ESTIMATING GENETIC VARIABILITY WITH RESTRICTION ENDONUCLEASES

RICHARD R. HUDSON[1]

Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Manuscript received September 8, 1981
Revised copy accepted January 6, 1982

ABSTRACT

The estimation of the amount of sequence variation in samples of homologous DNA segments is considered. The data are assumed to have been obtained by restriction endonuclease digestion of the segments, from which the numbers and frequencies of the cleavage sites in the sample are determined. An estimator, $\hat{p}$, of the proportion of sites that are polymorphic in the sample is derived without assuming any particular population genetic model for the evolution of the population. The estimator is very close to the EWENS, SPIELMAN and HARRIS (1981) estimator that was derived with the symmetric WRIGHT-FISHER neutral model. ENGELS (1981) has also recently proposed an estimator of the same quantity, and he arrived at his estimator without assuming a particular population genetic model. The sampling variance of $\hat{p}$ and ENGELS' estimator are derived. It is found that the sampling variance of $\hat{p}$ is lower than the sampling variance of ENGELS' estimator. Also, the sampling variance of $\hat{\theta}$, an estimate of θ $(=4Nu)$ is obtained for the symmetric WRIGHT-FISHER neutral model with free recombination and with no recombination.

A restriction endonuclease cuts DNA segments wherever the enzyme's recognition sequence occurs. When a number of homologous DNA segments are each treated with a particular restriction endonuclease, sequence heterogeneity is revealed if it is found that not all the segments are cut in the same locations. Quantitative estimates of the amount of sequence variation can be obtained with data derived from restriction endonuclease digestion of homologous DNA segments. EWENS, SPIELMAN and HARRIS (1981) and ENGELS (1981) have each proposed an estimator of p, the proportion of nucleotide sites which are polymorphic in a sample of segments examined with restriction endonucleases. The EWENS estimator, $\hat{p}_w$, was based on an analysis of the WRIGHT-FISHER neutral model with symmetric mutation. For a detailed description and analysis of the WRIGHT-FISHER model, see EWENS 1979. $\hat{p}_w$ differs from earlier intuitive estimates (e.g. JEFFREYS 1979) by a factor of two. The applicability of $\hat{p}_w$ to other models, neutral or otherwise, was unclear from EWENS' derivation. ENGELS derived his estimator, $\hat{p}_g$, with just one simple assumption (which will be stated in the next section); no other details of the population structure were needed. Engels suggested that $\hat{p}_g$ was to be preferred to $\hat{p}_w$ because it did not depend on

[1] Present address: Department of Genetics, University of California, Davis, California 95616.

detailed assumptions concerning the population from which the sample was taken.

In this note, I present a simple derivation of an estimator, $\hat{p}$, which is very close to $\hat{p}_w$. The derivation of $\hat{p}$ is not based on a particular model for the evolution of the population. With this derivation it is more easily seen under which conditions $\hat{p}_w$ and $\hat{p}$ apply to other models; $\hat{p}_w$ and $\hat{p}$ apply to a more restricted class of symmetric neutral models than does $\hat{p}_g$.

It is shown that the sampling variance of $\hat{p}$ is less than the sampling variance of $\hat{p}_g$. Estimators of the sampling variance of $\hat{p}$ and $\hat{p}_g$ are also obtained.

EWENS showed that the same data used to estimate p could be used, under the symmetric WRIGHT-FISHER neutral model, to estimate $\theta = 4Nu$, where N is the population size and u is the neutral mutation rate. The sampling variance of the EWENS estimator of θ is obtained under the no-recombination neutral model and the free-recombination neutral model.

