The Similarity Index and DNA Fingerprinting[1]

Michael Lynch
Department of Biology, University of Oregon

DNA-fingerprint similarity is being used increasingly to make inferences about levels of genetic variation within and between natural populations. It is shown that the similarity index—the average fraction of shared restriction fragments—provides upwardly biased estimates of population homozygosity but nearly unbiased estimates of the average identity-in-state for random pairs of individuals. A method is suggested for partitioning the DNA-fingerprint dissimilarity into within- and between-population components. Some simple expressions are given for the sampling variances of these estimators.

Introduction

DNA fingerprinting (Jeffreys et al. 1985*b*, 1985*c*) has attracted considerable attention as a possible means for rapidly assessing levels of genetic variation in natural and domesticated populations. It is common for series of "fingerprint loci" to share a core sequence, in which case multiple restriction-fragment-length polymorphisms can be visualized simultaneously on the same gel. Because such loci tend to exhibit high allelic diversity, random members of outbred populations rarely have the same fingerprint profiles. This suggests that DNA-fingerprint similarity may provide a sensitive indicator of relative levels of population homozygosity.

Comparative surveys of DNA-fingerprint similarity are now being pursued in several laboratories. Special attention is being given to levels of variation in small populations of endangered species and to the discrimination of breeds of domesticated species. However, since the statistical methods for the analysis of DNA-fingerprint data have received little attention (Lynch 1988; Cohen 1990), there is some uncertainty as to the interpretation of parameter estimates. The present paper is concerned solely with the statistical issues associated with DNA-fingerprint similarity analysis and starts with the assumption that the data to be analyzed are unambiguous. This is not meant to trivialize the numerous aspects of gel running, reading, and interpretation which may sometimes rival the statistical problems (Lander 1989).

In the following discussion, several generous assumptions are made with respect to the technical capabilities of the investigator. First, it is assumed that the DNA of individuals being compared is run in nearby lanes and/or with adequate controls, to minimize the errors in assigning identity to fragment pairs. Second, all individuals are assumed to be random members of the population. Third, it is assumed that any comigration of nonallelic markers can be resolved either by differences in band intensity or from other information. Fourth, the marker loci are assumed to be unlinked and in Hardy-Weinberg equilibrium within and between loci. Fifth, it is assumed that the same set of homologous loci is assayed completely for all individuals.

1. Key words: DNA fingerprinting, population genetic analysis.

Address for correspondence and reprints: Michael Lynch, Department of Biology, University of Oregon, Eugene, Oregon 97403.

Mol. Biol. Evol. 7(5):478–484. 1990.

0737-4038/90/0705-0008$02.00

The Meaning of DNA-Fingerprint Similarity

DNA-fingerprint similarity is generally defined as the fraction of shared bands. For individuals x and y, it is the number of common fragments in their fingerprint profiles (n_{xy}) divided by the average number of fragments exhibited by both individuals,

$$S_{xy} = \frac{2n_{xy}}{n_x + n_y}. \tag{1}$$

It would be useful if this index could be related to some standard population genetic parameter.

The parameters that would seem to be of greatest interest to those performing surveys of DNA-fingerprint similarity are the identity-in-state between pairs of individuals and the population homozygosity. Identity-in-state for two individuals can be defined by letting AA—AA and Aa—Aa comparisons indicate 100% identity and by letting AA—Aa and Aa—Aa′ comparisons indicate 50% identity. The expected genotypic identity-in-state for a random mating population is

$$E(I) = \frac{\sum_{k,i} p_{ki}^2 + p_{ki}^2(1 - p_{ki})^2}{L}, \tag{2}$$

where p_{ki} is the frequency of the ith allele at the kth locus and where L is the number of loci. Alternatively, identity-in-state can be defined from the standpoint of random gametes drawn from the two individuals under comparison. Under random mating, the expected gametic identity-in-state is equivalent to the population homozygosity,

$$E(H) = \frac{\sum_{k,i} p_{ki}^2}{L}. \tag{3}$$

Equations (2) and (3) show that $E(I) > E(H)$ and that the difference is greatest when there are a few alleles per locus at intermediate frequencies.

