Volume 4 Number 5 1977 Nucleic Acids Research

Estimation of DNA sequence divergence from comparison of restriction endonuclease digests*

William B. Upholt

Departments of Pediatrics and Biochemistry and the Joseph P. Kennedy, Jr. Mental Retardation Research Center, The University of Chicago, Box 413, Chicago, IL 60637, USA

Received 10 January 1977

ABSTRACT

An estimation of the DNA sequence divergence between defined DNA sequences of individuals or species may be made from comparison by gel electrophoresis of restriction endonuclease digests. This analysis is applicable to purified DNA sequences of moderate complexity (1-100 x 10^6 daltons) which have diverged by base substitution of 0.5 to 25% of nucleotides.

INTRODUCTION

Estimations of genetic relationships between organisms have been derived from a number of different biochemical measurements. Most of these procedures analyse protein relatedness by comparing amino acid sequences[1], electrophoretic mobility[2] or immunological crossreactivity[3] of equivalent proteins from two individuals or species. Amino acid sequence comparisons have been particularly useful for distant relationships[4], immunological crossreactivity for moderately close relationships and electrophoretic comparisons for closely related species.

Amino acid sequence comparisons are the most precise of the above three methods but cannot measure changes in the third or variable position of codons which may or may not be of importance in control steps related either to secondary structure of mRNA or to availability of isoacceptor tRNA species. Recent data of Salser and Isaacson[5] comparing rabbit and human globin mRNA sequences indicate that the acceptance rate of base substitution at the third nucleotide position is approximately ten times higher than at the other two. This estimate is substantially higher than previous estimates which have been used to convert amino acid changes into nucleotide sequence changes.

Electrophoretic and immunological comparisons have proved useful in that they provide ready access to larger amounts of data compared with sequence analysis. However, electophoretic analysis is estimated to score only about one-quarter of all amino acid differences between proteins[2] and some immunological differences are due to membrane carbohydrates or post-translational modifications rather than actual changes in primary amino acid sequences[6]. None of these three procedures take into account changes which occur in control or nonprotein coding DNA sequences.

The ideal method of comparison is the direct determination of base sequence in nucleic acids[7]. However at present this is not practical for large numbers of determinations. A simpler process has been the comparison of the extent of nucleic acid hybridization and stability of these hybrids formed between different species[8, 9]. This sort of analysis has worked well for nucleic acids of low kinetic complexity, but previous results comparing moderately repetitive and unique DNAs have proved to be more difficult to interpret due to variations in kinetic complexity and sequence divergence in this rather heterogeneous class of nucleic acid sequences[10].

With the current growing availability of restriction endonucleases[11] and the availability of cloned DNA sequences[11], a new possibility has become available for the comparison of purified sequences of moderate complexity (1-100 x 10^6 daltons). Closely related DNA sequences will have in common a differing number of restriction enzyme cleavage sites depending on the degree of sequence divergence between the DNAs. Analysis of the fraction of cleavage sites conserved between two DNAs may be used to estimate sequence divergence either when the sites are mapped or the fragment changes are sufficiently simple to be interpreted in terms of specific site changes. In cases where the digest provides a large number of fragments and where more extensive changes have occurred, it is not practical to map or analyse changes of specific sites and the comparison may be based on the fraction of conserved fragments.

Analysis of restriction digests requires that the DNA-restriction enzyme pair chosen be such that the data provide both a sufficient number of fragments for meaningful comparison and a sufficiently simple fragment pattern so that identities of fragments may be determined. It must also be assumed that the fragment changes arise by substitutions of single base pairs rather than rearrangements, deletions or gross organizational changes and that the cleavage sequence occurs with a distribution and frequency close to that expected in a random sequence of the same base composition. These conditions are met in the comparison of sheep and goat mitochondrial DNAs by Upholt and Dawid[12] and should be satisfactory for comparisons of cloned genes from closely related species.