DEFINITIONS AND ASSUMPTIONS

Consider a set of n homologous segments, each L nucleotides long. The data to be analysed are assumed to have been obtained by treating each segment with a restriction enzyme and determining the points at which each segment is cut. A block is defined as j consecutive nucleotide positions, where j is the length of the recognition sequence of the restriction enzyme being used. Usually j is 4 or 6. A cleavage site is defined as a block for which at least one DNA segment of the set has the recognition sequence and is thus cut. A block or cleavage site is referred to as "monomorphic" if all the DNA segments in the set are identical at the nucleotide positions in this block or cleavage site. Otherwise, the block or cleavage site is "polymorphic." Let m be the number of cleavage sites found, and k be the number of the cleavage sites which are polymorphic (that is the number of sites at which some, but not all, of the segments were cleaved.) Recall that p is the proportion of all nucleotide sites that are polymorphic in the set. Estimates of p are the central interest of his note.

The ENGELS estimator of p is

$$\hat{p}_g = [c-n(m-k)]/jc = (\sum_{i=1}^{k} c_i)/(j\{n(m-k) + \sum_{i=1}^{k} c_i\}) \quad , \qquad (1)$$

where c is the total number of cuts at all cleavage sites, and c_i is the number of cuts at the i^{th} polymorphic cleavage site. The simple assumption upon which $\hat{p}_g$ depends is: The probability that the sequence at a random block on any particular segment is the recognition sequence, is the same, whether or not the block is known to be monomorphic. Note that ENGELS' assumption may not hold exactly, under neutral models, when mutation is assymetric. With assymetric mutation, knowledge that a block is monomorphic may change the probability that a specified sequence occurs on a particular segment at that block.

The EWENS estimator of p is $\hat{p}_w = k/2mj$. (The "intuitive" estimator of p is k/mj.) The estimator $\hat{p}$ is defined as

$$\hat{p} = k/(2m-k)j \ . \qquad (2)$$

When $n = 2$, $\hat{p}$ equals $\hat{p}_g$. When $k \ll m$, $\hat{p}$ and $\hat{p}_w$ are nearly equal. The difference between $\hat{p}$ and $\hat{p}_w$ is of the order of terms which were ignored in EWENS' derivation of $\hat{p}_w$. In the next section, I will show that $\hat{p}$ can be easily derived under the assumption that a randomly picked block which is polymorphic in the sample is twice as likely to be a cleavage site as a randomly picked block which is monomorphic in the sample. At stationarity, under neutral models with symmetric mutation, dimorphic blocks are twice as likely to be cleavage sites as monomorphic blocks. Thus, under any symmetric neutral model where all polymorphic blocks are in fact dimorphic blocks, the above assumption will hold. Under the symmetric WRIGHT-FISHER neutral model with θ small, most polymorphic blocks will be dimorphic. We now see why the intuitive estimator is biased by a factor approximately equal to two. Since cleavage sites are twice as likely at dimorphic blocks as at monomorphic blocks, the proportion (k/m) of cleavage sites that are polymorphic is expected to be approximately twice the proportion of all blocks that are polymorphic. This result is obtained explicitly in the next section.

Even with neutral models in which polymorphic blocks are always dimorphic, the assumption that polymorphic blocks are twice as likely to be cleavage sites as monomorphic blocks, will not generally hold exactly when mutation is assymetric (just as ENGELS' assumption does not hold.) With symmetric neutral models, $\hat{p}$ will be seriously biased, if it frequently occurs that more than two sequences are present in the sample at polymorphic blocks. This is not true of $\hat{p}_g$, which does not require that polymorphic blocks are dimorphic.

DERIVING $\hat{p}$

I will first derive an estimate of q, the proportion of blocks (rather than nucleotide sites) that are polymorphic in the sample. Consider the conditional probability that a particular block is polymorphic given that the block is a cleavage site. This probability satisfies the following identity:

$$\text{Prob(polymorphism}|\text{cleavage site)} = \frac{\text{Prob(polymorphism)Prob(cleavage site}\,|\,\text{polymorphism)}}{\text{Prob(cleavage site)}} \quad (3)$$

where Prob(polymorphism) is the probability that a randomly chosen block is polymorphic in the sample, which is just q. Prob(cleavage site|polymorphism) is the probability that a randomly chosen polymorphic block is a cleavage site; this probability is by our assumption twice the probability (denoted r) that a monomorphic block is a cleavage site. The denominator, Prob(cleavage site), is the probability that a randomly chosen block is a cleavage site. This probability can be written as the sum of the probabilities of two mutually exclusive events, namely the event that the block is a polymorphic cleavage site and the event that the block is a monomorphic cleavage site. Thus it follows that

