

Separating the Effects of Genetic Drift and Natural Selection using a Modification of Tajima's D Statistic

Dianne Kopp, Karen O'Connell, Spencer Tipping,
Pamela J. Ryan, and Anton E. Weisstein

April 25, 2008

Abstract

Several statistical measures have been devised to infer past evolutionary forces from patterns of genetic diversity in sequence data. Tajima's D detects increases or decreases in overall genetic diversity, but cannot assess the relative contribution of genetic drift and natural selection to the variation observed. By applying Tajima's D analysis separately to synonymous (D_{SYN}) and nonsynonymous (D_{NON}) mutations, we can separate the effects of drift from those of selection. Specifically, (D_{SYN}) measures the effects of drift within a population, while ($D_{NON} - D_{SYN}$) measures the effects of selection for other organisms.

Critical values for (D_{SYN}) and ($D_{NON} - D_{SYN}$) will be generated under a model of neutral evolution. We will then analyze data obtained from computer simulations of specific evolutionary and demographic histories to measure our method's ability to correctly infer the population's history. Although our model reflects the details of HIV evolution, we anticipate our extension of Tajima's analysis will have a broader application.

1 Background

The D statistic is calculated from the formula $D = \frac{\Pi - \Theta}{S_d}$, where Π is obtained by dividing the pairwise distance of each pair of DNA sequences within the dataset by the total number of comparisons made. Θ is calculated by dividing the number of nucleotide locations, called sites, which have undergone at least one mutation (known as segregating sites and denoted S) by a correctional factor for sampling size. In this case, the correction factor is a_1 or $\sum_{i=1}^{n-1} \frac{1}{i}$ where n is the population size. [2]

A D value of approximately zero indicates the distribution of genetic variation is consistent with the neutral model, under which no selection or genetic drift occurs. A significantly high or low D indicates an increase or decrease in genetic variation within the population, respectively

$D < 0$	$D > 0$
Purifying selection or Population bottleneck	Diversifying selection or Population subdivision

Figure 1: Currently, the D analysis indicates only an increase or decrease in genetic variation within a population.

(see Figure 1). To identify evolutionary patterns like diversifying selection, purifying selection, population subdivision, and population bottlenecks, we use the new variables D_{SYN} and D_{NON} .

Because synonymous changes do not alter the amino acid that will be inserted in protein formation, they typically do not alter an organism's reproductive fitness within the population. These silent mutations generally have a neutral effect on the individual's reproductive fitness; therefore, they are typically under the influence of random evolutionary forces, not selective forces. A random evolutionary force called genetic drift comes about due to random fluctuations in allele frequencies within a population. Synonymous mutations are exclusively acted upon by this random evolutionary force and the calculation of D_{SYN} using exclusively synonymous mutations will represent only the effects of drift on the population. A D_{SYN} significantly less than zero indicates random evolutionary forces have decreased genetic diversity (i.e. population bottleneck). Likewise, a D_{SYN} that is significantly higher than zero suggests drift has increased diversity (i.e. population subdivision). However, random changes in nucleotides may or may not change the types of amino acids that compose a particular protein. Therefore, synonymous and nonsynonymous mutations are both acted on by genetic drift.

Nonsynonymous mutations, are nucleotide mutations that change the amino acid to be inserted during protein translation. These mutations can arise due to both genetic drift or non-random selection events related to the fitness of an individual. Because they can change the structure or function of a protein to varying intensities, these mutations are more likely to modify an organism's reproductive fitness and phenotypic expression within the environment. To determine these effects of selection on an individual, nonsynonymous mutations are used. Genetic drift also affects nonsynonymous mutations when random fluctuations in alleles include changes resulting in non-

synonymous mutations. Both genetic drift and natural selection act on nonsynonymous mutations and D_{NON} alone cannot represent the forces of selection; therefore, by eliminating D_{SYN} (which implies only the effects of drift) from D_{NON} (which includes the effects of selection and drift) we can obtain a value that isolates only the effects of selection on the population. We plan to examine four patterns: purifying selection, diversifying selection, population bottleneck, and population subdivision.

The two defining sets of characteristics of these evolutionary patterns are: whether natural selection or genetic drift is acting on the population, and if an overall increase or decrease in genetic diversity is observed. A population bottleneck occurs when a population crashes to a smaller number of individuals such as in dramatic natural disasters. This decrease in population size is not selective, meaning genetic variation does not raise or lower an individual's reproductive fitness. The small population size resulting from a bottleneck causes an overall decrease in genetic diversity and D_{SYN} would be significantly low. Population subdivision is a form of drift that maintains genetic variation among sub-populations. The original population divides into subpopulations that are not permitted to mate with each other. Because fitness does not have any bearing on how the population is randomly separated, this is a form of genetic drift. Both sub-populations accumulate mutations independently in distinct environmental conditions, causing an overall increase in genetic diversity and a significantly positive D_{SYN} .

