Published in final edited form as:
Biometrics. 2012 December ; 68(4): . doi:10.1111/j.1541-0420.2012.01775.x.

A Note on Two-Sample Tests for Comparing Intra-Individual Genetic Sequence Diversity between Populations

E. E. Giorgi[1,2,*] and **T. Bhattacharya**[1,3,**]

[1]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, U.S.A.

[2]Univeristy of Massachusetts, Amherst, Massachusetts 01002, U.S.A.

[3]The Santa Fe Institute, Santa Fe, New Mexico 87501, U.S.A.

Summary

present four U-statistic based tests to compare genetic diversity between different samples. The proposed tests improved upon previously used methods by accounting for the correlations in the data. We find, however, that the same correlations introduce an unacceptable bias in the sample estimators used for the variance and covariance of the inter-sequence genetic distances for modest sample sizes. Here, we compute unbiased estimators for these and test the resulting improvement using simulated data. We also show that, contrary to the claims in Gilbert et al., it is not always possible to apply the Welch–Satterthwaite approximate t-test, and we provide explicit formulas for the degrees of freedom to be used when, on the other hand, such approximation is indeed possible.

Keywords

HIV genetic diversity; Hypothesis testing; Nonparametric statistics; Two-sample test; U-statistic

1. Introduction

Gilbert, Rossini, and Shankarappa (2005) present a test for comparing intra-individual genetic sequence diversity between populations. This has proven to be a very useful test since intra-individual viral sequence diversity is a commonly used and readily calculated summary statistic that often has interesting biological correlates (Heath et al., 2010, 2009). Unfortunately, we discovered that the biased estimators for the variances and covariances of inter-sequence genetic distances used there make it difficult to apply the test to modestly sized samples, which typically occur in practice. Even though for small sample sizes a permutation test becomes feasible, in reality it is inconvenient to use one statistic for small samples and a different one for larger ones. In this note, we correct the bias in the original statistic and verify the resulting improvement by explicit simulation of the appropriate null model. We only explicitly address how to correct the first of the four tests proposed in Gilbert et al. (2005), but the correction for the other three tests follows by extension of the same methods and using the corrected variance.

Following the notation in Gilbert et al. (2005), let $\hat{\mu}$ be the sample mean, also called empirical mean, of the pairwise Hamming distances D_{ij} between sequences i and j from a sample of size N

[*]egiorgi@lanl.gov. [**]tanmoy@lanl.gov.

$$\widehat{\mu}=\left[\frac{N\left(N-1\right)}{2}\right]^{-1}\sum_{i<j}D_{ij}. \quad (1)$$

In the simplest case of two independent samples with mean Hamming distance, respectively, μ_1 and μ_2, Gilbert et al. (2005) define the following statistics for evaluating the null hypothesis $H_0 : \mu_1 = \mu_2$:

$$T_{\text{pooled}}=\frac{\widehat{\mu}_1-\widehat{\mu}_2}{\sqrt{\text{Var}\left(\widehat{\mu}_1\right)+\text{Var}\left(\widehat{\mu}_2\right)}}, \quad (2)$$

which is approximately normally distributed for large sample sizes. For small samples, they suggest the use of Welch–Satterthwaite approximation to calculate the appropriate degrees of freedom for a t-test.

This proposed test needs $\text{Var}\left(\widehat{\mu}\right)$, the sampling variance of $\widehat{\mu}$. The problem with estimating this variance is that the D_{ij}'s are not independent observations. To circumvent the problem, Gilbert et al. suggest to use the theory of U-statistics (Lee, 1990) to estimate it from the variances and pairwise covariances of the Hamming distances. The advantage of this is that the results are blind to the nature of the correlation, and only assume that the sequences are sampled independently in each subject. This independent sampling means that the population variance–covariance matrix can be written in terms of only two independent parameters. Again, following previous notation, let $\sigma_2^2=\text{Var}\left(D_{ij}\right)$ be the variance of the distribution of Hamming distances, and $\sigma_1^2=\text{Cov}\left(D_{ij},D_{ik}\right)$ be the covariance between the Hamming distances of pairs sharing one sequence.

Then, it is shown in Gilbert et al. (2005) that the required sampling variance is given by

$$\text{Var}\left(\widehat{\mu}\right)=\left[\frac{N\left(N-1\right)}{2}\right]^{-1}\left[2\left(N-2\right)\sigma_1^2+\sigma_2^2\right]. \quad (3)$$

This formula requires the estimation of σ_1^2 and σ_2^2 from the sample. The estimators used in Gilbert et al. (2005), however, have an $O(1/N)$ bias arising from the same nonindependence of the D_{ij}. Note that this bias is of the same order as the contribution of the σ_2 term itself, and turns out to be numerically important for the sample sizes of interest in many applications. Here we determine the corresponding unbiased estimators and use a simulation to compare our improvement over the test proposed in Gilbert et al. (2005).