THEORY AND CALCULATIONS

A. Analysis of divergence from site changes

Let p = the probability of a nucleotide substitution at a single nucleotide site, then 1 - p is the probability of a single nucleotide site remaining unchanged. If n is the number of base pairs involved in a restriction cleavage site, then the probability that a given restriction site remains unchanged after a fraction, p, of the nucleotides have undergone substitution is:

$$S = (1 - p)^n \qquad (1a)$$

This is also equal to the fraction of total cleavage sites which are conserved. Solving for p, the equation becomes

$$p = 1 - (S)^{1/n} \qquad (1b)$$

which is plotted in figure 1 for restriction cleavage sites containing 4 and 6 base pairs.

These equations take into account only the actual observed or minimum number of changes. At higher levels of divergence, multiple changes at a single site will occur and the actual number of changes per nucleotide site, p_a, may be estimated using the Poisson distribution function

$$pa = - \frac{1}{n} \ln S \qquad (1c)$$

The use of the Poisson distribution assumes that all nucleotides have an equal probability of undergoing substitution. If the majority of the observed substitutions occur in a subclass of sites, such as the third position of codons, equation 1c will still provide an underestimate of the actual number of changes.

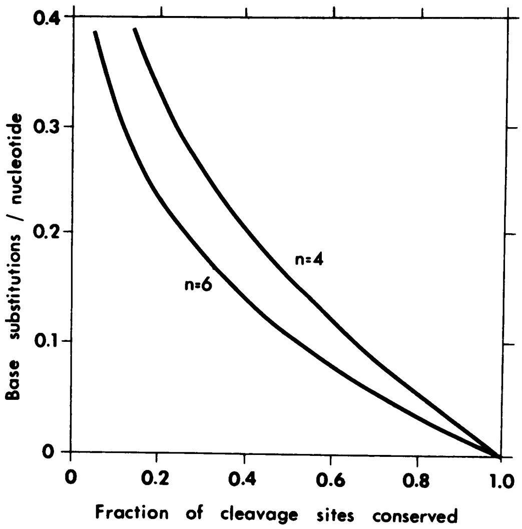

Figure 1. Estimated minimum number of base substitution per nucleotide as a function of fraction of cleavage sites conserved for cleavage sites containing 4 and 6 base pairs.

B. Analysis of sequence divergence from fragment changes

Two conditions must be met for the conservation of a restriction fragment: 1) the two existing external sites must remain unchanged and 2) a new site may not occur within the fragment.

The probability that one external cleavage site remains unchanged, S, after a fraction, p, of the bases have diverged was calculated in part A. The probability that both external sites remain unchanged is S^2.

The probability that new sites are not generated within an existing restriction fragment will first be considered in terms of the opposite possibility, i.e., the probability that a new site is generated at a potential n base site which is not initially a restriction site. Two events must occur in order for this to happen. First, one or more bases in this site must change and second, these changes must give rise to the correct sequence. The probability of one or more bases changing within an n base pair site is simply equal to one minus the probability that no base pairs change, or 1 - S. The probability of an n base site in a DNA of random sequence being the restriction site is:

$$a_{rs} = \left(\frac{X_{GC}}{2}\right)^{n_{GC}} \left(\frac{X_{AT}}{2}\right)^{n_{AT}} \qquad (2)$$

where n_{GC} and n_{AT} are respectively the numbers of GC and AT base pairs in the restriction site and X_{GC} and X_{AT} are the corresponding mole fractions of GC and AT base pairs in the DNA. The conditional probability of a site which is not initially the restriction site changing to the restriction site given that one or more base pairs have changed is:

$$a_c = \frac{a_{rs}}{1 - a_{rs}} \left\{ 2^n \left[\sum_{i=0}^{n} \binom{n}{i} \frac{\left(\frac{X_{AT}}{2}\right)^{n-i} \left(\frac{X_{GC}}{2}\right)^{i}}{1 - \left(\frac{X_{AT}}{2}\right)^{n-i} \left(\frac{X_{GC}}{2}\right)^{i}} \right] - \frac{a_{rs}}{1 - a_{rs}} \right\}$$

