DERLEME REVIEW

DOI: 10.5336/biostatic.2016-53446

Statistical Tests for Neutrality: Review

Nötralite İçin İstatistiksel Testler

İsmet DOĞAN,[a]
Nurhan DOĞAN[a]

[a]Department of
Biostatistics and Medical Informatics,
Afyon Kocatepe University
Faculty of Medicine, Afyonkarahisar

Geliş Tarihi/*Received:* 30.09.2016
Kabul Tarihi/*Accepted:* 14.11.2016

Yazışma Adresi/*Correspondence:*
Nurhan DOĞAN
Afyon Kocatepe University
Faculty of Medicine,
Department of
Biostatistics and Medical Informatics,
Afyonkarahisar,
TÜRKİYE/TURKEY
nurhandogan@hotmail.com

ABSTRACT Neutrality tests are used to test whether there is balance in the populations. And they are also used to test whether there is selection on an allele from a large number of alleles in the populations. Statistical tests for neutrality are important and useful tools for population genetics. Since the development of molecular genetics techniques allowed to obtain nucleotide sequences for the study of population genetics, a number of neutrality tests have been developed with the objective to facilitate the interpretation of an increasing volume of molecular data. Since Motoo Kimura (1968) first suggested that most polymorphisms are selectively neutral testing the neutral hypothesis has been one of the prime objectives of molecular population genetics. A number of statistical tests for neutrality have been developed in recent decades. Statistical tests for neutrality have been widely used in population genetics analyses, not only to reject the neutral theory but also as summary statistics to facilitate the interpretation of DNA sequence data in populations. Although most of these tests were originally developed to detect the effect of positive selection, they are also affected by demographic processes. An important family of neutrality tests is based on the frequency spectrum of nucleotide polymorphisms. The classical tests of this family are proposed by Tajima (1989), Fu and Li (1993), Fay and Wu (2000) and by Zeng (2006). This article focuses on describing these statistical methods and their rationale. Simulations indicate that Tajima's test is generally most powerful among tests within this class.

Keywords: Neutrality tests; nucleotide polymorphism; population genetics

ÖZET Nötralite testleri, populasyonların dengede olup olmadığını ve populasyonlarda var olan çok sayıdaki alel içerisindeki herhangi bir alel üzerinde seçilim olup olmadığını test etmek amacıyla kullanılmaktadır. Nötralite için istatistiksel testler popülasyon genetiği için önemli ve yararlı araçlardır. Popülasyon genetiği çalışmaları için nükleotid dizilerinin elde edilmesine izin veren moleküler genetik tekniklerinin geliştirilmesinden itibaren, büyük hacimli moleküler verilerin yorumlanmasını kolaylaştırmak amacı bir dizi nötralite testi geliştirilmiştir. İlk kez Motoo Kimura (1968) tarafından önerilmesinden itibaren polimorfizmlerin çoğunun seçici nötral testi için nötral hipotezler, popülasyon genetiğinin temel amaçlarından biri olmuştur. Son yıllarda nötralite için bir dizi istatistiksel test geliştirilmiştir. Popülasyon genetiği analizlerinde yaygın olarak kullanılan istatistiksel testler, sadece nötral teoriyi reddetmek için değil aynı zamanda özetleyici istatistikler gibi popülasyonlarda DNA dizisi verilerinin yorumlanmasını kolaylaştırmak için de kullanılmaktadırlar. Bu testlerin çoğu başlangıçta pozitif seçimin etkisini tespit etmek için geliştirilmiş olmasına rağmen, demografik süreçler tarafından etkilenmektedirler. Nötralite testlerinin önemli bir ailesi nükleotid polimorfizmlerinin frekans spektrumuna dayanmaktadır. Bu ailenin klasik testleri Tajima (1989), Fu ve Li (1993), Fay ve Wu (2000) ile Zeng (2006) tarafından önerilen testlerdir. Çalışma, bu istatistiksel yöntemlerin tanımlanmasına ve bunların gerekçelerinin açıklanmasına odaklanmaktadır. Simülasyonlar, Tajima testinin bu sınıf içindeki testler arasında en güçlü test olduğunu göstermektedir.