When it is noted from Jeffreys et al. (1985*a*) and Lynch (1988) that

$$E(S) = \frac{\sum_{k,i} p_{ki}^2(2 - p_{ki})}{L}, \tag{4}$$

the preceding formulas can be rearranged to

$$E(I) = E(S) + \frac{\sum_{k,i} p_{ki}^3(p_{ki} - 1)}{L} \tag{5}$$

and

$$E(H) = E(S) + \frac{\sum_{k,i} p_{ki}^2(p_{ki} - 1)}{L}. \tag{6}$$

Thus, the similarity index is always an upwardly biased estimator of both I and H, more so for the latter, and the magnitude of this bias is greatest when most alleles are at intermediate frequencies (and consequently when S is intermediate). The maximum bias occurs when $p = .5$ for all alleles, in which case $E(S) - E(I) = .125$ and $E(S) - E(H) = .25$. When all allele frequencies are low, the expected similarity is approximately twice the homozygosity.

To provide a more empirical evaluation of the relationship between similarity, identity-in-state, and homozygosity, several imaginary populations were examined, each consisting of 10 loci each having 1–10 alleles. The gene-frequency distributions employed were similar to those estimated for natural populations. The expected identity-in-states, homozygosities, and similarities were computed with equations (2)–(4). Figure 1 shows that, as expected, $\bar{S}$ usually gives substantially upwardly biased estimates of the population homozygosity. On the other hand, $\bar{S}$ overestimates the average identity-in-state only slightly.

The Sampling Variance of Fingerprint Similarity

Through the use of two or three probes with nonoverlapping sequence homology, it should not be difficult to sample 30–40 loci by DNA fingerprinting. This raises a useful statistical property. When large numbers of polymorphic loci are sampled, the distribution of similarity is expected to be approximately normal by the central limit theorem. The standard errors can then be used to construct confidence limits and for other applications associated with hypothesis testing.

The sampling variance for the mean population similarity can be estimated directly from the observational data,

$$\mathrm{Var}(\bar{S}) = \frac{N\,\mathrm{Var}(S_{xy}) + 2N'\mathrm{Cov}(S_{xy}, S_{xz})}{N^2}, \qquad (7)$$

where N is the total number of similarity measures used to estimate $\bar{S}$ and where N' is the number of pairs of those measures that share an individual. For example, if all

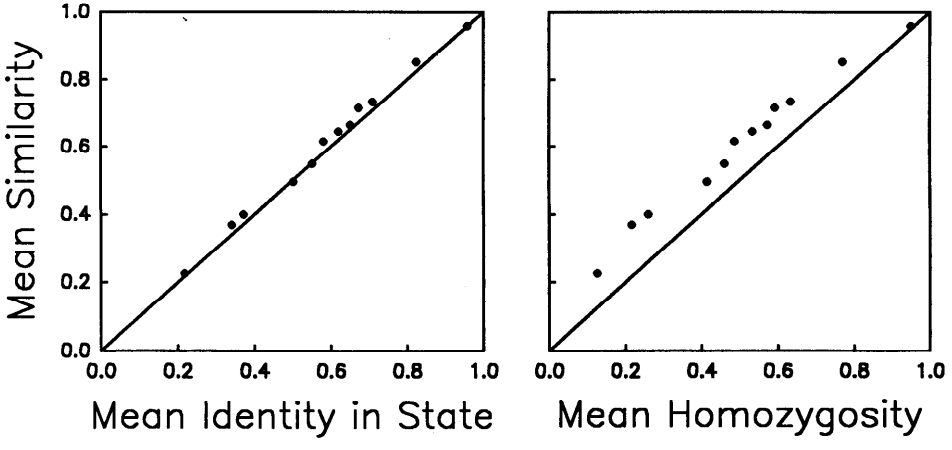

FIG. 1.—Comparison of mean DNA-fingerprint similarity with the mean homozygosity and mean identity in state in 12 simulations.

possible comparisons between four individuals have been made, $N = 6$ and $N' = 12$. The standard error of $\bar{S}$ is estimated by the square root of this quantity.

