DNA Sequence Homology Estimation by Combinatorial Analysis of Endonuclease Restriction Data

(Accepted 1 August 1977)

SUMMARY

A combinatorial analysis of the products of DNA cleavage by restriction endonucleases permits an estimation of the sequence homology between closely related linear DNAs. The main assumption to be made is that fragments of identical length represent equivalent sections of the genomes. All patterns of common and varying restriction sites compatible with the results of the electrophoretic analysis of fragment sizes can be enumerated, each yielding an estimate of the proportion of sequences of restriction site length common to both DNAs. From this, sequence homology is derived by using a model of randomly distributed point mutations relating the two genomes. Possible limitations of the approach are discussed.

One of the many possible applications of restriction endonucleases (surveyed by Roberts, 1976) is the determination of biological relationship among various DNAs, as undertaken for poxviruses in the previous report (Müller et al. 1978). In this work, we define a theoretical basis for the quantitative assessment of these relationships. This is only possible for linear and closely related DNAs as some major assumptions must be made which are inadmissible otherwise.

As restriction enzymes have specific recognition sequences of, usually, four to six nucleotides, they can be said to analyse a selective sample of the full DNA sequence and the 'fragment patterns' obtained after electrophoretic separation of the cleavage products are characteristic of the original DNAs. Also, when one compares genomes of different relatedness, as in the previous paper (Müller et al. 1978), this is well reflected by the resulting fragment patterns (for a review of recent comparative analyses see Roberts, 1976). Closely related genomes yield almost identical patterns, as the enzyme picks its specific samples from two DNAs of high sequence homology. Thus, in this case, fragments of identical length can be assumed to represent equivalent sections of the corresponding genomes. Less closely related and entirely unrelated DNAs will perhaps also have a few fragments in common, especially if an enzyme yielding a large number of fragments is used. However, this will be due entirely to chance and by no means could the above assumption be applied.

If two individual fragments have the same size in two DNAs, we shall call them 'matching' or, as we are considering closely related DNAs, 'homologous' fragments. This does not mean that they share all their nucleotide sequences, but only that they represent equivalent sections of the two genomes. Under these circumstances, the difference between two related DNAs will be described by the way the two sets of non-matching (or non-homologous) fragments relate.

Two mechanisms can be used to describe how one set of non-matching fragments can be mapped on to the other: (a) some of the original restriction sites can be eliminated by alteration; (b) non-specific sequences can change into recognition sites. Recognition sequences can only be eliminated between adjacent non-homologous fragments, as such an
Short communications

Fig. 1. Possible distributions of the three non-homologous fragments $X_A$, $Y_A$ and $Z_A$ on the strand of DNA A. Case 1: all the fragments are separated by one or several homologous fragments; case 2: two non-homologous fragments are adjacent, away from the third one; case 3: the three fragments are connected.

elimination alters the lengths of both fragments thus rendering them non-homologous. New restriction sites, however, can be added anywhere within the non-homologous fragments. An essential condition is that only those patterns of eliminations and additions need be investigated after which no fragment remains invariant, as only the non-matching segments are considered. Of course, the number of possible eliminations and additions is subject to another requirement, as they must be compatible with the difference in the number of non-matching fragments indicated by electrophoresis results. Combinatorial analysis yields an enumerable set of possible elimination/addition patterns, corresponding to the relative positions of invariant and non-matching fragments on the genomes. Each of these possible patterns can be translated into a corresponding estimate of base sequence homology.

In order to demonstrate how the possible numbers of eliminations and additions relating the sets of non-matching fragments of two related DNAs can be determined, we shall consider a hypothetical example in which endonuclease digestion of two related DNAs yields 10 and 11 fragments respectively, 7 of which are homologous. The 3 non-homologous fragments of DNA A can be transformed into the 4 non-homologous fragments of DNA B in several different ways, depending on how the homologous and non-homologous fragments of DNA A are distributed on the genome.

