Sites of cleavage by restriction enzymes in viral DNAs: comparison of observed and expected frequencies

R.A.Elton

Medical Research Council Virology Unit, Church Street, Glasgow G11 5JR, UK

Received 23 August 1974

ABSTRACT

Observed frequencies of restriction enzyme cleavage in viral DNAs are compared to those expected from the cleavage site sequence and the known size and nearest neighbour frequencies of the DNAs. It is found that a number of significant deviations from expectation occur, and possible explanations are put forward.

INTRODUCTION

The frequency of occurrence of short sequences in DNAs has until recently been measurable only in the case of certain general types of sub-sequence such as nearest neighbour base doublets^{1,2} or pyrimidine runs³. With the advent of restriction enzyme analysis⁴, it has become possible to measure directly the frequency with which specific tetramers, pentamers and hexamers occur in viral and other DNA genomes. In the main, these frequencies have been found to be roughly consistent with the size and base composition of the DNAs concerned when a random base sequence is assumed⁵⁻⁸. However, it is known from doublet analyses that the random model is usually inadequate to describe real DNA sequences, and a more realistic calculation of expected restriction site frequency can be made by taking the doublet data into account. In this paper, calculations of this type are applied to published site frequency analyses, and a number of cases are identified where discrepancies occur between theory and observation. Among the explanations considered for these discrepancies are undetected protection of cleavage sites, selection against the presence of sites, incorrect site sequence identification and genuine non-randomness in the distribution of sites in the DNA.

EXPECTED FREQUENCIES

The expected frequencies are calculated on the assumption that the DNA is a random sequence of bases with nearest neighbour dependence defined by the independently determined base doublet frequencies of the DNA. The probability of a sequence $X_1X_2 \dots X_n$, where each X_i represents a particular base, is calculated as $P(X_1)P(X_2/X_1) \dots P(X_n/X_{n-1})$ where $P(X_i)$ is the proportion of base X_i in the DNA and $P(X_i/X_j)$ is the proportion of times that the base following X_j is X_i . The expected site frequency is the product of this probability with the number of bases in the DNA molecule. For site sequences which are not self-complementary, the expected frequency in double-stranded DNA is the sum of those for the two complementary sequences. The probability distribution of the number of sites in a statistically homogeneous DNA is Poisson, and a confidence interval for the true expected number can be set up from a given observed number by using standard tables (e.g. ref. 9). A significant difference between observed and expected frequency occurs when such an interval at the 95% level does not include the calculated expected frequency.

Most doublet frequencies are determined on double-stranded DNA, and are hence only averages of those in the two single strands. If the strands differ substantially in doublet frequency, their expected site frequencies may also differ, and it is worth checking whether this introduces an appreciable bias into the double-stranded calculation. The single-stranded doublet frequencies from ϕ X174 DNA have been used to carry out this check for several different site sequences. While it is found that the expected frequencies from the double-stranded doublet frequencies consistently exceed those calculated from the single-stranded data, the excess is never more than 3%, and we can probably conclude that this bias contributes little to any deviations from expectation in observed site frequencies.

Another possible source of bias arises from compositional heterogeneity in viral genomes. For example, the A+T-rich nature of the EcoRI site implies that it should occur less frequently in high -G+C DNAs than in low -G+C ones. DNAs which have intramolecular G+C content heterogeneity, such as λ ¹⁰ or Adeno¹¹, should thus show a biased distribution of EcoRI cleavage sites along the genome, and both of them have in fact been shown

to do so^{12,13}. In the case of λ , the sites cluster as expected in the A+T-rich half of the molecule. The doublet frequencies of the two halves of λ DNA have been determined¹⁴, and expected site frequencies for each half can be calculated and added to give a more accurate estimate of the total expected frequency. The resulting value of 14.77 differs only slightly from the 16.25 calculated from the overall doublet frequencies, suggesting that intramolecular heterogeneity also probably only contributes a small bias to the calculations.