Anahtar Kelimeler: Nötralite testleri; nükleotid polimorfizm; populasyon genetiği

DNA polymorphisms are powerful sources of information for studying the evolution of a population. Whether a locus or region from which a DNA sample has been taken evolves neutrally or under natural selection is of considerable interest in evolutionary study and can be examined using a statistical test designed for DNA polymorphisms.[1] Comparison of DNA sequences within and between species is a powerful approach not only for determining the evolutionary forces acting in specific gene regions but also for determining relevant aspects of the evolutionary history of the species.[2] The amount and pattern of polymorphism in DNA sequence samples from a population reflects not only mutations in the ancestors of the sequences but also random genetic drift as well as other evolutionary forces, such as natural selection. How to detect the presence of natural selection in molecular population genetics and evolution is an important issue. It is possible to detect the presence of natural selection because natural selection often causes the pattern of polymorphism to differ from the under the neutral mutation hypothesis, which postulates that the majority of mutations that have contributed significantly to the genetic variation in natural populations is neutral or nearly neutral.[3]

The ascertainment of the demographic and selective history of populations has been a major research goal in genetics for decades. To that end, numerous statistical tests have been developed to detect deviations between expected and observed frequency spectra, *e.g.*, Tajima's *D*, Fu and Li's *D** and Fay and Wu's *H*.[4] Whether the observed pattern of polymorphism in a set of DNA sequences is consistent with a neutral model of evolution is of great interest to the study of evolution. Several statistical tests are available for testing, for a sample of DNA sequences from a population. These tests are often referred to as tests of neutrality.[5] Assumptions of statistical tests of neutrality are [6]:

- Polymorphic sites are selectively neutral
- Random mating
- Samples from a large, constant, diploid population size *N* individuals
- Non-overlapping generations
- No migration
- No recombination within a locus
- No repeated mutations at the same site for inter-specific divergence, that is, no multiple hits between species
- Infinite sites (every mutation is scored as a polymorphism, that is, no multiple hits within species)
- Constant neutral mutation rate

An increasing number of statistical tests have been developed to detect departures of DNA sequence variability from the expectations of the neutral theory of evolution. Several tests have been proposed to detect departures of nucleotide variability patterns from neutral expectations. However, very different kinds of evolutionary processes, such as selective events or demographic changes, can produce similar deviations from these tests, thus making interpretation difficult when a significant departure of neutrality is detected.[7]

A number of different statistics are used for detecting natural selection using DNA sequencing data including statistics that are summaries of the frequency spectrum, such as Tajima's D, Fay and Wu's H, Fu and Li's D* and Zeng's E. In the past decade there has been considerable interest in detecting natural selection in humans and other organisms from DNA sequence data. An often used approach for detecting selection is to use a neutrality test statistic based on allele frequencies. Such frequency based tests have been used to identify a number of genes, that have undergone selection.[8]

The advent of DNA sequence data now makes possible a quantitative comparison of levels of nucleotide polymorphism within populations and similar levels of sequence divergence between populations. Thus it is now important to consider statistical tests of the neutral theory that are based on both kinds of data. For example, one might ask, given different levels of nucleotide divergence between species in two or more regions of DNA, whether the levels of within-species polymorphism in the corresponding regions differ in the appropriate way, as predicted by the neutral theory.[9] Whether a locus evolves neutrally or under the influence of natural selection is of great interest in molecular evolutionary study. Statistical tests can be used to test if the observed polymorphisms in a DNA sample are consistent with the prediction of neutral evolution.[10] A number of authors have developed several methods of statistical inference and statistical tests using different approaches. We focus on few of the most commonly used tests. In this study, neutrality tests which are using population genetic data have been based on frequency distribution of segregation sites at multiple loci are introduced.