2. Unbiased Estimators

Let N be the number of sequences sampled, and D_{ij} the Hamming distance between sequence i and sequence j. We define

$$\begin{aligned}\tilde{\sigma}_1^2\equiv&\sum_{i<j<l}\left[\left(D_{ij}-\widehat{\mu}\right)\left(D_{il}-\widehat{\mu}\right)+\left(D_{ij}-\widehat{\mu}\right)\left(D_{jl}-\widehat{\mu}\right)\right.\\&\left.+\left(D_{il}-\widehat{\mu}\right)\left(D_{jl}-\widehat{\mu}\right)\right]\end{aligned} \quad (4)$$

and

$$\tilde{\sigma}_2^2\equiv\sum_{i<j}\left(D_{ij}-\widehat{\mu}\right)^2. \quad (5)$$

The expression for $\hat{\sigma}_1^2$ is proportional to a symmetrized version of the estimator $\hat{\sigma}_1^2$ used in Gilbert et al. (2005). This symmetrization is purely cosmetic, though our definition has a lower sampling variance than the one used in Gilbert et al. (2005). The improvement demonstrated later is, however, not sensitive to this precise choice. Note also that both $\hat{\sigma}_1^2$ and $\hat{\sigma}_1^2$ can be negative.

We aim to find unbiased estimators of σ_1^2 and σ_2^2 in the linear span of these two statistics. In terms of $E(D_{ij} \equiv \mu$, we can write $E\left(D_{ij}^2\right)=\sigma_2^2+\mu^2, E\left(D_{ij}D_{ik}\right)=E\left(D_{ij}D_{ki}\right)=\sigma_1^2+\mu^2$ and $E(D_{ij}\ D_{kl}) = E(D_{ij})E(D_{kl}) = \mu^2$, when i, j and k, l refer to nonoverlapping indexes. A little algebra shows that

$$S^2 \equiv \frac{4\left(2\tilde{\sigma}_1^2+\tilde{\sigma}_2^2\right)}{N\left(N-1\right)\left(N-2\right)\left(N-3\right)} \quad (6)$$

is an unbiased estimator of $\mathrm{Var}\left(\hat{\mu}\right)$, whereas $\hat{S}^2$, the statistic defined in Gilbert et al. (2005), has a $O\left(\frac{1}{N}\right)$ bias :

$$E\left(\hat{S}^2\right)=\mathrm{Var}\left(\hat{\mu}\right)+4\frac{N^2-2N+4}{N^3-3N^2+2N-3}\sigma_2^2. \quad (7)$$

In Table 1, we show the performance of our variance estimator S^2 by comparing it with the true $\mathrm{Var}\left(\hat{\mu}\right)$ and the estimator $\hat{S}^2$ defined in Gilbert et al. (2005) for different values of N.

Notice that neither S^2 nor $\hat{S}^2$ is guaranteed to be positive since both $\tilde{\sigma}_1^2$ and $\hat{\sigma}_1^2$ can be negative and large.

We note that the inter-sequence distances are not normally distributed, in fact they are always positive. Nevertheless, we show by simulations below that, even for rather small samples, the use of equation (6) to estimate the denominator in (equation 2) leads to an acceptable test statistic.

3. Simulations

To test our formulas, we simulate a population of independently evolved random sequences, i.e., sequences that can be thought of as being drawn from a star phylogeny. We then randomly draw two samples and test the difference in means using both the statistic defined in Gilbert et al. (2005) and our improved one. We then repeated the simulation drawing from a nonstar phylogeny population. To compare directly with the explicit formulas provided in Gilbert et al. (2005), we display and discuss later the result of using a normal approximation.

4. Welch–Satterthwaite Approximation

To extend these results, we attempt to design an approximate t-test following the Welch–Satterthwaite approximation advocated in Gilbert et al. (2005). Recall that a random variate T is t-distributed if there exist independent variates Z and C such that $T=\frac{z}{\sqrt{C/\nu}}$, where Z is normally distributed, and C is χ^2-distributed with degrees of freedom ν.