There are three basically different distributions of the 3 non-homologous fragments of DNA A along the genome (Fig. 1): (1) the 3 fragments $X_A$, $Y_A$ and $Z_A$ are separated by one or several homologous fragments; (2) 2 non-homologous fragments are adjacent and the third one is isolated; (3) the 3 fragments are connected. Case 1 is not compatible with the measured number of fragments and can be excluded: all existing restriction sites must remain, as they terminate homologous fragments; but adding restriction sites within every non-homologous fragment raises their number to at least six, in contradiction with the required four fragments.

In case 2 only the restriction site separating the two non-homologous fragments can be eliminated; indeed, it must be eliminated if the impasse of case 1 is to be avoided. As we see, separating two non-homologous fragments by a homologous sequence or by a non-eliminated restriction site gives identical results, as, in both cases, the termination of the non-homologous fragments is conserved. Let us thus assume that the restriction site separating the two non-homologous fragments $X_A$ and $Y_A$ has been eliminated. A non-homologous set of fragments will result if fragment $Z_A$ receives at least one additional
Short communications

As a total of four fragments must result, two new cleavage sites will have to be introduced, at least one of which must divide fragment \( Z_a \).

In the third case, all the three non-homologous fragments are adjacent. If no internal restriction site is eliminated, we have, again, case 1. If only one restriction site is eliminated, case 2 results. Finally, if both internal restriction sites are eliminated, three new ones will be required to yield the necessary four fragments. These additional recognition sequences can be introduced anywhere within fragments \( X_a, Y_a \) and \( Z_a \), with the exception of former restriction sites; this again would lead to case 1 or 2.

Thus, a combinatorial analysis permits enumeration of all possible combinations of additions and eliminations of restriction sites compatible with given numbers of non-homologous fragments of two related DNAs. In our case, DNA A can be transformed into DNA B either by eliminating one site and introducing two, or by eliminating two and introducing three.

The general case where \( k \) non-homologous fragments are mapped on to \( k+n \) can be analysed mathematically. The number of additions (\( s \)) and eliminations (\( r \)) of restriction sites must always differ by \( n \) in order to agree with the observed fragment numbers:

\[
s = r+n. \tag{1}
\]

As only recognition sites lying between adjacent non-homologous fragments can be eliminated, a maximum number of

\[
r_{\text{max}} = k-1 \tag{2}
\]

eliminations is possible, namely, if all \( k \) non-homologous fragments are grouped in a single cluster.

The minimal number of eliminated restriction sites can also be derived. According to equation (1), the number of additions can be split into the two terms \( n \) and \( r \). A number of \( n \) new restriction sites modifies at most \( n \) fragments, namely in the case where each addition operates on a different original fragment. If \( n \) is at least equal to \( k \), all fragments have been modified, and the minimum number of eliminations required is zero; if \( n \) is smaller than \( k \) \( (k-n) \) fragments remain to be modified. The remaining number of modifications, if any is required, must result from an equal number (\( r \)) of eliminations and additions. An elimination modifies the two adjacent fragments, while an addition modifies one fragment. Thus, \( r \) eliminations and \( r \) additions modify at best three \( r \) fragments. The smallest non-negative integer at least equal to \( \frac{1}{2}(k-n) \) will therefore be the minimal number of eliminations (\( r_{\text{min}} \)). Summarizing, \( r \) must always lie within the range

\[
\begin{align*}
k-1 &\geq r \\
&\geq 0 \quad (n \geq k) \\
&\geq \frac{1}{2}(k-n) \quad (n < k).
\end{align*} \tag{3}
\]

This range of possible restriction site alterations can now be used to estimate the sequence homology of DNAs A and B. The restriction sites are considered as a sample of the entire population of sequences of six nucleotides. Such a sequence can either be altered or conserved. The probability \( P_a \) for no sequence alteration can be estimated from a binomial model. The total sample consists of three classes of 'sites': unaltered restriction sites, A-specific restriction sites to be eliminated when mapping A on to B, and non-restriction-site sequences of six nucleotides which become B-specific restriction sites by what we call additions. It is therefore not important whether we map A on to B or vice versa, as additions and eliminations can be interchanged, both representing sequence alterations in a six-nucleotide sequence. If \( a \) is the number of restriction sites in DNA A,
the sample size is therefore \((a + s)\), and \((r + s)\) sequences of six nucleotides are non-identical. The binomial probability for the conservation of any six-nucleotide sequence is therefore

\[
P_6 = \frac{a - r}{a + s},
\]

(4)

This can be calculated both for \(r_{\text{min}}/s_{\text{min}}\) and for \(r_{\text{max}}/s_{\text{max}}\) and confidence limits can be determined from well-established statistical tables.