TAJIMA'S D[11]

For nucleotide data, one of the most popular tests is Tajima's $D\ test.$ Tajima's D is the scaled difference in the estimate of $\theta = 4N_e\mu$ (N_e= effective population size, μ= mutation rate per generation) based on the number of pairwise differences and the number of segregating sites in a sample of nucleotide sequences. It is defined as

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{Var(\theta_\pi - \theta_W)}} = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{d}{\sqrt{e_1 S + e_2 S(S-1)}}$$

$$e_1 = \frac{c_1}{a_1}$$

$$e_2 = \frac{c_2}{a_1^2 + a_2}$$

$$c_1 = b_1 - \frac{1}{a_1}$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

$$b_1 = \frac{n+1}{3(n-1)}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

$$M = \frac{S}{a_1}$$

$$d = the\ average\ number\ of\ polymorphisms - M$$

n : is the sample size (the number of DNA sequences studied),

S : the number of segregating sites,

If the value of D is too large or too small the neutral null hypothesis is rejected. A very rough rule of thumb to significance is that values greater than +2 or less than -2 are likely to be significant. This rule is based on an appeal to asymptotic properties of some statistics, and thus +/- 2 does not actually represent a critical value for a significance test. The critical values are given in Fu and Li (1993).

$D < 0$: Rare alleles present at low frequencies / Recent selective sweep, population expansion after a recent bottleneck, linkage to a swept gene,

$D = 0$: Observed variation similar to expected variation / Population evolving as per mutation-drift equilibrium. No evidence of selection,

$D > 0$: Multiple alleles present, some at low, others at high frequencies / Balancing selection, sudden population contraction,

The foregoing is illustrated in the example given below. Consider the following data set:

Seq1 AACTGTGCACTGCATGATGA
Seq2 AACTGTGCACTGCATGATGA
Seq3 AAGTGTGCACTGCCTGATGA
Seq4 AAGTGTGCACTGCCTGATGA
Seq5 AACTGTGCACTGCATGATGA
Seq6 AACTGTGCACTGCATGATGA
Seq7 AACTGTGCACTGCATGCTGA
Seq8 AACTGTGCACTGCATGATGA

$n = 8 \quad S = 3$

The difference values obtained in pairwise comparisons of the sequences are given in Table 1.

From this data, we can calculate,

$$a_1 = \sum_{i=1}^{7} \frac{1}{i} = 2.592857$$

$$a_2 = \sum_{i=1}^{7} \frac{1}{i^2} = 1.511797$$

$$b_1 = \frac{n+1}{3(n-1)} = 0.428571$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)} = 0.297619$$

$$c_1 = b_1 - \frac{1}{a_1} = 0.042896$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2} = 0.040398$$

$$e_1 = \frac{c_1}{a_1} = 0.016544$$

$$e_2 = \frac{c_2}{a_1^2 + a_2} = 0.004906$$

TABLE 1: Numbers of difference for pairwise sequence comparisons.

Pairwise comparison of sequences	Number of comparison	Number of pairwise difference
Seq1-Seq2	1	0
Seq1-Seq3	2	2
Seq1-Seq4	3	2
Seq1-Seq5	4	0
Seq1-Seq6	5	0
Seq1-Seq7	6	0
Seq1-Seq8	7	0
Seq2-Seq3	8	2
Seq2-Seq4	9	2
Seq2-Seq5	10	0
Seq2-Seq6	11	0
Seq2-Seq7	12	1
Seq2-Seq8	13	0
Seq3-Seq4	14	0
Seq3-Seq5	15	2
Seq3-Seq6	16	2
Seq3-Seq7	17	3
Seq3-Seq8	18	2
Seq4-Seq5	19	2
Seq4-Seq6	20	2
Seq4-Seq7	21	3
Seq4-Seq8	22	2
Seq5-Seq6	23	0
Seq5-Seq7	24	1
Seq5-Seq8	25	0
Seq6-Seq7	26	1
Seq6-Seq8	27	0
Seq7-Seq8	28	0
Total	**29**	

$$M = \frac{S}{a_1} = \frac{3}{2.592857} = 1.157025$$

$$d = the\ average\ number\ of\ polymorphisms - M = \frac{29}{28} - 1.157025 = -0.121311$$

$$D = \frac{d}{\sqrt{e_1 S + e_2 S(S-1)}} = \frac{-0.121311}{\sqrt{0.016544 * 3 + 0.004906 * 3 * 2}} = \frac{-0.121311}{0.281190} = -0.431420$$

According to this result, it could be said that polymorphism between sequences to be negligible.