$$P(\text{cleavage site}) = 2rq + (1-q)r \ . \quad (4)$$

Substituting into (3), one obtains

$$\text{Prob}(\text{polymorphism}\,|\,\text{cleavage site}) = 2rq/[2rq+(1-q)r] \ , \tag{5}$$

which, when q is small, is nearly twice the unconditional probability of polymorphism, q. Solving (5) for q and estimating Prob(polymorphism|cleavage site) by k/m, the observed frequency of polymorphism at cleavage sites, the following estimator of q is obtained:

$$\hat{q} = k/(2m-k) \ . \tag{6}$$

Following EWENS and ENGELS I assume that a given block may be polymorphic at no more than one its j positions. Thus I estimate p by $\hat{q}/j$, and obtain the estimator

$$\hat{p} = k/(2m-k)j \ . \tag{7}$$

One can also obtain $\hat{p}$ from ENGELS' estimator if one assumes that $E(c_i)$ equals $n/2$. Substituting $n/2$ for c_i in (1) leads to $\hat{p}$. For neutral models with symmetric mutation and for which polymorphic blocks are always dimorphic, $E(c_i)$ equals $n/2$.

SAMPLING PROPERTIES OF $\hat{p}$ AND $\hat{p}_g$

As just mentioned, $\hat{p}$ is just $\hat{p}_g$ with the c_i (which are variable quantities) replaced by $n/2$ (which is constant.) Clearly the sampling variance of $\hat{p}$ is less than the sampling variance of $\hat{p}_g$. It should be emphasized that $\hat{p}$ applies to a more restricted class of symmetric neutral models than does $\hat{p}_g$. When polymorphic blocks are frequently not dimorphic, $\hat{p}$ could be seriously biased. In this section, explicit expressions for the sampling variance of $\hat{p}$ and $\hat{p}_g$ are obtained. Estimators of these sampling variances are also obtained. Throughout this section it is assumed that polymorphic blocks are twice as likely to be cleavage sites as monomorphic blocks. Also it is assumed that $E(c_i)$ equals $n/2$.

Sampling variances depend on how one conceives of performing hypothetical repetitions of the experiment. Each repetition of our experiment could consist of sampling n homologous segments and applying a different restriction enzyme. The sampling variances of $\hat{p}$ and $\hat{p}_g$ using this repetition scheme reflect the precision with which they estimate the probability, p^*, that a random site is polymorphic in a random sample of n segments. Note that p, the proportion of sites which are polymorphic in a sample, would vary from sample to sample. The expectation of p is p^*.

Alternatively, each repetition of the experiment could consist of applying a different restriction enzyme to the same set of n segments. The sampling variances of $\hat{p}$ and $\hat{p}_g$, with this repetition scheme, reflects the precision with which $\hat{p}$ and $\hat{p}_g$ estimate p, the unknown proportion of sites which are polymorphic in the particular set of n segments under study. It is this latter type of repetition scheme which is considered in detail in this section.

Consider a set of n homologous segments with a fixed but unknown value of p. Each segment in this set consists of $L - j + 1$, or approximately, L blocks. These blocks overlap with neighboring blocks, so that the presence of the recognition sequence at one block is not independent of the presence or absence of the recognition sequence at neighboring blocks. I will assume that this dependence has negligible effect on the random variables k and $m - k$, since $m << L$, so that we may assume that we are dealing with approximately L independent blocks. I will again assume that polymorphic blocks are polymorphic at only one of the j sites, so that there are jpL polymorphic blocks and $(1-jp)\ L$ monomorphic blocks. With these assumptions the number of polymorphic cleavage sites (k) and the number of monomorphic cleavage sites $(m-k)$ are independent binomially distributed random variables. Let r be the probability that the sequence at a monomorphic block is the recognition sequence. By assumption, the probability that the recognition sequence occurs in the sample at a random polymorphic block is $2r$. Since r is fairly small, if jpL, the number of polymorphic blocks, is not too small, the binomial distributions of k and $m - k$ will be well approximated by POISSON distributions:

