6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

6.047 LECTURE 11, MOLECULAR EVOLUTION/COALESCENCE/SELECTION/KAKS OCTOBER 09, 2008

1. Introduction

Evolution is the shaping of phenotypic and genotypic variation by natural selection. The relative importance of selection and mutation has been a long standing question in population genetics. And, we'd like to understand the relative importance of each of these forces. In particular, we'd like to know which genes are undergoing active selection. *neutrality tests* are a key tool for addressing this question. There are many different evolutionary forces that might cause deviations from neutrality such as differential mutation rates, recombination, population structure, drift in addition selection and others. Therefore, it is easier to develop test of neutrality rather than directly searching for signatures of selection. From a historical perspective, most of these ideas were developed theoretically in the last century. And only in the last few decades that we were able to gather data to directly test these theories. 1983 is the first time that we have molecular polymorphism data.

First, we need to understand the neutral model, focusing on inheritance alone. Historically, this area of study dates back to Darwin's time. Scientists in the 1860s believed in *blending inheritance* which stated that each organ was determined by a different *gemmule*. In this model, children inherit a blending of their parents' corresponding *gemmules*, so if mom had a "yellow" gemmule and dad had a "blue" gemmule, then baby would inherit a "greenish" gemmule. As Darwin's critics pointed out, this model predicts that everyone's gemmules blend and blend until all gemmules are a drab shade of gray, losing all genetic diversity. Specifically, Fleeming Jenkins in 1867 showed that the total genetic variation will be halved assuming this mode of inheritance.

However, Mendel had developed a theory particulate inheritance around the same time but was not widely recognized. Only in the beginning of 20th century, researchers appreciated the importance of his results. Mendel's more accurate model, developed as a result of his famous study of pea plants, states that each trait is determined by two corresponding *alleles*, which together determine a person's *phenotype*. Offspring receive one allele from each parent, selected randomly from the parent's two copies. The allele model is quite close to what actually happens when gametes pair during meiosis, and correctly predicts the observed phenomena of dominance and recessivity. These ideas are summarized as the Law of Segregation and the Law of Independent Assortment.

2. Hardy-Weinberg Law

A natural question to ask was how genetic variation changes under the particulate theory of inheritance. The *Hardy-Weinberg Law* (1908) answers this question in an ideal population. This model assumes infinite population size, completely random mating, and an absence of selection, mutation, or migration. Considering two versions A and a of an allele, let u_0, v_0, and w_0 be the respective frequencies of genotypes AA, Aa, and aa, respectively. Then the frequency of A is $p = u_0 + v_0/2$, and the frequency of a is $q = w_0 + v_0/2$. A Punnett square predicts that after a single generation, we observe frequencies $u = p^2$, $v = 2pq$, and $w = q^2$, and that these frequencies remain fixed over successive generations. Because $q = 1 - p$, our entire model is characterized by a single parameter p.

In practice, several of the Hardy-Weinberg Law's assumptions are hardly ever met but it still provides a general framework for thinking about genetic variation. Consider slide 3 page 2 as an example of the application of the HW Law, and let us denote recessive allele causing sickle-cell anemia by s. We observe that many more people have genotype Ss than is expected under the HW equilibrium. This discrepancy is indeed due to selection, because people with one copy of s possess immunity to the dreaded tropical disease malaria.

We next study two different approaches to neutral theory: the *prospective approach* of classic population genetics, and the *retrospective approach* focusing on the *coalescent.*

3. Prospective method

3.1. **Neutral Theory History.**

In the 1960s, people thought that all mutations differ in their fitness, so selection would rule out bad mutations and fix good mutations, with variation kept by balancing selection. So when Neutral Theory was first proposed by Motoo Kimura, a Japanese population geneticist, it was shocking and controversial at the time. Having a background in physics, Kimura incorporated diffusion approximations to the field of population genetics, focusing on finite populations. The main premise of Neutral Theory is that most of the mutations are neutral or nearly neutral, and the change in allelic frequencies is a result of random genetic drift in finite populations (Slide 6, Page 2). All the mutations will eventually become extinct or fixed by this process unless directly opposed by other evolutionary forces. Hence, the genetic variation seen in a population are generally caused by mutations which are on their way to fixation/extinction (Slide 1, Page 3). As a consequence, if one knows the rate of mutations and how quickly they fix in the population, one could have a good idea of how the allelic distribution in the population should be.

3.2. **Ewens Sampling Formula.**

In 1972, Warren Ewens proposed the famous Ewens Sampling Formula, which is based on diffusion theory and introduced the *infinite alleles* model. The infinite alleles model claims that there are an infinite number of states into which an allele can mutate, so each mutation generates a unique allele. Ewens used the concept of *identity by descent (IBD)* as opposed to *identity by kind.* IBD is a concept that is defined with respect to the allele in the ancestor. Imagine that you and your sibling received the same copy of chr7 from your mother. In this particular chromosome even if there is a mutation in one of the genes, your copy and your siblings copy will be IBD but not identitical by kind.