The sampling variance of the S_{xy} can be estimated with

$$\mathrm{Var}(S_{xy}) = \frac{N^*(\overline{S_{xy}^2} - \bar{S}^2)}{N^* - 1}, \tag{8}$$

where the mean square $\overline{S_{xy}^2}$ is computed from a set of N^* pairwise comparisons that do not share a member (i.e., if S_{wx} and S_{yz} are included, S_{wy} should not be). A simple way to compute the mean square is to use nonoverlapping pairs of individuals on gels (i.e., lane 1 vs. lane 2, lane 3 vs. lane 4, etc.), such that N^* equals one-half the number of individuals assayed.

The second term in equation (7) arises because there is a positive correlation between similarity measures that involve a common member. If the shared member happens to exhibit several bands that are common in the population, it will tend to have high similarities with all other members of the population—and vice versa if it happens to contain rare alleles. Failure to account for this will lead to underestimates of the sampling variance of $\bar{S}$ when multiple comparisons are made with the same individuals. This sampling covariance can also be estimated directly from the data,

$$\mathrm{Cov}(S_{xy}, S_{xz}) = \frac{N^*(\overline{S_{xy}S_{xz}} - \bar{S}^2)}{N^* - 1}, \tag{9}$$

where N^* is now the number of pairs of comparisons involving shared members. The mean cross product can be computed most efficiently by focusing on adjacent triplets on gels (i.e., lanes 1–3 yield $S_{12}S_{23}$, lanes 4–6 yield $S_{45}S_{56}$, etc.).

Strictly speaking, the above formulations estimate the sampling variance and covariance associated with the loci that happened to be included in the fingerprint survey. They do not account for the error arising from the sampling of a finite number of loci. An alternative approach is to assume that the sampled loci have gene-frequency distributions that are representative of other such loci throughout the genome. Since the similarity index estimates the probability that two random individuals share any fragment, it is reasonable to expect the sampling variance of S_{xy} to be approximated by the binomial sampling variance estimator $\bar{S}(1 - \bar{S})/2L$, where $2L$ is the number of genes sampled per individual.

To use this expression in practical applications, an estimate of L is required. A direct estimate of the number of loci is difficult to obtain unless a detailed segregation analysis can be carried out, which is usually not the case. However, the similarity index provides some information on L. If $\bar{n}$ is the average number of bands exhibited by an individual, then $\bar{n} \simeq \bar{S}L + 2(1 - \bar{S})L$. Rearranging, and recalling that $\bar{S}$ overestimates the homozygosity, we obtain $L \leq \bar{n}/(2 - \bar{S})$. On the other hand, the number of loci must be $> \bar{n}/2$. Averaging these two values, we obtain $L \simeq \bar{n}(4 - \bar{S})/[4(2 - \bar{S})]$. When observed values are substituted for expected values, an approximate estimator for the sampling variance of S_{xy} for a random pair of individuals is then

$$\mathrm{Var}'(S_{xy}) = \frac{2\bar{S}(1 - \bar{S})(2 - \bar{S})}{\bar{n}(4 - \bar{S})}. \tag{10}$$

This expression is very similar to a formula derived by Nei and Tajima (1983) for the case in which the number of loci is known.

Equation (10) should be used in place of equation (8) when the mean similarities of different populations are being compared and when it is uncertain whether the same loci have been sampled in both populations. It should also be used when one is using the set of sampled loci to make inferences about genome-wide properties. Since the covariance between similarity measures, $\mathrm{Cov}(S_{xy}, S_{xz})$, is proportional to the sampling variance, equation (7) can be corrected for locus sampling by multiplying by $\mathrm{Var}'(S_{xy})/\mathrm{Var}(S_{xy})$.