where $\binom{n}{i}$ is the number of combinations of n things taken i at a time and is equal to $\frac{n!}{i!(n-i)!}$. a_c has been evaluated for a number of extreme cases to determine the error in the estimated sequence divergence if a_c is assumed to be equal to a_{rs}. In a comparison of two DNAs of 90% G+C which have undergone 0.26 substitutions/nucleotide an error of 0.006 substitutions/nucleotide is made if the cleavage site consists of 4 GC base pairs. This error is considerably smaller than other errors in the determination. (For comparison of recognition sites of less than 4 base pairs, the assumption of a_c being equal to a_{rs} should be reevaluated.) In subsequent equations, a_{rs} will be used in place of a_c and the subscript rs will be dropped. The conditional probability of any n base site which is not the restriction site both undergoing a change, (1 - S), and changing to the restriction site is thus:

$$a(1 - S)$$

and the probability of such a change not occurring is:

$$1 - a(1 - S)$$

If i is the number of base pairs in the fragment under consideration, then

$$[1 - a(1 - S)]^{i-n+1}$$

is the probability that a new site does not occur within this fragment. The exponent, i-n+1, is equal to the number of possible n base sites in a fragment of i base pairs. The product of this probability and the probability that the two external site sequences are conserved, S^2, is the probability that the fragment is conserved:

$$S^2[1 - a(1 - S)]^{i-n+1}$$

The expected fraction of total fragments which are i base pairs in length is:

$$\frac{a(1 - a)^{i-1}}{\sum_{i=n}^{\infty} a(1 - a)^{i-1}}$$

The denominator is a normalising factor to correct for the exclusion of fragments smaller than the restriction site which are assumed not to cccur. Upon evaluation of the summation, this expression simplifies to:

$$a(1 - a)^{i-n+1} \qquad (3)$$

To obtain the fraction of total fragments which are conserved, F, the probability that a fragment of i base pairs is conserved is multiplied by the fraction of total fragments which are of this length and this product is summed for all possible fragment lengths:

$$F = \sum_{i=n}^{m} S^2a(1 - a)^{i-n+1}[1 - a(1 - S)]^{i-n+1} \qquad (4)$$

where m equal the maximum fragment length when the DNA contains no sites. When m is large this simplifies upon summation to

$$F = \frac{S^2}{2 - S - a + aS} \qquad (5)$$

Since a is small compared to 1 and S must be between 0 and 1, a and aS in the denominator may be set equal to zero and the equation becomes:

$$F \simeq \frac{S^2}{2 - S} \qquad (6a)$$

Substituting $(1-p)^n$ for S and solving for p, the fraction of base substitutions as a function of the fraction of conserved fragments, F, is:

$$p \simeq 1 - \left[\frac{-F + \sqrt{F^2 + 8F}}{2}\right]^{(1/n)} \qquad (6b)$$

This relationship is plotted in figure 2 for n equal to 4 and 6.

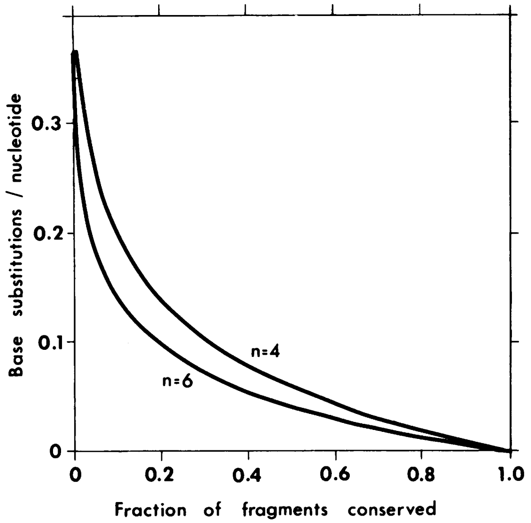

Figure 2. Estimated minimum number of base substitutions per nucleotide as a function of fraction of cleavage fragments conserved for cleavage sites containing 4 and 6 base pairs.