RESULTS

Cleavage site frequencies have been found for a number of viral DNAs and restriction enzymes, and Table 1 lists those cases so far published. The notation for the enzymes follows the convention of Smith and Nathans²⁵, and the sites are shown as single strands running in the 5' $\rightarrow$ 3' direction, with alternative bases indicated by the symbols: Y = pyrimidine, R = purine, Z = T or A. The frequencies themselves are shown in Tables 2 and 3, together with the corresponding expected frequencies calculated as above, a significant difference between observed and expected frequency being indicated by an asterisk. Doublet frequencies are not available for the DNAs of T3 and fl, and expected frequencies were calculated in these cases on the assumption that their doublet patterns are the same as those of T7 and ϕ X174 respectively.

Table 1. Viral DNAs for which restriction enzyme site frequencies are known.

Enzyme	Site	Viral DNAs
EcoRI	GAATTC ^{15,16}	λ ¹⁵ , Herpes I ¹⁷ , Adeno 2 ¹³ , SV40 ¹⁸ , Polyoma ¹⁹
EcoRII	CCZGG	λ ⁶ , SV40 ²⁰
EcoPI	AGATCT ¹⁶	λ ¹⁶
HindII	GTYRAC ⁵	λ ²¹ , T7 ²¹ , P22 ⁵ , ϕ X174 ⁸ , fl ²² , SV40 ¹⁸
HindIII	AAGCTT ²³	λ ²¹ , T7 ²⁴ , Herpes I ¹⁷ , SV40 ¹⁸ , Polyoma ¹⁹
Hae	GGCC ²⁵	ϕ X174 ²⁶
Hap/HpaII	CCGG ^{7,27}	λ ²⁸ , T3 ⁷ , ϕ X174 ²⁹ , fl ⁷ , SV40 ¹⁸ , Polyoma ¹⁹
HpaI	GTTAAC ²⁷	SV40 ¹⁸

Table 2. Cleavage Site frequencies in bacteriophage DNAs.

Virus	λ	T3	T7	P22	ϕ X174	fl
Doublet reference	1	+	30	30	2	+
EcoRI	5* (16.25)					
EcoRII	20* (87.87)					
EcoPI	10 (9.50)					
HindII	34 (38.13)		51 (40.80)	30 (29.06)	13 (4.81)	1* (5.77)
HindIII	6 (12.82)		0* (10.38)			
Hae					11 (14.32)	
Hap/HpaII	> 50* (185.26)	50* (90.52)			8 (10.56)	13 (12.67)

+ Not available (see text).

In many cases, the expected frequencies calculated here do not differ to any great extent from those obtained using base composition alone, but there are some examples where the difference is crucial. For example, the expected frequency of HpaII (CCGG) sites in SV40 DNA is greatly reduced by the extreme shortage of the doublet CG in this DNA (expected frequency of 7.20 in random DNA of 40% G+C, as compared to 1.54 in SV40 DNA). The doublet prediction for SV40 is in agreement with the observed frequency (Table 3), as is the somewhat larger frequency of 5.40 for Polyoma DNA predicted from the higher G+C content and less severe shortage of CG in this DNA.

The majority of the site frequencies accord satisfactorily with expectation, suggesting that the model of random sequence with nearest neighbour dependence is adequate in these cases, but the situations of special interest are those in which a significant deviation occurs between observed and expected frequency.

Table 3. Cleavage site frequencies in animal virus DNAs.

Virus	Herpes I	Adeno 2	SV40	Polyoma
Doublet Reference	30	31	31	32
EcoRI	13 (20.05)	5 (7.38)	1 (1.26)	1 (1.24)
EcoRII			17* (8.45)	
HindII			5 (3.81)	
HindIII	11 (12.94)		6 (3.10)	2 (2.13)
Hap/HpaII			1 (1.54)	8 (5.40)
HpaI			3 (1.19)	