The effects of natural selection also may result in an increase or decrease of genetic diversity. In diversifying selection, the most genetically distinct individuals within the population have the greatest reproductive fitness within the population. In this situation, the difference $D_{NON} - D_{SYN}$ is significantly positive and indicates an increase of diversity. However, a significantly negative difference implies a decrease in diversity due to natural selection. Purifying selection, occurs when mutations causing more variation within the population are selected against. Genetically different individuals within the population have a lower reproductive fitness. In this case, natural selection has caused a decrease in diversity and $D_{NON} - D_{SYN}$ is significantly negative (see Figure 2).

	D < 0	D > 0
$D_{\text{NON}} - D_{\text{SYN}}$	Purifying Selection	Diversifying Selection
D_{SYN}	Population Bottleneck	Population Subdivision

Figure 2: Measuring separate D values for synonymous and nonsynonymous data can indicate one of the four evolutionary patterns shown here.

2 Current State of the Project

We are in the process of building simulation software to help us determine the accuracy of our procedure under different conditions. At present, our simulation code successfully generates populations under each of the different evolutionary constraints (e.g. purifying selection, diversifying selection, population bottlenecking, etc.), or any linear combination of those. It also lends feasibility to high-volume data generation, as it is 2700 times as fast as the existing Perl code for the neutral model, 900 times as fast for the other evolutionary trends, and supports massively parallel execution on a UNIX cluster.

One of the important aspects of this implementation is the orthogonality of evolutionary constraints. Each aspect of deviation from the neutral model can be activated with a given intensity, and multiple models may be combined. There are two ways this is done: First, selection models are treated as arbitrary functions acting on the individual within a population. Thus, any set of selection models may be combined. Second, the ability of each member of the population to generate offspring is determined by a matrix of the form:

$$\begin{array}{c}
 \text{Child Generation} \\
 \left[\begin{array}{ccccc}
 c_{1,1} & c_{1,2} & c_{1,3} & \cdots & c_{1,n} \\
 c_{2,1} & c_{2,2} & c_{2,3} & \cdots & c_{2,n} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 c_{n,1} & c_{n,2} & c_{n,3} & \cdots & c_{n,n}
 \end{array} \right]
 \end{array}
 \begin{array}{c}
 \text{Parental Generation}
 \end{array}$$

in which each $c_{i,j}$ determines the probability that member i of the parent generation produced member j of the child generation. So, for instance, to simulate a subdivision of a population of four members into two equally-sized groups, we would create a matrix that looks like this:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Children 1 and 2 (represented by rows 1 and 2 within the matrix) must have come from parents 1 and 2 (represented by columns 1 and 2 within the matrix), and children 3 and 4 from parents 3 and 4. Because no mating may occur across groups, the two population subgroups are genetically isolated. One of the advantages of using a continuous approach is that the two populations can be separated to various degrees. In some cases, we could divide the population completely as shown above so that the two sub-populations could not mate with each other at all. In other simulations, we may partially separate the two sub-populations so that they may occasionally mate with each other. We will use this flexibility to test the sensitivity of our model.

3 Future Directions

HIV has been chosen as our model organism because the selective pressures of the human body and HIV's high mutability and short generation time make it possible for more than one evolutionary force to influence an HIV population simultaneously. For example, diversifying selection and population bottlenecks may both occur simultaneously within an infected patient after medicinal treatment. HIV strains that have a unique genetic makeup, allowing them to remain unaffected by the treatment, will be able to survive and continue reproducing (i.e. diversifying selection). At the same time however, the treatment may eliminate a large percentage of the population and induce a population bottleneck. Using our analysis should enable us to discern the simultaneous forces of selection and drift.

Once the simulations modeling the various evolutionary patterns are functioning correctly and specific distributions of D_{SYN} and $D_{NON} - D_{SYN}$ are found, we plan to find an appropriate sample size to use D 's calculation. In practice, large sample sizes of HIV genetic code are about 20 sequences; therefore, we need to ensure this small sample size sufficiently identifies evolutionary forces acting upon the population of HIV. Obtaining distributions of the same evolutionary pattern

using various sample sizes (e.g. $N = 20, 50, 250,$ and 500) should allow a thorough analysis of the effects of sample size on the distributions of D values.

Prior to considering the evolutionary forces acting on the simulated populations, the validity of the neutral model will be rigorously analyzed. While testing the accuracy of the neutral model, simulations with population sizes of $N = 500$ and $N = 10,000$ will be executed and their D distributions will be compared. A small discrepancy between the smaller and larger sample sizes would indicate that population size is not contributing to variation in the D distributions.

Another inconsistency in the neutral model may reside in the way the D statistic measures mutations at one site. The D statistic was constructed under an infinite sites model, which assumes there is only one mutation per site in the population. Due to HIV's high mutation rate, a finite sites model, which allows more than one mutation to occur at each site, more accurately reflects the observed variation. We will create distributions of both D_{SYN} and $D_{NON} - D_{SYN}$ for the finite and infinite sites models in each of our evolutionary patterns by increasing and decreasing sequence length. If little discrepancy is found, we can assume the infinite D calculations can accurately represent finite HIV data; however, if a large discrepancy is found, we will search for a calculation modification that can relate finite sites data to infinite sites data.