Consider a set of N sequences with pairwise Hamming distances $\{D_{ij}\}$ drawn from a population described by μ, σ_1, σ_2 defined in the Introduction. Let σ be the variance–

covariance matrix of $\mathbf{D} \equiv \{D_{ij}\}$. It is easy to see that $\mathbf{\Sigma = -\sigma_1^2 L}_{\mathrm{N}} + \left[\boldsymbol{\sigma_2^2 + 2\,(N-2)\,\sigma_1^2}\right]\mathbf{I_N}$, where $\mathbf{I_N}$ is the N-dimensional identity matrix, and $\mathbf{L_N}$ is the graph Laplacian of the edge graph of the complete graph on N vertices. σ can be easily diagonalized by diagonalizing $\mathbf{L_N}$ using the representations of the permutation group of N elements. Using Cochran's theorem (Cochran, 1934), one can then find independent chi-square variates C_{N-1} and $C_{N(N-3)/2}$ with degrees of freedom $N - 1$ and $N(N - 3)/2$, respectively, such that

$$S^2 = \frac{1}{a+b}\left[a\frac{C_{N-1}}{N-1} + b\frac{C_{N(N-3)/2}}{N\,(N-3)\,/2}\right], \quad (8)$$

$$\text{where } a = \frac{2\,(N-1)}{N-2}\left[\sigma_2^2 + (N-4)\,\sigma_1^2\right], \quad (9)$$

$$b = -\frac{N}{N-2}\left(\sigma_2^2 - 2\sigma_1^2\right). \quad (10)$$

However, note that b can be negative, and hence S^2 is not positive semidefinite. This problem is shared by the expression used in Gilbert et al. (2005), and distinguishes this problem from that considered by Satterthwaite and Welch (Welch, 1947; Satterthwaite, 1946). The simple case $b = 0$ is when the intersequence Hamming distances D_{ij} can be written as the sum $D_i^0 + D_j^0$ of distances of each sequence from a common ancestor; i.e., when the underlying sequence set is *phylogenetically* independent, or, in other words, when the phylogeny is "star-like." In the more general case, S^2 is not positive definite but has the same unit mean and the same variance as a scaled chi-square variate with degrees of freedom given by

$$\nu = \frac{(a+b)^2}{\frac{a^2}{\nu_2} + \frac{b^2}{\nu_3}} \quad (11)$$

$$= \frac{\left[\sigma_2^2 + 2\,(N-2)\,\sigma_1^2\right]^2}{\frac{4(N-1)}{(N-2)^2}\left[\sigma_2^2 + (N-4)\,\sigma_1^2\right]^2 + 2\frac{N}{(N-2)^2(N-3)}\left(\sigma_2^2 - 2\sigma_1^2\right)^2}. \quad (12)$$

A simple estimate of ν (not unbiased) in terms of $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ is

$$\nu = \frac{(N-2)^2\left(\tilde{\sigma}_2^2 + 2\tilde{\sigma}_1^2\right)}{4\,(N-3)\left(\tilde{\sigma}_2^2 + \tilde{\sigma}_2^1\right) + 2\frac{N}{N-3}\left[(N-4)\,\tilde{\sigma}_2^2 - 2\tilde{\sigma}_1^2\right]}. \quad (13)$$

Even though this is not the Welch–Satterthwaite approximation in sensu stricto because of the negative coefficient (equation (10)), in Table 2, we show the efficacy of using this estimate of the degrees of freedom to improve the z-test to a t-test for the small sample case with a sample with typical distances set to a moderate value of 20. In the case of star topology, for which the problematic term vanishes, this t-test has almost its nominal size (0.05), and even the normal approximation (z-test) using the unbiased estimators performs reasonably down to a sample size of about 30. In contrast, using the test presented in Gilbert et al. (2005), rather large deviations are seen even at a sample size of 50. For the nonstar case, the t-test becomes noticeably over-conservative below a sample size of about 20, but the normal approximation errs in the opposite direction, and both our corrections perform noticeably better than the test presented in Gilbert et al. (2005).

5. Discussion

We have improved the statistics proposed in Gilbert et al. (2005) and used a simulation to validate it. We have also computed the degrees of freedom for performing an approximate t-test using the Welch–Sattherthwaite formula. The t-test has its nominal size when the underlying sequences have no phylogenetic structure, and is conservative otherwise. The original test, on the other hand, provides an inadequate control of type I errors even for moderate sample sizes. For example, we requested the data used by Heath et al. (2009) and were able to compare our p-values to the ones the authors had obtained using the statistic proposed by Gilbert et al. (2005). One particular comparison of two samples, of sizes 13 and 8, respectively, is illustrative: it yielded a very low p-value (0.003) using the test proposed in Gilbert et al., whereas our statistic gives $p = 0.058$ with the normal approximation and $p = 0.059$ with the Satterthwaite–Welch t-test. We also found a pair where the normal approximation failed due to the nature of the data: the statistics in Gilbert et al. (2005) and the one developed here yielded $p = 0.07$ and $p = 0.0002$, respectively, but the p-value from the Satterthwaite–Welch t-test was $p = 0.68$.