FAY AND WU'S H STATISTIC[12]

$$H = \theta_\pi - \theta_H$$

$$\theta_\pi = \sum_{i=1}^{n-1} \frac{2S_i i(n-i)}{n(n-1)}$$

$$\theta_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)}$$

S_i : the number of derived variants found i times in a sample of n chromosomes,

NORMALIZED FAY AND WU'S H STATISTIC[13]

$$H = \frac{\theta_\pi - \theta_L}{\sqrt{Var(\theta_\pi - \theta_L)}}$$

$$Var(\theta_\pi - \theta_L) = \frac{n-2}{6(n-1)}\theta + \frac{18n^2(3n+2)b_{n+1} - (88n^3 + 9n^2 - 13n + 6)}{9n(n-1)^2}\theta^2$$

$$\theta = \theta_w$$

$$\theta_w = \frac{1}{a_n}\sum_{i=1}^{n-1}\xi_i$$

$$\theta^2 = \frac{s(s-1)}{(a_n^2 + b_n)}$$

$$\theta_\pi = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i(n-i)\xi_i$$

$$\theta_L = \frac{1}{n-1}\sum_{i=1}^{n-1} i\xi_i$$

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i}$$

$$b_n = \sum_{i=1}^{n-1}\frac{1}{i^2}$$

n : is the sample size (the number of DNA sequences studied),

ξ_i : number of segregating sites where the mutant type occurs i times in the sample,

ZENG'S E STATISTIC[13]

$$H = \frac{\theta_L - \theta_W}{\sqrt{Var(\theta_L - \theta_W)}}$$

$$Var(\theta_L - \theta_W) = \left[\frac{n}{2(n-1)} - \frac{1}{a_n}\right]\theta + \left[\frac{b_n}{a_n^2} + 2\left(\frac{n}{n-1}\right)^2 b_n - \frac{2(nb_n - n + 1)}{(n-1)a_n} - \frac{3n+1}{n-1}\right]\theta^2$$

$$\theta_L = \frac{1}{n-1}\sum_{i=1}^{n-1} i\xi_i$$

$$\theta_w = \frac{1}{a_n}\sum_{i=1}^{n-1}\xi_i$$

$$\theta = \theta_w$$

$$\theta^2 = \frac{s(s-1)}{(a_n^2 + b_n)}$$

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i}$$

$$b_n = \sum_{i=1}^{n-1}\frac{1}{i^2}$$

n : is the sample size (the number of DNA sequences studied),

ξ_i : number of segregating sites where the mutant type occurs i times in the sample,

FU AND LI'S D* STATISTIC[14]

$$D^* = \frac{\eta - a_n \eta_e}{\sqrt{u_D \eta + v_D \eta^2}}$$

$$v_D = 1 + \frac{a_n^2}{b_n + a_n^2}\left(\frac{2na_n - 4(n-1)}{(n-1)(n-2)} - \frac{n+1}{n-1}\right)$$