$$\mathrm{Prob}(k=i) = \left[{jpL \atop i}\right] (2r)^i(1-2r)^{jpL-i} \simeq \exp(-2rjpL)\,(2rjpL)^i/i! \qquad (8)$$

and

$$\begin{aligned}\mathrm{Prob}(m-k=l) &= \left[{(1-jp)L \atop l}\right] r^l(1-r)^{(1-jp)L-l} \\ &\simeq \exp\{-(1-jp)rL\}\{(1-jp)rL\}^l/l! \qquad (9)\end{aligned}$$

Using the POISSON distributions, one can calculate maximum likelihood estimates of the pair of parameters, rL and p, in terms of the observations, m and k. The maximum likelihood estimate of p obtained in this way is just $\hat{p}$.

The approximate mean and variance of $\hat{p}$ and $\hat{p}_g$ can be found using TAYLOR expansions and ignoring third ad higher order moments. Using this method and the POISSON distributions of (8) and (9), it was found that each estimate has a bias of order p^2. The variance of $\hat{p}$ was found to be

$$\mathrm{Var}(\hat{p}) \simeq p/2rLj\ , \qquad (10)$$

which can be estimated from the data by

$$\widehat{\mathrm{Var}}(\hat{p}) = \hat{p}^2/k\ . \qquad (11)$$

ENGELS obtained this result (his equation 18), when he calculated $\mathrm{V\dot{a}r}(\hat{p}_g)$ for $n = 2$ (because when $n = 2$, $\hat{p}_g = \hat{p}$). For $n > 2$, ENGELS did not calculate the sampling variance of $\hat{p}_g$. I do so now.

With p fixed, $m - k$ and $\sum_{i=1}^{K} c_i$ are independent, so the variance of $\hat{p}_g$ can be found approximately with

$$\mathrm{Var}(\hat{p}_g) \simeq \left[\frac{\partial \hat{p}_g}{\partial(m-k)}\right]^2 \mathrm{Var}(m-k) + \left[\frac{\partial \hat{p}_g}{\partial \Sigma c_i}\right]^2 \mathrm{Var}(\Sigma c_i)\ . \qquad (12)$$

If, in addition, the c_i are assumed mutually independent, then the $\mathrm{Var}(\Sigma c_i)$ can be written

$$\mathrm{Var}(\Sigma c_i) = \{E(c_i)\}^2\mathrm{Var}(k) + E(k)\mathrm{Var}(c_i) \quad . \tag{13}$$

By assumption, the expectation of c_i is $n/2$. Taking derivatives of $\hat{p}_g$ and using (13), one finds

$$\mathrm{Var}(\hat{p}_g) \simeq \frac{p}{2rL}\left\{1 + \frac{4}{n^2}\mathrm{Var}(c_i)\right\} \ , \tag{14}$$

ignoring terms in p^2. The variance of the c_i depends on the mechanism maintaining the polymorphisms. $\mathrm{Var}(c_i)$ could be estimated by its observed variance, $\sum_{i=1}^{k}(c_i - \bar{c})^2/(k-1)$, where $\bar{c} = \Sigma\, c_i/k$. It is clear that $\mathrm{Var}(c_i)$ is less than $n^2/4$, so the maximum variance of $\hat{p}_g$ is twice the variance of $\hat{p}$.

ESTIMATING θ

The analysis, so far, has concerned estimates of p, a property of a particular set of segments. One may also, as suggested by EWENS, SPIELMAN and HARRIS (1981), use the data to estimate θ, a population parameter. Under the symmetric WRIGHT-FISHER neutral model, EWENS proposed estimating θ with $\hat{\hat{\theta}} = \hat{p}_w/\log(n)$. In what follows the nearly identical estimator $\hat{\theta} = \hat{p}/\log(n)$, will be considered. The sampling variance of this estimator will be derived under the following scheme of hypothetical repetitions. Each repetition would consist of the application of the same restriction enzyme to a new sample of n homologous segments obtained from a completely independent population with the identical parameter θ. The populations will be assumed to be at stationarity with respect to allele frequencies. With this repetition scheme, p is a random variable. The distribution of p, under the neutral model, depends on θ and the amount of recombination. Note that with p a random variable, k and $m - k$ are no longer independent, nor, necessarily POISSON distributed.