Using equation (4), we can now estimate \(P_6\) for our example. Restriction endonuclease digestion yields 10 fragments for A and 11 for B, 7 of which are homologous. As shown, one or two restriction sites of A are eliminated, while two or three are added to yield the restriction pattern of B. Therefore, we obtain

\[
P_{6\text{min}} = \frac{9 - 2}{9 + 3} = \frac{7}{12} = 58.33\% \quad (27.67 \text{ to } 84.83\%),
\]

\[
P_{6\text{max}} = \frac{9 - 1}{9 + 2} = \frac{8}{11} = 72.73\% \quad (39.03 \text{ to } 93.98\%),
\]

where the values in parentheses represent confidence intervals for an uncertainty of 5%. With limits of 27.67 and 93.98%, the range obtained for \(P_6\) is obviously very large. However, not all values in this range have the same probability of representing the actual value:

(a) The maximum number of restriction site alterations used to calculate the lower value of \(P_6\) is not very probable since it requires all non-homologous fragments to be grouped in a single cluster. Of course, unless restriction site maps are available, we have no \textit{a priori} knowledge of how the homologous and non-homologous fragments are distributed on the genome. However, considering all possible distributions, a single cluster can in most cases be shown to occur with a low probability. Further, even in the case of a maximum length cluster containing all non-homologous fragments, it is by no means necessary that all internal restriction sites be altered. Thus, on a probabilistic basis, the maximum number of restriction site alterations seems less likely than any other value.

(b) The range of \(P_6\) is considerably increased by taking into account confidence limits for the highest and lowest possible values of \(P_6\). Within a confidence interval, however, actual values are more probable near the estimated value than near the edges of the interval. For instance, had we considered confidence limits for an uncertainty of 20% instead of 5%, a range between approx. 36.8% and 88.5% would have been obtained.

We conclude that within the range obtained for \(P_6\), some extreme values, especially in the lower part of the range, are much less probable than others. It is important to bear this in mind, since the mathematical model used below to estimate sequence homology is possibly biased inasmuch as it might yield values which are too high. Further, as we are taking into account confidence limits for an estimated proportion, the ranges can be narrowed by:

(a) Using enzymes which produce more fragments, to increase the sample size. However, when many fragments are obtained, electrophoretic separation may often cause problems and co-migration of fragments which do not represent equivalent sections of the genomes may occur.

(b) Using data from several digestions with different enzymes, again in order to obtain a larger sample.

Having estimated the probability with which any sequence of six nucleotides is conserved, we can now calculate the proportion of conserved single bases. To do this, a mathematical model has to be used, defining how one sequence is transformed into another. Although no accurate model can be given (Schwarz, 1976), we can assume randomly distributed
Short communications

I65

single base mutations if we are aware of possible biases introduced should this model be incorrect. If all bases have the same probability ($P_1$) of being conserved, a six-nucleotide sequence will be conserved with a probability ($P_6$) equal to the sixth power of $P_1$. Inversely, sequence homology (i.e. the probability of nucleotide conservation) is

$$SH = P_1 = \sqrt[6]{P_6}.$$  (5)

Again, this step has to be carried out for the lower and upper confidence limits of $P_6$.

As mentioned, the model used is valid for single base mutations occurring at random but two DNAs with perhaps little heterogeneity in the distribution of mutations or with small deletion/addition type mutations may also be regarded as fitting the model. If there is considerable heterogeneity as to the distribution of base mutations or if large sequences have been deleted or inserted, the values for sequence homology obtained with our model will be too high. However, considering that low values of $P_6$ have a low probability as demonstrated above, we will, despite the possible bias introduced by using this model, in most cases still obtain an interval for sequence homology which contains the actual value. Taking this into account, it is obvious, however, that no error probability can be given for the lower limit of the range, whereas the upper limit can still be said to be exceeded in only less than 2.5% of the cases.