Each evolutionary scenario will be run multiple times with different parameter values such as strength of selection, bottleneck intensity, and subdivision duration. In both selection models, the fitness (W), of individuals will be calculated to determine how likely each individual is to reproduce (see figure 3). How much these fitnesses alter an organism's survival and proliferation is a factor that will be varied to represent the strength of selection (s) acting on the population. In purifying selection, q quantifies how genetically deviant an individual is from the pre-determined majority of the population. At the beginning of the purifying selection simulation, $q = 0$ across the entire population and mutations are subsequently allowed to alter the genetic make-up and fitnesses of individuals within the population. Diversifying selection will determine an organism's fitness based on its relative diversity within the population (see figure 3). Relative diversity is represented by the variable r which is calculated by summing all possible pairwise distances for each individual in the population, and as r increases, so does the fitness of the individual. In both fitness equations

Neutral Model: Large population (N = 500), largest population (N = 10000)			
Selection Models: Strong (0.5), moderate (0.08), and weak (0.01) selection strengths		Drift Models: Where duration and end time are measured in generations	
Purifying	$W = (1 - s)^q$	Bottleneck	Intensity: 0.99, 0.92, 0.50 Duration: 500, 100, 20 End time: 1000, 800, 500
Diversifying	$W = (1 - s)^{-r}$	Subdivision	Split ratio: 50:50, 90:10 Duration: 800, 300, 100 End time: 1000, 800

Figure 3: Models and their Parameters

represented in Figure 3, the individual with the highest fitness value has the highest likelihood to reproduce.

Diversifying selection will determine an organism's fitness based on its relative diversity within the population due to its fitness equation (see Figure 3). Relative diversity is represented by the variable r which is calculated by summing all possible pairwise distances for each individual in the population. The sequence with the highest fitness will be more likely to have offspring.

Evolutionary trends that are related to genetic drift are population bottlenecks and population subdivisions. Distinct types of these trends will be run with varying parameters as described in detail in Figure 3. Their parameters will be modified to try to represent various types of evolutionary patterns populations may undergo. For population bottleneck models, the intensity (percent of the population eliminated in the crash), duration, and end time of the bottleneck will be adjusted.

Two types of D distributions (D_{SYN} and $D_{NON} - D_{SYN}$) will be created using 10,000 data points for each simulated evolutionary pattern. The critical values of these distributions will be the distinguishing characteristics of the evolutionary trends. However, we suspect the numerical ranges outside these various critical values will overlap, causing ambiguity in the interpretation of HIV's evolutionary history. One way to approach this ambiguity is by using Bayesian statistics. The Bayesian perspective uses likelihoods to estimate which model a given observed data point best fits. Under a frequentist model, we would simply decide to reject or fail to reject a given model depending on the p -value of an observed data point. Figure 4 better illustrates the benefit in using

Model	P-value	Bayesian Likelihoods	Frequentist Likelihoods
Neutral	0.021	0.053	Reject
Purifying/Directional	0.135	0.3435	Fail to Reject
Diversifying	0.006	0.015	Reject
Population Bottleneck	0.217	0.5521	Fail to Reject
Population Subdivision	0.014	0.0356	Reject

Figure 4: Bayesian and Frequentist Approaches

a Bayesian versus a frequentist approach.

The Bayesian approach’s likelihoods are calculated by dividing the data’s p-value, derived from the various evolutionary models, by the sum of all of the p-values. In this example, we can see that the observed value is most likely a population bottleneck with a likelihood of 0.5521 under the Bayesian perspective. The frequentist perspective simply “fails to reject” a population bottleneck model. The Bayesian approach elicits more information than can be derived from frequentist statistics.

A Bayesian interpretation of the distributions of D_{SYN} and $D_{NON} - D_{SYN}$ will allow us to determine the likelihood that certain evolutionary patterns have occurred. One assumption of this analysis is that synonymous data can only represents the effects of genetic drift. It is now known however, that synonymous changes can have a selective affect on the organism in specific instances (such as when they occur within crucial consensus sequences at splice sites). These types of synonymous mutations typically occur at splicing sites, cause an irregular loss or gain of exons within an mRNA before translation, and can greatly affect the functionality of the resulting protein. Accounting for splice sites is not currently incorporated into simulations, but will be an area of intense research and examination in the near future. Accounting for these splice sites will provide more detailed models and still allow our short term goal of separating the effects of genetic drift and natural selection. A long-term goal is to make our methods applicable to any system, as it is currently catered for HIV. Ideally, by using our new innovation, we will provide the means to unveil information about the evolutionary history of any given population.

References

- [1] K. L. Simonson, G. A. Churchill, and C. F. Aquadro. Properties of statistical tests of neutrality for dna polymorphism data. *Genetics*, 141:413–429, 1995.
- [2] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123:585–595, 1989.
- [3] Fumio Tajima. The amount of dna polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, 143:1457–1465, 1996.