For the two extreme cases of free recombination and no recombination, the distribution of p is easily described. In either case, the expected value of pL, the number of polymorphic nucleotide positions, is approximately $L\theta\log(n)$. With free recombination, pL is approximately POISSON distributed. With no recombination, WATTERSON (1975) gives the distribution of the number of segregating sites for the infinite-site model. For θ small, his result should apply approximately to the distribution of pL. Thus, for either the free-recombination or no-recombination neutral model, the joint probability generating function of k and $m - k$ can easily be written down. The variance of $\hat{\theta}$ can be obtained directly with the identity from conditional probability

$$\mathrm{Var}(\hat{p}) = E[\mathrm{Var}(\hat{p}|p)] + \mathrm{Var}[E(\hat{p}|p)] \quad . \tag{15}$$

$\mathrm{Var}(\hat{p}|p)$ is given by equation (10). $E(\hat{p}|p)$ is approximately p. Thus, regardless of the amount of recombination,

$$\begin{aligned}\mathrm{Var}(\hat{p}) &\simeq E(p/2rLj) + \mathrm{Var}(p)\\ &\simeq \theta \log(n)/2rLj + \mathrm{Var}(p) \ . \end{aligned} \tag{16}$$

For the free recombination case, $\mathrm{Var}(p)$ is just $\theta\log(n)/L$, which (since r is small) is very small compared to the first term on the right hand side of (16). So, we have the approximation

$$\mathrm{Var}(\hat{p}) \simeq \theta \log(n)/2rLj \tag{17}$$

and

$$\mathrm{Var}(\hat{\theta}) \simeq \theta/[2rLj\log(n)] \ , \tag{18}$$

which can be estimated by

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \hat{\theta}^2/k \ . \tag{19}$$

The variance in p contributes insignificantly to the variance in $\hat{p}$ and $\hat{\theta}$.

For the no-recombination case, the variance of pL is (WATTERSON 1975):

$$\mathrm{Var}(pL) \simeq L\theta \log(n) + (L\theta)^2 \sum_{i=1}^{n-1} 1/i^2 \ . \tag{20}$$

Substituting into (16), one obtains

$$\begin{aligned}\mathrm{Var}(\hat{p}) &\simeq \theta \log(n)/2rLj + \theta \log(n)/L + \theta^2 \sum_{i=1}^{n-1} 1/i^2\\ &\simeq \theta \log(n)/2rLj + \theta^2 \sum_{i=1}^{n-1} 1/i^2 \ , \end{aligned} \tag{21}$$

and

$$\mathrm{Var}(\hat{\theta}) \simeq \frac{\theta}{2rLj\log(n)} + \frac{\theta^2}{\{\log(n)\}^2} \sum_{i=1}^{n-1} 1/i^2 \ , \tag{22}$$

which may be estimated by

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{\hat{\theta}}{k} + \frac{\hat{\theta}^2}{\{\log(n)\}^2} \sum_{i=1}^{n-1} 1/i^2 \ . \tag{23}$$

The variance of p, may contribute substantially to the variance of $\hat{p}$ and $\hat{\theta}$, when there is no recombination.

One can now calculate standard errors of the estimate, $\hat{\theta}$, with (19) and (23) for the case of free recombination and no recombination respectively.

APPLICATIONS

Up to this point, data derived from the use of a single enzyme have been considered. The results extend directly to data from several restriction enzymes provided all have recognition sequences of the same length. In this case, k is to be interpreted as the total number of polymorphic cleavage sites, and m is the total number of cleavage sites.