Standard deviations of the mean estimated sequence divergence may be evaluated by:

$$\sigma = \left[\frac{p(1 - p)}{N}\right]^{1/2} \qquad (7)$$

where p is the determined fraction of bases substituted and N is the number of independent nucleotide positions in cleavage sites (nucleotides in common cleavage sites may be counted only once).

DISCUSSION

The application of restriction digest gel patterns to the estimation of base substitution between related DNAs is limited by certain conditions. If 20% conservation of either sites or fragments is taken as an arbitrary lower limit below which assignment of fragment identity may be questionable, figures 1 and 2 show that the analysis of cleavage sites is valid only for DNAs which have undergone less than 25% substitution and the analysis of fragment conservation is valid only below about 15% base substitution. Furthermore, changes in sites must be assumed to arise from simple base substitutions rather than deletions, substitutions or other rearrangements of nucleotide sequence and derivation of the equation for comparison of fragments assumes that the number of fragments and the fragment size distribution are equivalent to that expected for a DNA composed of a random sequence of the same base composition.

Certain aspects of the restriction fragment gel patterns may be used to evaluate the suitability of specific DNAs for this type of analysis. Equation 2 may be used to predict the expected number of cutting sites (a_{rs} x number of base pairs in the DNA) and equation 3 provides the expected fragment size distribution. Quantitation of the number and sizes of the fragments unique to each DNA in a comparison can give some indication of whether these changes are the result of simple base substitutions or more complex changes. The loss of a single restriction site due to base substitution gives 3 fragment changes resulting in the loss of 2 fragments and the addition of a third fragment whose molecular weight is equal to the sum of the other two. Deletions within a restriction fragment will result in the loss of the fragment plus the addition of a smaller fragment whereas deletions containing a site will result in the loss of 2 fragments and the addition of a new fragment smaller than the sum of the lost fragments. An inversion of a segment including a site will result in the loss of 2 fragments and the addition of 2 new fragments of the same total molecular weight as the lost fragments.

Further verification of the nature of the sequence changes may be obtained by electron microscopic heteroduplex analysis as discussed by Upholt and Dawid[12]. Their analysis of changes in mitochondrial DNAs from 3 individual sheep and 2 goats using 3 different restriction endonucleases, EcoRI, Hind III and HaeIII, suggests that most of the observed changes in these DNAs are the result of single base substitutions. The number and size distribution of the cleavage sites they obtained agree well with the predicted distribution suggesting that the cleavage sites for these 3 enzymes are not part of special sequences which might occur at abnormal frequencies, e.g. termination codons, promotors, etc. The sequence divergences as measured by each of the three enzymes were also consistent.

ACKNOWLEDGMENTS

I thank I.B. Dawid and E. Westfall for helpful discussions.

REFERENCES

*Dedicated to Jerome Vinograd.

1 Fitch, W.M. (1976) J. Mol. Evol. 8, 13-40

2 King, M.-C. and Wilson, A.C. (1975) Science 188, 107-116

3 Prager, E.M. and Wilson, A.C. (1971) J. Biol. Chem. 246, 7010-7017

4 deHaën, C., Neurath, H. and Teller, D.C. (1975) J. Mol. Biol. 92, 225-259

5 Salser, W. and Isaacson, J.S. (1976) in Progress in Nucleic Acid Research Vol 19 pp. 205-220. Academic Press. New York

6 Nei, M. (1975) Molecular Population Genetics and Evolution pp. 145-146 North Holland. Amsterdam

7 Salser, H., Bowen, S., Browne, D., El Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B. and Whitcome, P. (1976) Federation Proceedings 35, 23-35

8 Laird, C.D., McConaughy, B.L. and McCarthy, B.J. (1969) Nature (London) 224, 149-154

9 Britten, R.J. and Davidson, E.H. (1976) Federation Proceedings 35, 2151-2157

10 McCarthy, B.J. and Farquhar, M.N. (1972) Brookhaven Symp. Biol. 23, 1-43

11 Nathans, D. and Smith, H.O. (1975) Ann. Rev. Biochem. 44, 273-293

12 Upholt, W.B. and Dawid, I.B. (1977) Cell, in press.