The R code implementing the new proposed statistic is available for download at “ftp://ftp-t10.lanl.gov/pub/TwoSampleTTest/.”

Acknowledgments

The authors thank Peter Hraber and Bette Korber for discussions and suggestions, and Peter Gilbert for a critical reading of the manuscript. This work was supported by the Center for HIV/AIDS Vaccine Immunology, NIH grants U19-AI067854-05, AI28433-19, and RR06555-18.

References

Cochran W. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. Mathematical Proceedings of the Cambridge Philosophical Society. 1934; 30:178–191.

Gilbert B, Rossini AJ, Shankarappa R. Two-sample tests for comparing intra-individual genetic sequence diversity between populations. Biometrics. 2005; 61:106–117. [PubMed: 15737083]

Heath L, Fox A, McClure J, Diem K, Van’t Wout A, Zhao H, Park D, Schouten J, Twigg HLI, Corey L, Mullins J, Mittler J. Evidence for limited genetic compartmentalization of HIV-1 between lung and blood. PLoS One. 2009; 4:e6949. [PubMed: 19759830]

Heath L, Conway S, Jones L, Semrau K, Nakamura K, Walter J, Decker W, Hong J, Chen T, Heli M, Sinkala M, Kankasa C, Thea D, Kuhn L, Mullins J, Aldrovandi G. Restriction of hiv-1 genotypes in breast milk does not account for the population transmission genetic bottleneck that occurs following transmission. PLoS One. 2010; 5:e10213. [PubMed: 20422033]

Lee, A. U-Statistics-Theory and Practice. Marcel-Dekker; New York: 1990.

Satterthwaite F. An approximate distribution of estimates of variance components. Biometrics Bulletin. 1946; 2:110–114. [PubMed: 20287815]

Welch B. The generalization of “student’s” problem when several different population variances are involved. Biometrika. 1947; 34:28–35. [PubMed: 20287819]

Table 1

Comparison between $\mathrm{Var}\,(\hat{\mu})$, the $\widehat{S}^2$ estimator defined in Gilbert et al. (2005), and our estimator S^2 for different sample size N. All quantities were calculated through 10,000 runs from the simulation described in Section 3. In the star topology, the branch lengths of the phylogenetic tree were distributed according to a Poisson of mean 10. The non-star topology consisted of two stars with Poisson distribution with means 10 and 5 connected by a branch of length 12.

	Star topology			**Non-star topology**		
N	**Var($\hat{\mu}$)**	$\hat{S}^2$	S^2	**Var($\hat{\mu}$)**	$\hat{S}^2$	S^2
4	9.939	1.972	9.96	19.6995	6.9244	19.281
5	7.889	2.571	7.969	14.395	7.425	15.144
8	4.98	2.713	5.008	8.291	5.738	8.158
10	3.973	2.513	4.002	6.27	4.899	6.223
20	2.051	1.61	2.003	2.972	2.856	2.955
30	1.312	1.158	1.331	1.893	1.97	1.92
50	0.799	0.737	0.7998	1.135	1.215	1.13
100	0.399	0.384	0.3998	0.558	0.623	0.559

Table 2

Fraction of 10,000 samples that result in p ≤ 0.05 for a variety of tests. The column N gives the sample size, whereas the columns GRS, z-test, and t-test refer to the tests as described in Gilbert et al. (2005), the z-test implemented with our unbiased variance estimator, and the t-test implemented using the Satterthwaite and Welch approximation. In the star topology, the branch lengths of the phylogenetic tree were distributed according to a Poisson of mean 10. The nonstar topology consisted of two stars with Poisson distribution with means 10 and 5 connected by a branch of length 12. The Satterthwaite–Welch approximation would have been exact for the star topology if the branch lengths were distributed normally.

	Star topology			Nonstar topology		
N	GRS	z-test	t-test	GRS	z-test	t-test
4	0.280	0.101	0.050	0.099	0.070	0.020
5	0.260	0.084	0.049	0.126	0.097	0.016
8	0.151	0.069	0.048	0.123	0.085	0.023
10	0.126	0.067	0.054	0.106	0.070	0.032
20	0.081	0.058	0.051	0.067	0.055	0.046
30	0.070	0.055	0.053	0.054	0.054	0.047
50	0.061	0.053	0.049	0.045	0.051	0.045
100	0.053	0.048	0.053	0.039	0.049	0.051