$$u_D = a_n - 1 - v_D$$

$$\eta = \sum_{i=1}^{m} s_i$$

$$\eta_e = \sum_{i=1}^{m} e_i$$

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

$$b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

η : total number of mutations,

s_i : number of different nucleotides minus one at site i among the n sequences,

e_i : number of singleton nucleotides,

CONCLUSION

Several genomic sequencing projects have been recently completed or are close to completion. Such data are perfectly suited for scanning the genome for sites at which positive selection has occurred. Several authors have argued that positive selection might be frequent in the genomes of humans and other organisms. If this is true, we have the necessary statistical methods for identifying which sites have undergone selection based on comparative data. Identifying selection in the genome might very well become one of our most powerful tools for identifying causes for species specific differences and for identifying genomic regions of functional, and perhaps, medical importance.[15] Since Motoo Kimura (1968) first suggested that most polymorphisms are selectively neutral testing the neutral hypothesis has been one of the prime objectives of molecular population genetics.[16] Statistical tests of neutrality use a snapshot of the present to infer the past. These tests allow us to assess selection at the molecular level directly from the standing variation in natural populations, rather than extrapolating from laboratory-adapted and/or inbred populations to the wild. Furthermore, statistical tests of neutrality can be used to detect selection in species less suited to laboratory culture, because these tests do not require genetic manipulation of the organism.[6] Tajima's D statistic is probably the most widely used test of neutrality. Simulations indicate that Tajima's test is generally most powerful against the alternative hypotheses of selective sweep, population bottleneck, and population subdivision, among tests within this class.[17] Tajima's D is optimal against a spectrum with an excess of intermediate alleles and a defect of low- frequency alleles, while Fu and Li's D^* is sensitive to very rare alleles and Fay and Wu's H is optimal against an excess of high-frequency alleles and a defect of low- frequency alleles.[4] The practical advantage of the Tajima test is that it can be conducted on sequences from any locus (coding or noncoding) of any species (no outgroup is required). Hence, conservation geneticists engaged in sequence studies of various species can readily conduct this test.[18]

Authorship Contributions
Idea/Concept: *İsmet Doğan*
Design: *İsmet Doğan*
Control/Supervision: *İsmet Doğan, Nurhan Doğan*
Data Collection and/or Processing: *İsmet Doğan, Nurhan Doğan*
Analysis and/or Interpretation: *İsmet Doğan, Nurhan Doğan*
Literature Review: *İsmet Doğan, Nurhan Doğan*
Writing the Article: *İsmet Doğan, Nurhan Doğan*
Critical Review: *İsmet Doğan, Nurhan Doğan*
References and Fundings: *İsmet Doğan, Nurhan Doğan*
Materials: *İsmet Doğan, Nurhan Doğan*
Conflict of Interest Statement
Approve that they do not derive any personal benefits (financial, etc.) from this study

REFERENCES

1. Fu YX. New statistical tests of neutrality for DNA samples from a population. Genetics 1996;143(1):557-70.
2. Ramos-Onsins SE, Rozas J. Statistical properties of new neutrality tests against population growth. Mol Biol Evol 2002;19(12):2092-100.
3. Li H, Zhang Y, Zhang YP, Fu YX. Neutrality tests using DNA polymorphism from multiple samples. Genetics 2003;163(3):1147-51.
4. Ferretti L, Perez-Enciso M, Ramos-Onsins S. Optimal neutrality tests based on the frequency spectrum. Genetics 2010;186(1):353-65.
5. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 1997;147(2):915-25.
6. Wayne ML, Simonsen KL. Statistical tests of neutrality in the age of weak selection. Trends Ecol Evol 1998;13(6):236-40.
7. Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics 2008;179(1):555-67.
8. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 2013;14:289.
9. Hudson RR, Kreitman M, Aguadé M. A test neutral molecular evolution based on nucleotide data. Genetics 1987;116(1):153-9.
10. Fu YX. Neutrality and selection in molecular evolution: statistical tests. 2001. Doi:10.1038/npg.els.0001809.
11. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123(3):585-95.
12. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 2000;155(3):1405-13.
13. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 2006;174(3):1431-9.
14. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993;133(3):693-709.
15. Nielsen R. Statistical tests of selective neutrality in the age of genomics. Heredity (Edinb) 2001;86(Pt 6):641-7.
16. Kimura M. Evolutionary rate at the molecular level. Nature 1968;217(5129):624-6.
17. Simonsen KL, Churchill GA, Aquadro CF. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 1995;141(1):413-29.
18. Rand DM. Neutrality tests of molecular markers and the connection between DNA polymorphism, demography, and conservation biology. Conservation Biology 1996;10(2):665-71.