The data of BROWN (1980) will illustrate the use of the estimators. BROWN studied mtDNA from 21 individuals, with seven tetranucleotide restriction enzymes. The data as summarized by ENGELS are: $m = 244$, $k = 45$, and $c = 4672$. The estimate, $\hat{p}$, of the proportion of nucleotide sites that are polymorphic in this sample of DNA segments is $\frac{45}{4(488-45)} = 0.0254$, with standard error from (11) of 0.0038. (The ENGELS estimate, $\hat{p}_g$, is 0.0264. $\hat{p}_w$ is 0.0231.)

Estimating θ, we have $\hat{\theta} = p/\log(n) = 0.0083$; assuming no recombination, the standard error, from (23), is 0.037. As noted by ENGELS, for mitochondrial DNA, θ must be defined as Nu, rather than $4Nu$.

When data are obtained with restriction enzymes with different length recognition sequences, an estimate of the proportion of nucleotide positions that are polymorphic is still easily obtainable. Following ENGELS, p can be estimated with exactly the same expression as before, but j is redefined to be the weighted average of the lengths of the recognition sequences,

$$j = \frac{\Sigma i m_i}{\Sigma m_i} \quad , \tag{24}$$

where m_i is the number of cleavage sites in the sample that correspond to recognition sequences of length i. j must now be considered a random variable, but as discussed by ENGELS, the variance and covariances involving j are small, so results concerning variances which assumed j to be constant are still essentially correct.

I illustrate with the same example used by EWENS, SPIELMAN and HARRIS (1981) which they derived from the data of JEFFREYS (1979). One hundred and twenty homologous DNA segments were examined with eight restrictions enzymes, seven with recognition sequences of length 6, and one enzyme with recognition sequence of length 4. The number of cleavage sites with recognition sequence of length 4 was 7, that is, $m_4 = 7$, and none of these cleavage sites was polymorphic. There were 47 cleavage sites for the other enzymes, $m_6 = 47$, and 3 of these sites were polymorphic. Thus we have $j = [(4)(7) + (6)(47)]/(7+47) = 5.74$, and

$$\hat{p} = (0+3)/[5.74(14+94-3)] = 0.0051 \quad ,$$

with standard error of 0.0029 from (11).

Estimating θ, we have $\hat{\theta} = 0.0051/\log(120) = 0.0011$. The standard error depends on the amount of recombination. Assuming free recombination, (19) gives standard error of 0.0006. Assuming no recombination, (23) gives standard error of 0.0007, only slightly higher than the free recombination result.

CONCLUSIONS

The estimator, $\hat{p}$, can be derived with an assumption which holds under symmetric neutral models if it is also true that polymorphic blocks are always dimorphic. ENGELS' derivation of $\hat{p}_g$ does not require that polymorphic blocks be dimorphic. However, the sampling variance of $\hat{p}_g$ is larger than that of $\hat{p}$ (but less than twice the sampling variance of $\hat{p}$.) The sampling variance of $\hat{p}$ can be estimated with (11).

The estimate, $\hat{p}/\log(n)$, of the population parameter θ, has sampling variance which can be estimated by (19) and (23) for the free-recombination and no-recombination neutral models, respectively.

The helpful comments of WILLIAM ENGELS and WARREN EWENS are gratefully acknowledged. The author is supported by Public Health Service Grant T32 GM 07517-04.

LITERATURE CITED

BROWN, W. M., 1980 Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. U.S.A. **77**: 3605–3609.

ENGELS, W. R., 1981 Estimating genetic divergence and genetic variability with restriction endonucleases. Proc. Natl. Acad. Sci. U.S.A. **78**: 6329–63333.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3**: 87–112. ——, 1979 *Mathematical Population Genetics*. Springer Verlag, Berlin.

EWENS, W. J., SPIELMAN, R. S., and HARRIS, H., 1981 Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. U.S.A. **78**: 3748–3750.

JEFFREYS, A. J., 1979 DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ,- and β-globin genes of man. Cell. **18**: 1–10.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7**: 256–276.

Corresponding editor: W. J. EWENS