Ewens model depends on parameters N, the population size, and μ, the mutation rate. We assume a diploid population (two alleles/ person), so there are $2N$ alleles in total. In each generation, $2N\mu$ new alleles are introduced to the population, each with an initial frequency of $1/2N$. The Ewens Sampling Formula calculates the probability that a sample of $n(n << N)$ gene copies contains k alleles such that there are $a_1, a_2, \ldots a_k$ alleles represented $1, 2, \ldots, n$ times in the sample:

$$P(a_1, a_2, \ldots, a_n) = \frac{n!\Theta^k}{\Theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!},$$

where

$$\Theta_{(n)} = \Theta(\Theta + 1) \ldots (\Theta + n - 1).$$

The main purpose of the Ewens Sampling Formula is to get an expected site frequencies spectrum (Slide 6, Page 2). This formula allows us to consider questions like how many singletons do we expect given that we sample 100 individuals from an infinite population.

4. Retrospective approach

4.1. **Coalescent Theory.** Coalescent Theory was first introduced by Kingman in the early 1980s. Coalescent theory is a sample-based approach to population genetics and is specifically concerned with making inferences under a rigorous statistical framework. The word coalescent refers to the occurrence, backward in time, of a common ancestor between a pair of lineages that are both ancestral to a present-day sample. The coalescent is a backward-time stochastic process that models the ancestry of a sample of size n, back to the most recent common ancestor (MRCA) of the entire sample. The coalescent makes detailed predictions about patterns of genetic variation in the sample. All polymorphisms in the sample can then be demonstrated by putting mutation events on different branches of the coalescent tree (Slide 1, Page 4). The expected time to MCRA of a population could be found at approximately T=4N generations before, but in practice, the MCRA of a 20-invdividual sample can be very close to that of the total population. The ancestry of a sample of finite size (n) is composed of $n - 1$ independent, exponentially-distributed coalescence times, each ending with a coalescent event between a random pair of lineages. When there are i ancestral lineages the rate of coalescence is $i(i-1)/2$ and the expected time to a coalescent event is $2/(i(i-1))$.

4.2. **tests of selection within species.** We construct three polymorphism summary statistics:

(1) S, the number of segregating sites in the sample;
(2) π, the average number of pairwise differences;
(3) ν_i, the number of sites that divide the sample into i and $n - i$ sequences.

These statistics can be used to estimate the important statistic Θ (Slide 6, Page 4). Under neutrality all of the estimates of Θ should be equal. However, selection, population structure and other evolutionary forces could change the allelic frequency spectrum in population and lead to differences in these estimates. Based on this inference, three selection tests are proposed by Tajima and Fu and Li respectively. (Slide 1, Page 5). In these tests, neutral expectation is used as null hypothesis (Slide 3, Page 5), an alternative hypothesis (positive selection, balancing selection, population structure/subdivision or population expansion) is tested

(Slide 4,5,6, Page 5 and Slide 1, Page 6). For example, positive selection would lead to an excess of singletons while balancing selection would lead to a depletion of singletons, doublets and triplets. However, population subdivision would also yield a similar profile to balancing selection and population expansion will cause a signal similar to positive selection. Therefore, these tests do not provide direct evidence of selection.

Several other tests have been developed that tries to improve on the ideas of the Tajima's D and Fu and Li's statistics. HKA Test, for example, measures the polymorphism and divergence in two loci, and tests whether there is excess in one of the two classes (Slide 2, 3, Page 6). The MK test, on the other hand, measures the synonymous and nonsynonymous polymorphism and divergence in one locus, and tests whether there is an excess in one class (Slide 4, 5, Page 6). Hence, MK test requires information from a single locus but polymorphism data from 2 species/

4.3. **Rate-based selection metric.** To determine if a certain gene is under selection, we may use the rate-based selection metric. Recall that an amino acid is determined by a codon of three adjacent nucleotides, and that different codons, such as CGC and CGA, can represent the same amino acid, arginine in this case. A *synonymous mutation* leaves the amino acid sequence intact, while a *non-synonymous* mutation changes it.

We use various methods, notably PAML, to compute d_N, the rate at which synonymous mutations occur, and d_S, the rate at which synonymous mutations occur. We define the *rate-based selection metric* to be d_N/d_S. There are three cases of interest:

(1) $d_N/d_S < 1$, purifying selection
(2) $d_N/d_S = 1$, neutral expectation
(3) $d_N/d_S > 1$, positive selection

In general, genes performing essential roles, such as those encoding hemoglobin, undergo purifying selection, because changes tend to disrupt the gene's function and cause harm. Genes that perform peripheral functions can undergo positive selection if some change is useful, but not necessary for survival. For instance, changes to a mammals hair coloring could provide useful camouflage.

Rate-based tests could be affected if synonymous sites are not completely neutral. There is growing evidence for the use of preferred codons for certain amino acids instead of the other equivalent codons. This phenomenon is known as codon bias. There is evidence for positive correlation between d_N/d_S and dispensability, gene length, and negative correlation with expression level, protein abundance, codon bias, number of protein-protein interactions and centrality in interaction networks.