Xinjian Wang, Zhuo Li, Sarah Dell, Jenice Brown, Stella Maris Chen, Yin-Hsiu Chien, Wuh-Liang Hwu, Pi-Chuan Fan, Lee-Jun Wong, Paldeep S. Atwal, and Taosheng Huang. Biparental inheritance of mitochondrial dna in humans. *Proceedings of the National Academy of Sciences USA*, 115(51):13039–13044, 2018.

[13] Garrett J. McKinney, Wesley A. Larson, Lisa W. Seeb, and James E. Seeb. Radseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on breaking rad by lowry et al. (2016). *Molecular Ecology Resources*, 17(3):356–361, 2017.

[14] Nora Mitchell, Paul O. Lewis, Emily Moriarty Lemmon, Alan R. Lemmon, and Kent E. Holsinger. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of protea l. *American Journal of Botany*, 104(1):102–115, 2017.

[15] David B. Neale and Ronald R. Sederoff. Inheritance and evolution of conifer organelle genomes. In James W. Hanover, Daniel E. Keathley, Claire M. Wilson, and Gregory Kuny, editors, *Genetic Manipulation of Woody Plants*, pages 251–264. Springer US, Boston, MA, 1988.

[16] Caroline Obert, Jack Sublett, Deepak Kaushal, Ernesto Hinojosa, Theresa Barton, Elaine I Tuomanen, and Carlos J Orihuela. Identification of a Candidate Streptococcus pneumoniae Core Genome and Regions of Diversity Correlated with Invasive Pneumococcal Disease. *Infection and Immunity*, 74(8):4766–4777, 2006.

[17] V M Sarich and A C Wilson. Immunological time scale for hominid evolution. *Science*, 158:1200–1203, 1967.

[18] Jonathan Sebat, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Par Lundin, Susanne Maner, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, T Conrad Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-Scale Copy Number Polymorphism in the Human Genome. *Science*, 305(5683):525–528, 2004.

[19] Marie Touchon, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, Odile Bouvet, Alexandra Calteau, Hélène Chiapello, Olivier Clermont, Stéphane Cruveiller, Antoine Danchin, Médéric Diard, Carole Dossat, Meriem El Karoui, Eric Frapy, Louis Garry, Jean Marc Ghigo, Anne Marie Gilles, James Johnson, Chantal Le Bougénec, Mathilde Lescat, Sophie Mangenot, Vanessa Martinez-Jéhanne, Ivan Matic, Xavier Nassif, Sophie Oztas, Marie Agnès Petit, Christophe Pichon, Zoé Rouy, Claude Saint Ruf, Dominique Schneider, Jérôme Tourret, Benoit Vacherie, David Vallenet, Claudine Médigue, Eduardo P C Rocha, and Erick Denamur. Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLoS Genet*, 5(1):e1000344, 2009.

[20] A C Wilson and V M Sarich. A molecular time scale for human evolution. *Proceedings of the National Academy of Sciences U.S.A.*, 63:1088–1093, 1969.

[21] 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 License. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.