Edgell et al⁸ have found that ØX174 DNA is cleaved at thirteen sites by nuclease 'Endo R'. It is not clear whether this enzyme corresponds to HindII alone or is a mixture of HindII and III activities, but even in the latter case the expected frequency is only 7.06, considerably less than observed. Edgell et al themselves comment on the unexpectedly high number of cleavages. The contrast with f1 DNA is marked: a somewhat longer DNA of the same base composition contains only one site, significantly less than expected. A similar situation occurs with the cleavage of SV40 and λ DNAs by EcoRII. The SV40 DNA is only one-tenth the size of that of λ, but contains almost as many sites, and the different doublet frequencies of the two species cannot account for the discrepancy. λ DNA also shows significantly fewer sites than expected with EcoRI and HpaII, as does T3 with Hap and T7 with HindIII (in the latter case, no cleavages occur at all where over ten are expected). The case of λ and HpaII is not clear-cut: only about 50 cleavage fragments can be discerned, but there may well be substantially more than this²⁸.

DISCUSSION

Of the 27 observed frequencies, 8 lie outside the 95% prediction interval for expected frequency. This is highly significantly more than the one in twenty expected by chance alone, and the small effects of single-strand bias and heterogeneity are unlikely to account for the majority of these cases. Before we conclude that they therefore reflect a genuine excess or shortfall in the occurrence of the site sequence, however, it should be pointed out that observed frequencies on the low side could also result from the sites being protected from nuclease action. This is trivially true for the DNAs of the bacteria which code for the restriction enzymes: their DNAs are specifically modified to prevent self-degradation. A so-far undetected modification could also explain the lack of any sites for HindIII cleavage in T7 DNA, and possibly even the reduced occurrence of some sites in, for example, λ .

Another possible cause of low observed frequencies is direct selection against the presence of cleavage sites in a DNA. This has been observed in laboratory experiments with λ (K. Murray, personal communication), and is explained by the fact that the rate of degradation of the DNA is dependent on the overall number of sites present. If such selection has taken place in the wild, it could show up in the form of low frequencies of cleavage sites for enzymes from potential host bacteria (such as λ /E. Coli), but presumably not for species outwith the host range (λ /H. influenzae).

The most interesting alternative is that deviation from expectation reflects a genuine non-randomness in the distribution of the sites in the DNA which is not related to direct selection on their presence. This could result from the involvement of the precise site sequence in a very specific functional role in the virus life cycle. Such a situation would be unlikely to occur in the polypeptide-specifying part of the genome, since the genetic code redundancy allows considerable freedom in the sequence of bases coding for a given amino acid sequence. Recognition signals for the attachment of enzymes (e.g. for transcription) are attractive candidates for such functional sequences, especially since the

self-complementary nature of the restriction enzyme cleavage sites could place them at the extremities of regions of base-paired protuberance in DNA⁴. This suggests the possibility of cleaving DNAs into natural genetic units by restriction enzyme digestion, although it is clearly too early to identify this sort of situation with any of the deviations discussed here. Recent results^{33,34} do, however, indicate that some *E. coli* RNA polymerase promoter sites contain HindII sequences.

Tables 2 and 3 show no consistent excesses or shortfalls with respect to expectation for any individual enzyme. Were this to be the case, it might suggest the possibility of an error in the site sequence (as occurred in the original identification of the EcoRI site as ZGAATTCZ¹⁵, leading to much lower expected frequencies than those for the correct sequence). While the statistical evidence alone would not justify the rejection of a proposed site sequence, a consistent bias over a number of different DNAs should encourage reappraisal of the direct experimental evidence for the sequence in question.

ACKNOWLEDGEMENTS

I am grateful to J. B. Clements, N. M. Wilkie, G. J. Russell, J. H. Subak-Sharpe and K. Murray for discussion and for access to unpublished data.