To verify the utility of equation (10), the sets of loci described above were sampled with replacement so that the estimated similarities of pairs of individuals within each population were based on different sets of loci. $\mathrm{Var}'(S_{xy})$ was computed for 2,000 pairs of individuals and was averaged to obtain a population-wide mean. Figure 2 shows that the variance computed by equation (1) approximates the true variance (computed directly from the simulated data by using the usual variance definition) reasonably well. The estimated variances tend to overestimate the actual variances slightly, so equation (10) yields conservative standard errors.

Population Subdivision

In some situations, it may be of interest to evaluate whether there is significantly less similarity between samples than would be expected by chance, a result which would indicate population subdivision. However, a measure of between-population similarity corrected by the within-population similarity is

$$\bar{S}_{ij} = 1 + \bar{S}'_{ij} - \frac{\bar{S}_i + \bar{S}_j}{2}, \qquad (11)$$

where $\bar{S}_i$ is the average similarity of individuals within population i and where $\bar{S}'_{ij}$ is

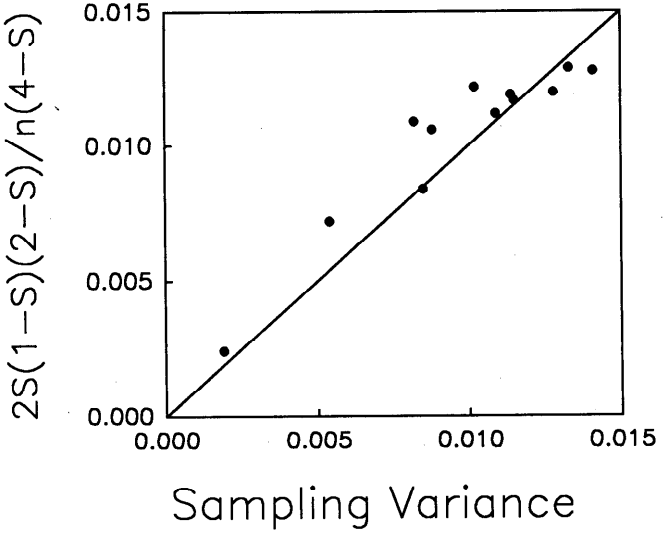

FIG. 2.—Comparison of actual sampling variance of S_{xy} (over individuals and loci) with that approximated by eq. (10).

the average similarity between random pairs of individuals across populations i and j. Some readers may be more comfortable using $\bar{D}_{ij} = 1 - \bar{S}_{ij}$ as an index of dissimilarity. Note that when $\bar{S}'_{ij} = \bar{S}_i = \bar{S}_j$, $\bar{D}_{ij} = 0$, indicating that the populations are homogeneous. A simple way of obtaining the information necessary to solve equation (11) is to run alternating pairs of individuals from populations i and j across the gel. By this means, each individual (except those on the ends of gels) can be compared with a member of its own population and with a member of the alternate population.

The sampling variance of $\bar{S}_{ij}$ (and $\bar{D}_{ij}$) is given by

$$\mathrm{Var}(\bar{S}_{ij}) = \mathrm{Var}(\bar{S}'_{ij}) + \tfrac{1}{4}[\mathrm{Var}(\bar{S}_i) + \mathrm{Var}(\bar{S}_j)] - \mathrm{Cov}(\bar{S}'_{ij}, \bar{S}_i) - \mathrm{Cov}(\bar{S}'_{ij}, \bar{S}_j)\,. \quad (12)$$

Expressions for the sampling variance of $\bar{S}_i$ and $\bar{S}_j$ have been given above, and the sampling variance of $\bar{S}'_{ij}$ can be estimated by use of the same formulas, with individual x being from one population and individual y being from the other. The sampling covariances between the uncorrected between-population similarity and the within-population similarities are nonzero whenever the same individuals are used in the computation of each parameter estimate. These covariances can again be estimated from the data directly:

$$\mathrm{Cov}(\bar{S}'_{ij}, \bar{S}_i) = \frac{N^*\mathrm{Cov}(S_{xi,yj}, S_{xi,zi})}{N_{ij}N_i}\,, \quad (13)$$

where N_i is the number of similarity indices computed for population i and where N_{ij} is the number computed from cross-population comparisons. N^* is the number of combinations of within- and between-population comparisons that share an individual from population i, and the covariance term is computed from those combinations of indices.