Returning to our example, we can determine the sequence homology of the two DNAs as

$$SH_{\text{min}} = \sqrt[6]{0.2767} = 80.72\%,$$
$$SH_{\text{max}} = \sqrt[6]{0.9398} = 98.97\%.$$

A quantitative estimation of the sequence homology of two closely related DNAs from endonuclease digestion data has its limitations which must be borne in mind.

In the case of circular DNAs, the procedure has to be modified, and for DNAs with repeated or inverted sequences, the approach becomes more complicated. A thorough study of the structure of the DNAs under investigation should thus precede any quantitative analyses of genome relatedness.

Further, it is not possible to prove our basic assumption that co-migrating fragments actually represent equivalent sections of the genomes, unless physical maps of the enzyme-specific restriction sites are available. However, this view is strongly supported by numerous studies of virus genomes and their mutants for which physical maps have been established (Allet et al. 1973; Osborn et al. 1974; Brockman, Gutai & Nathans, 1975; Griffin & Fried, 1975; Lee, Brockman & Nathans, 1975; Murray & Murray, 1975; Sambrook et al. 1975; Thomas & Davis, 1975; Giessmann & Zur Hausen, 1976; Ito, Kawamura & Yanofsky, 1976). We think that this assumption of our theory is thus valid for closely related DNAs if a good resolution of the fragments is achieved and if analyses with several enzymes yield similar results.

Another problem in this mathematical approach is that there is no model, at present, of the way one sequence is actually transformed into another. Viable mutations cannot be expected to be randomly distributed and there are several mutational mechanisms involved, probably with different frequencies. However, although estimates of sequence homology by our method tend to be too high if two DNAs are not related by randomly distributed single-base alterations, we will, for the reasons discussed above, still obtain ranges containing the actual value of sequence homology with considerable certainty.

As, in contrast to a recent publication (Sugden, Summers & Klein, 1976), mainly combinatorial methods are used in this approach, relatively wide ranges of possible values of sequence homology are obtained. However, we think that, even though these ranges may overlap widely for various DNAs (Müller et al. 1978), endonuclease restriction being a very
Short communications

sensitive method, small differences in the upper and lower limits of sequence homologies obtained by this method may be regarded as indicative of an actual difference.

The sequence homology information obtained from restriction enzyme fragmentation patterns of closely related DNAs differs from the information resulting from nucleic acid hybridization or electron microscope hetero-duplex analyses (Roberts, 1976). DNA restriction is a sampling technique and even small alterations, such as single base changes, which affect enzyme-specific recognition sequences are readily detected by this method. On the other hand, there are changes in DNAs which can be undetectable by the use of restriction enzymes and which might prove readily detectable by hybridization. Still other alterations might remain undetected by all the methods mentioned and can only be found when the DNA sequences are known nucleotide by nucleotide. However, we feel that our method offers a potentially useful new approach.

We are indebted to Dr Robert Elton for his substantial suggestions in reviewing this paper. We gratefully thank Mrs Elaine Haenggeli-Browse, Dr Ernst Eggengerber, Dr Ernst Peterhans, Dr Walter Schaffner, Dr Riccardo Wittek and Dr Robert Wyler for helpful discussions and critically reading the manuscript.

Institut für Virologie der Universität Zürich
Winterthurerstrasse 266a
8057 Zürich

and * Institut für Technische Physik
Eidg. Technische Hochschule
Hönggerberg, 8049, Zürich, Switzerland

REFERENCES


SUGDEN, B., SUMMERS, W. L. & KLEIN, G. (1976). Nucleic acid renaturation and restriction endonuclease cleavage analyses show that the DNAs of a transforming and a non-transforming strain of Epstein–Barr virus share approximately 90 % of their nucleotide sequences. Journal of Virology 13, 765–775.


(Received 19 January 1977)