REFERENCES

- 1 Josse, J., Kaiser, A.D. and Kornberg, A. (1961) *J. Biol. Chem.* 236, 864-875
- 2 Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962) *J. Biol. Chem.* 237, 1961-1967
- 3 Burton, K. and Peterson, G.B. (1960) *Biochem. J.* 75, 17-27
- 4 Meselson, M., Yuan, R. and Heywood, J. (1972) *Ann. Rev. Biochem.* 41, 447-466
- 5 Kelly, T.J. and Smith, H.O. (1970) *J. Mol. Biol.* 51, 393-409
- 6 Bigger, C.H., Murray, K. and Murray, N.E. (1973) *Nature New Biol.* 244, 7-10
- 7 Sugisaki, H. and Takanami, M. (1973) *Nature New Biol.* 246, 138-140
- 8 Edgell, M.H., Hutchinson, C. A. and Sclair, M. (1972) *J. Virol.* 9, 574-582
- 9 Documenta Geigy Scientific Tables (1968), 6th edn., p. 107, Geigy (U.K.), Macclesfield
- 10 Skalka, A., Burgi, E. and Hershey, A.D. (1968) *J. Mol. Biol.* 34, 1-16
- 11 Doerfler, W. and Kleinschmidt, A.K. (1970) *J. Mol. Biol.* 50, 579-593
- 12 Allet, B., Jeppesen, P.G.N., Katagiri, K. J. and Delius, H. (1973) *Nature (London)* 241, 120-123

- 13 Petterson, U., Mulder, C., Delius, H. and Sharp, P.A. (1973) Proc. Nat. Acad. Sci. U.S. 70, 200-204
- 14 Maitra, U., Cohen, S.N. and Hurwitz, J. (1966) Cold Spr. Harb. Symp. Quant. Biol. 31, 113-122
- 15 Hedgpeth, J., Goodman, H.M. and Boyer, H.W. (1972) Proc. Nat. Acad. Sci. U.S. 69, 3448-3452
- 16 K. Murray, Unpublished data
- 17 Wilkie, N.M., Clements, J.B., MacNab, J.C.M. and Subak-Sharpe, J.H. (1974) Cold Spr. Harb. Symp. Quant. Biol. 39 (in the press)
- 18 Danna, K. J., Sack, G.H. and Nathans, D. (1973) J. Mol. Biol. 78, 363-376
- 19 Griffin, B.E., Fried, M. and Cowie, A. (1974) Proc. Nat. Acad. Sci. U.S. 71, 2077-2081
- 20 Boyer, H.W., Chow, L.T., Dugaiczky, A., Hedgpeth, J. and Goodman, H.M. (1973) Nature New Biol. 244, 40-43
- 21 Landy, A., Ruedisueli, E., Robinson, L., Foeller, C. and Ross, W. (1974) Biochemistry 13, 2134-2142
- 22 Takanami, M. and Kojo, H. (1973) FEBS Lett. 29, 267-270
- 23 Roy, P.H. and Smith, H.O. (1973) J. Mol. Biol. 81, 445-459
- 24 Roy, P.H. and Smith, H.O. (1973) J. Mol. Biol. 81, 427-444
- 25 Smith, H.O. and Nathans, D. (1973) J. Mol. Biol. 81, 419-423
- 26 Middleton, J.H., Edgell, M.H. and Hutchinson, C.A. (1972) J. Virol. 10, 42-50
- 27 Garfin, D.E. and Goodman, H.M. (1974) Biochem. Biophys. Res. Commun. 59, 108-116
- 28 Allet, B. (1973) Biochemistry 12, 3972-3977
- 29 Johnson, P.H., Lee, A.S. and Sinsheimer, R.L. (1973) J. Virol. 11, 596-599
- 30 G. J. Russell, unpublished data
- 31 Morrison, J.M., Keir, H.M., Subak-Sharpe, J.H. and Crawford, L.V. (1967) J. Gen. Virol. 1, 101-108
- 32 Subak-Sharpe, J.H., Bürk, R.R., Crawford, L.V., Morrison, J.M., Hay, J. and Keir, H.M. (1966) Cold Spr. Harb. Symp. Quant. Biol. 31, 737-748
- 33 Allet, B., Roberts, R.J., Gesteland, R.F. and Solem, R. (1974) Nature (London) 249, 217-221
- 34 Maurer, R., Maniatis, T. and Ptashne, M. (1974) Nature (London) 249, 221-223