Since the similarity index does not yield unbiased estimates of population homozygosity, some care needs to be taken in using it in the estimation of the usual measure of population subdivision: Wright's (1951) F statistics. However, if $\bar{S}'_{ij}$, $\bar{S}_i$, and $\bar{S}_j$ are all biased to approximately the same degree, then these biases will cancel out in equation (11), leaving $\bar{D}_{ij}$ as a nearly unbiased estimator of the between-population gene diversity (heterozygosity). If D_b is the average value of $\bar{D}_{ij}$ over all i, j and if D_w is the average value of $1 - \bar{S}_i$ over all i, then

$$F' = \frac{D_b}{D_w + D_b} \quad (14)$$

should provide a downwardly biased—and hence conservative—estimate of population subdivision. F' takes on a value of 1 when populations are fixed for different alleles and takes on a value of 0 when there is no subdivision. A standard error for F' can be obtained by use of a Taylor expansion approximation that takes into account the sampling variance-covariance structure of D_b and D_w (Chakraborty 1974; Lynch and Crease 1990).

Discussion

The main point of the present paper has been to put the DNA-fingerprint similarity index (S) in the context of population genetic parameters and to provide approximate expressions for the sampling variance of S in terms of observable quantities. The traditional measure of population uniformity—and the one that fits most naturally into most population genetic formulations—is the mean homozygosity. Unfortunately, the similarity index does not provide a good estimate of this quantity. Rather, it closely approximates the average identity-in-state for random pairs of individuals. A comparison of equations (2) and (3) shows that there is no simple relationship between the average identity-in-state and the population homozygosity. However, the two parameters do tend to be highly correlated. Thus, when this distinction is kept in mind, the similarity index may yield adequate information for some practical applications.

As noted above, provided that large numbers of polymorphic loci are examined, it seems reasonable in hypothesis testing to treat S_{xy} as a normally distributed variable with approximate mean $\bar{S}$ and approximate variance $\mathrm{Var}(S_{xy})$. It is then possible to use the standard errors to identify populations that are exceptionally depauperate in genetic variation, under the assumption that the variation exhibited by fingerprinting loci is proportional to that in the remainder of the genome. When $\bar{S}_{ij}$ is treated as a normally distributed variable, it is also possible to test the null hypothesis of no population subdivision.

Acknowledgments

This work has been supported by NSF grant BSR 86-00487 and PHS grant R01 GM36827-01. I give many thanks to two anonymous reviewers for helpful comments.

LITERATURE CITED

CHAKRABORTY, R. 1974. A note on Nei's measure of gene diversity in a substructured population. Humangenetik **21**:85–88.

COHEN, J. E. 1990. DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. Am. J. Hum. Genet. **46**:358–368.

JEFFREYS, A. J., J. F. Y. BROOKFIELD, and R. SEMEONOFF. 1985*a*. Positive identification of an immigration test-case using human DNA fingerprints. Nature **317**:818–819.

JEFFREYS, A. J., V. WILSON, and S. L. THEIN. 1985*b*. Hypervariable 'minisatellite' regions in human DNA. Nature **314**:67–73.

———. 1985*c*. Individual-specific 'fingerprints' of human DNA. Nature **316**:76–79.

LANDER, E. S. 1989. DNA fingerprinting on trial. Nature **339**:501–505.

LYNCH, M. 1988. Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol. **5**:584–599.

LYNCH, M., and T. J. CREASE. 1990. The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. **7**:000–000.

NEI, M., and F. TAJIMA. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105**:207–217.

WRIGHT, S. 1951. The genetical structure of populations. Ann. Eugenics **15**:323–354.

WALTER M. FITCH, reviewing editor

Received February 22, 1990; revision received April 27, 1990

Accepted April 30, 1990